Collecting Data

Instructor:

Emmanuel Thompson

Statistics

- **Statistics** is a collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions.

- **Statistics** provides tools that you need in order to react intelligently to information you hear or read.
  - In this sense, Statistics is one of the most important things that you can study.

Populations and Samples

- **Before collecting and analyzing data in any statistical work, we first must identify and characterize the population to be studied.**
  - If we want to study the amount of money spent on textbooks by a typical first-year college student, our population might be all first-year students at your college.

Population or Target Population:

- The population of a study is the group the collected data is intended to describe.
- Gathering data on an entire population is often impractical, we usually select a sample to study.

Sample

- A sample is a smaller subset of the entire population, ideally one that is representative of the whole population.

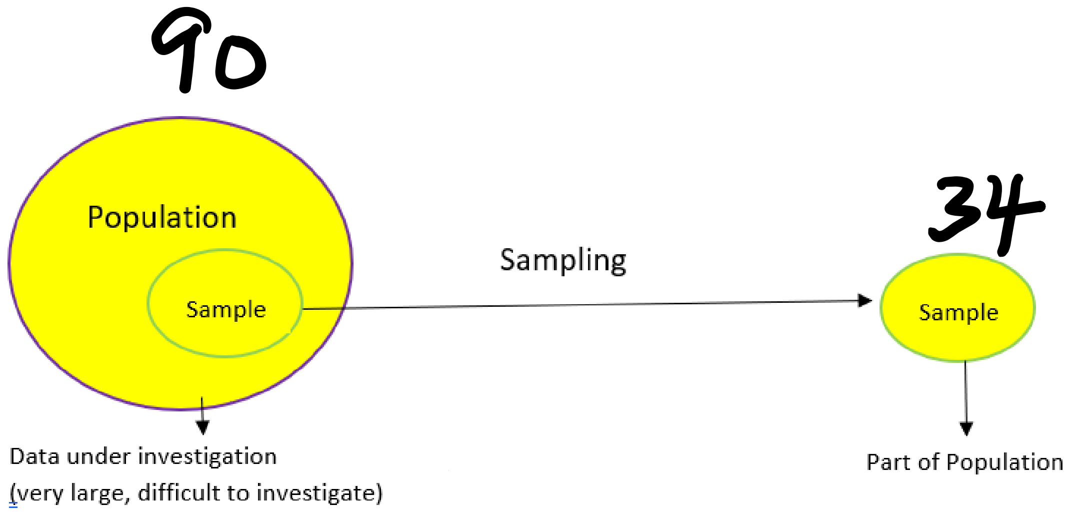
Populations and Samples

Q1:

A political scientist surveys 34 of the current 90 representatives in a state's legislature.

What is the size of the sample: 34

What is the size of the population: 90



90

Population

Sample

Sampling

34

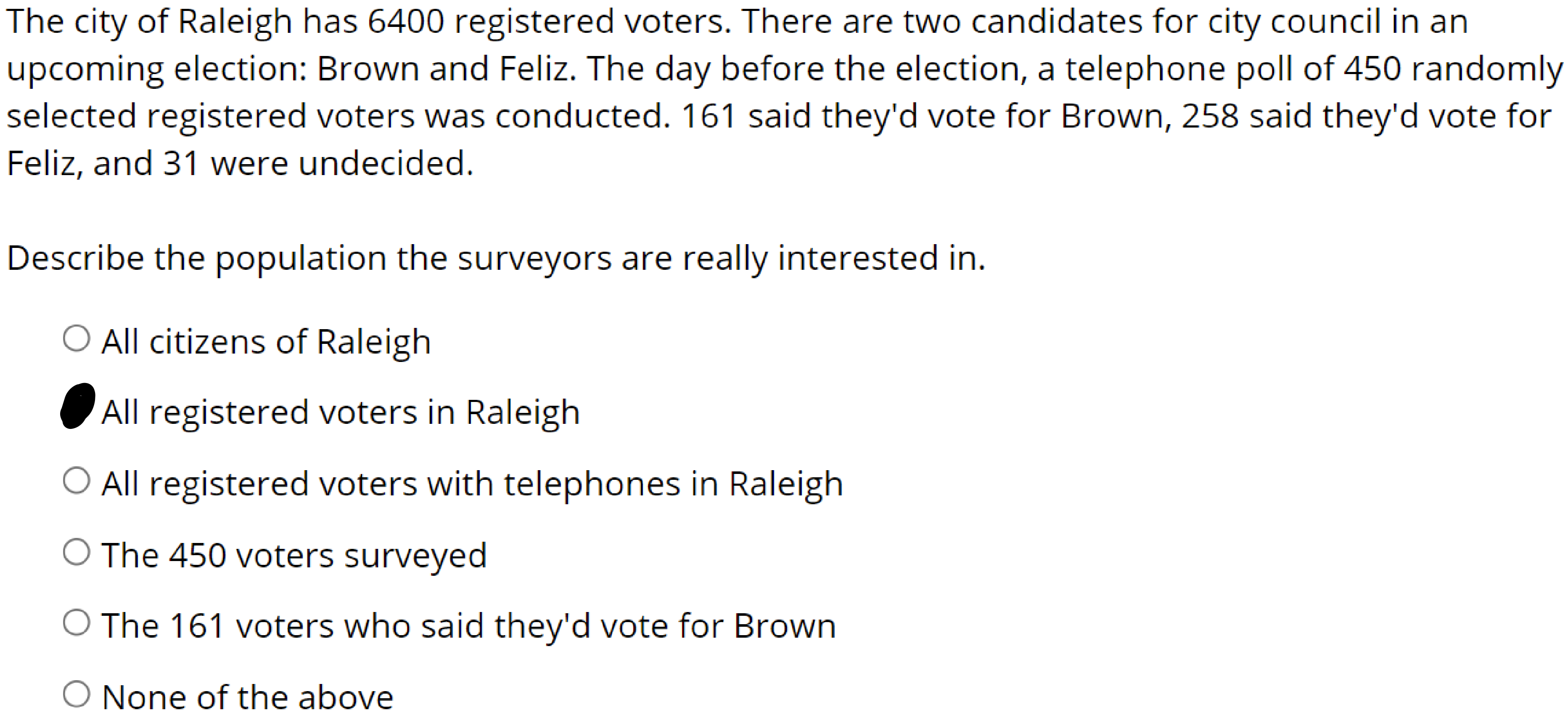
Sample

Data under investigation
(very large, difficult to investigate)

Part of Population

Populations and Samples

Q2:



The city of Raleigh has 6400 registered voters. There are two candidates for city council in an upcoming election: Brown and Feliz. The day before the election, a telephone poll of 450 randomly selected registered voters was conducted. 161 said they'd vote for Brown, 258 said they'd vote for Feliz, and 31 were undecided.

Describe the population the surveyors are really interested in.

○ All citizens of Raleigh

● All registered voters in Raleigh

○ All registered voters with telephones in Raleigh

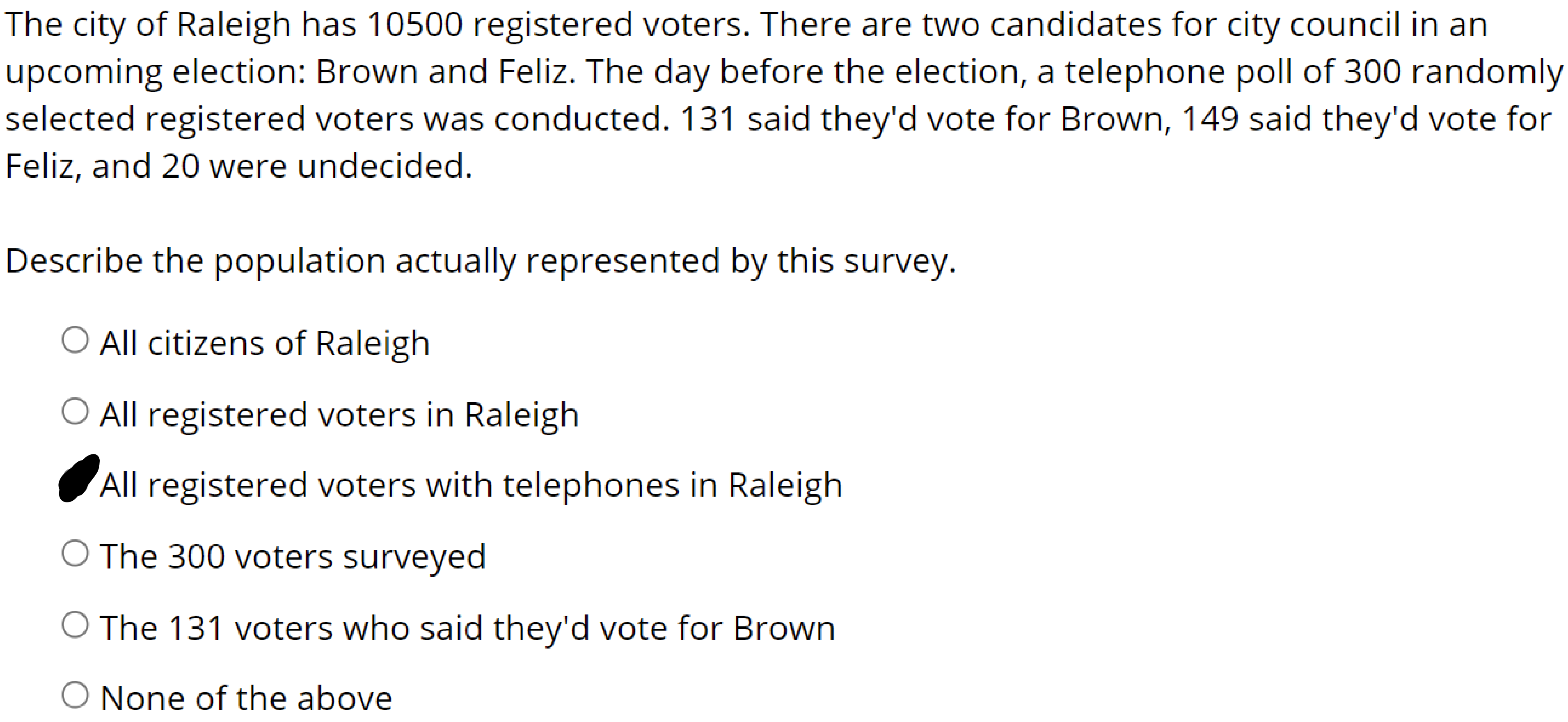
○ The 450 voters surveyed

○ The 161 voters who said they'd vote for Brown

○ None of the above

Populations and Samples

Q3:



The city of Raleigh has 10500 registered voters. There are two candidates for city council in an upcoming election: Brown and Feliz. The day before the election, a telephone poll of 300 randomly selected registered voters was conducted. 131 said they'd vote for Brown, 149 said they'd vote for Feliz, and 20 were undecided.

Describe the population actually represented by this survey.

○ All citizens of Raleigh

○ All registered voters in Raleigh

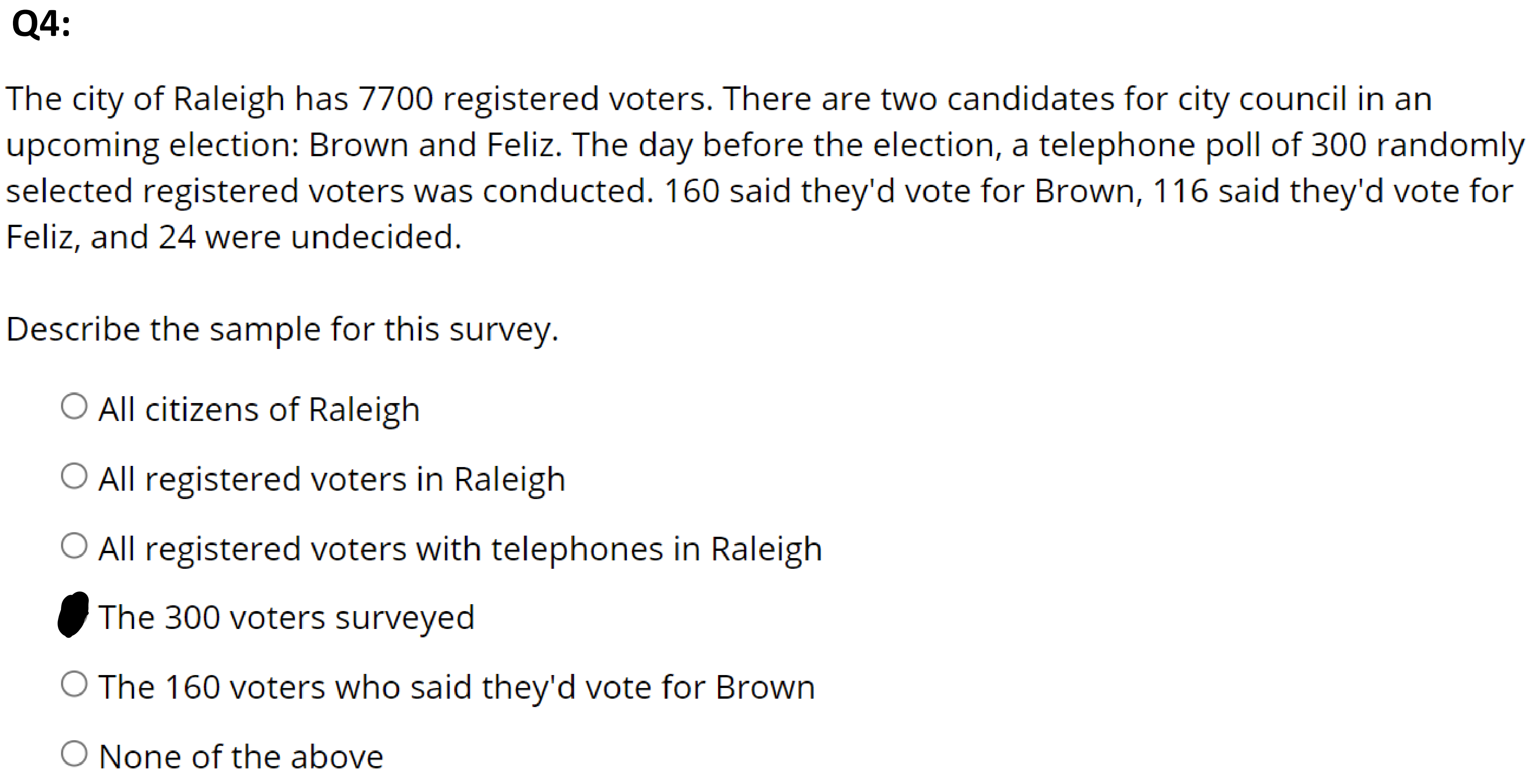
● All registered voters with telephones in Raleigh

○ The 300 voters surveyed

○ The 131 voters who said they'd vote for Brown

○ None of the above

Populations and Samples



Q4:

The city of Raleigh has 7700 registered voters. There are two candidates for city council in an upcoming election: Brown and Feliz. The day before the election, a telephone poll of 300 randomly selected registered voters was conducted. 160 said they'd vote for Brown, 116 said they'd vote for Feliz, and 24 were undecided.

Describe the sample for this survey.

- ○ All citizens of Raleigh
- ○ All registered voters in Raleigh
- ○ All registered voters with telephones in Raleigh
- ● The 300 voters surveyed
- ○ The 160 voters who said they'd vote for Brown
- ○ None of the above

Parameters and Statistics

Parameter

- A parameter is a value (average, percentage, etc.) calculated using all the data from a population.

Census

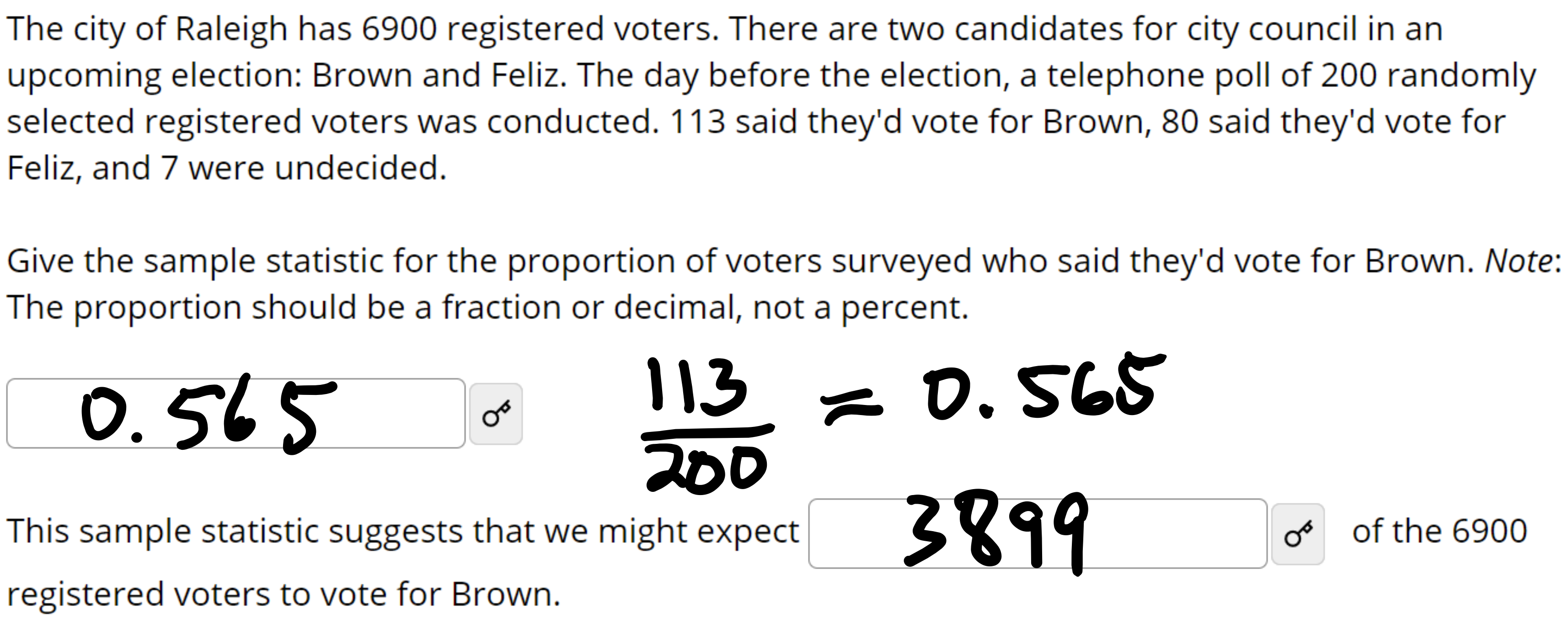
- A survey of an entire population is called a **census**.

Statistic

- A statistic is a value (average, percentage, etc.) calculated using the data from a sample.

Parameters and Statistics

Q5:



The city of Raleigh has 6900 registered voters. There are two candidates for city council in an upcoming election: Brown and Feliz. The day before the election, a telephone poll of 200 randomly selected registered voters was conducted. 113 said they'd vote for Brown, 80 said they'd vote for Feliz, and 7 were undecided.

Give the sample statistic for the proportion of voters surveyed who said they'd vote for Brown. *Note*: The proportion should be a fraction or decimal, not a percent.

[0.565]

This sample statistic suggests that we might expect [3899] of the 6900 registered voters to vote for Brown.

$$\frac{113}{200} = 0.565$$

$$0.565 \cdot 6900 \simeq 3898.5 \approx 3899$$

Candidate	Votes
Brown	113
Feliz	80
Undecided	7
Total	200

Categorical and Quantitative Data

- Once we have gathered data, we might wish to classify it.
- Data can be classified as categorical or quantitative data.

Categorical (Qualitative) data

- Categorical (qualitative) data are pieces of information that allow us to classify the objects under investigation into various categories.

Quantitative (Numerical) data

- Quantitative data are responses that are numerical in nature and with which we can perform meaningful arithmetic calculations.

Categorical and Quantitative Data

Q6:

a) We might conduct a survey to determine the name of the favorite movie that each person in a math class saw in a movie theater.

Categorical data

Q7:

A survey could ask the number of movies you have seen in a movie theater in the past 12 months (0, 1, 2, 3, 4, ...)

Quantitative data

Q8:

Classify each measurement as categorical or quantitative

a. Eye color of a group of people Categorical data

b. Daily high temperature of a city over several weeks Quantitative data

c. Annual income Quantitative data

Sampling Methods

- In any statistical study, once we have identified the population of interest, the question now is, how do we choose an appropriate sample?

- There are a lot of ways to sample a population, but there is one goal we need to keep in mind:
  - we would like the sample to be representative of the population.

- One way to ensure that the sample reflects the population is to employ randomness.
- The most basic random method is **simple random sampling**.

Simple random sample

- A random sample is one in which each member of the population has an equal probability of being chosen.
- A simple random sample is one in which every member of the population and any group of members has an equal probability of being chosen.

Sampling Methods

Sampling variability

- The natural variation of samples is called sampling variability.
  - This is unavoidable and expected in random sampling, and in most cases is not an issue.

- To help account for variability, statisticians might instead use a stratified sample.

Stratified sampling

- In stratified sampling, a population is divided into several subgroups (or strata).
  - Random samples are then taken from each subgroup with sample sizes proportional to the size of the subgroup in the population.

- A variation on this technique is called quota sampling.

Quota sampling

- Quota sampling is a variation of stratified sampling, where samples are collected in each subgroup until the desired quota is met.

Sampling Methods

- Another sampling method is cluster sampling, in which the population is divided into groups, and one or more groups are randomly selected to be in the sample.

Cluster sampling

- In cluster sampling, the population is divided into subgroups (clusters), and a set of subgroups are selected to be in the sample

- Other sampling methods include systematic sampling.

Systematic sampling

- In systematic sampling, every *n*th member of the population is selected to be in the sample.

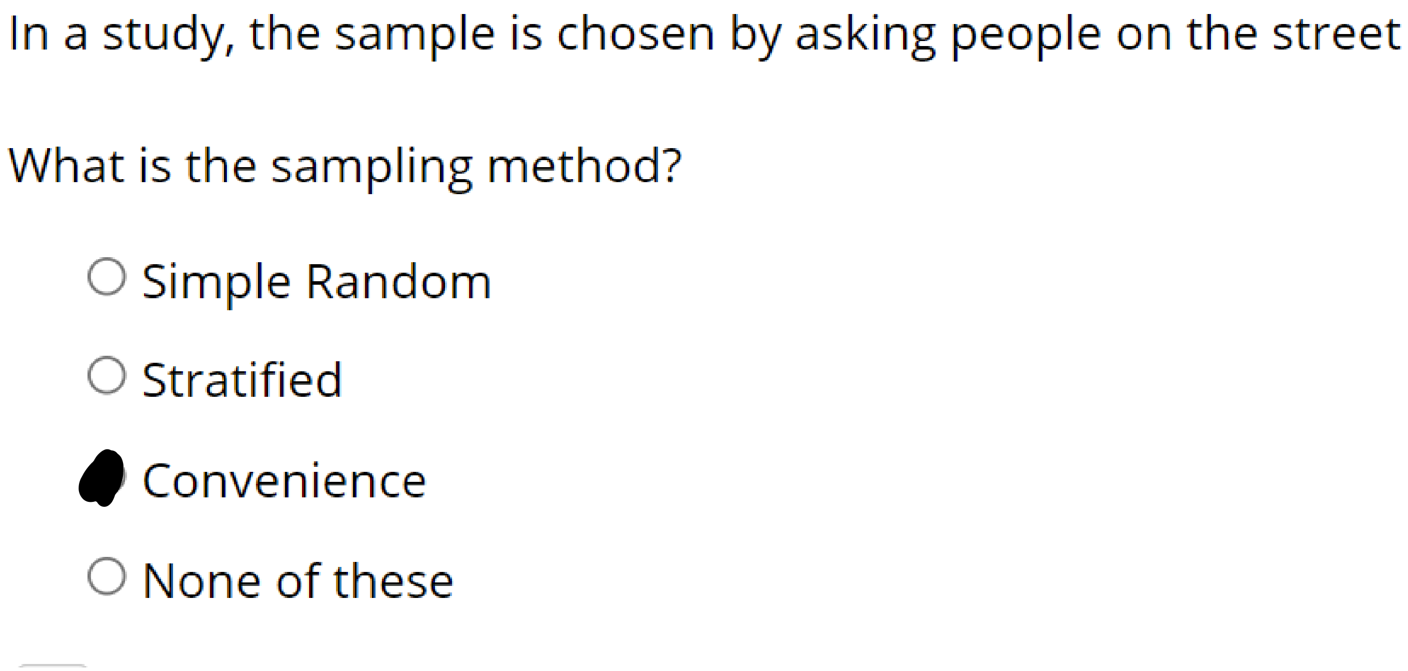
- **The worst types of sampling methods include convenience samples and voluntary response samples**.

Convenience sampling and voluntary response sampling

- Convenience sampling is samples chosen by selecting whoever is convenient.
- Voluntary response sampling is allowing the sample to volunteer.

Sampling Methods

Q9:



In a study, the sample is chosen by asking people on the street

What is the sampling method?

○ Simple Random

○ Stratified

● Convenience

○ None of these

Sampling Methods

Sources of bias

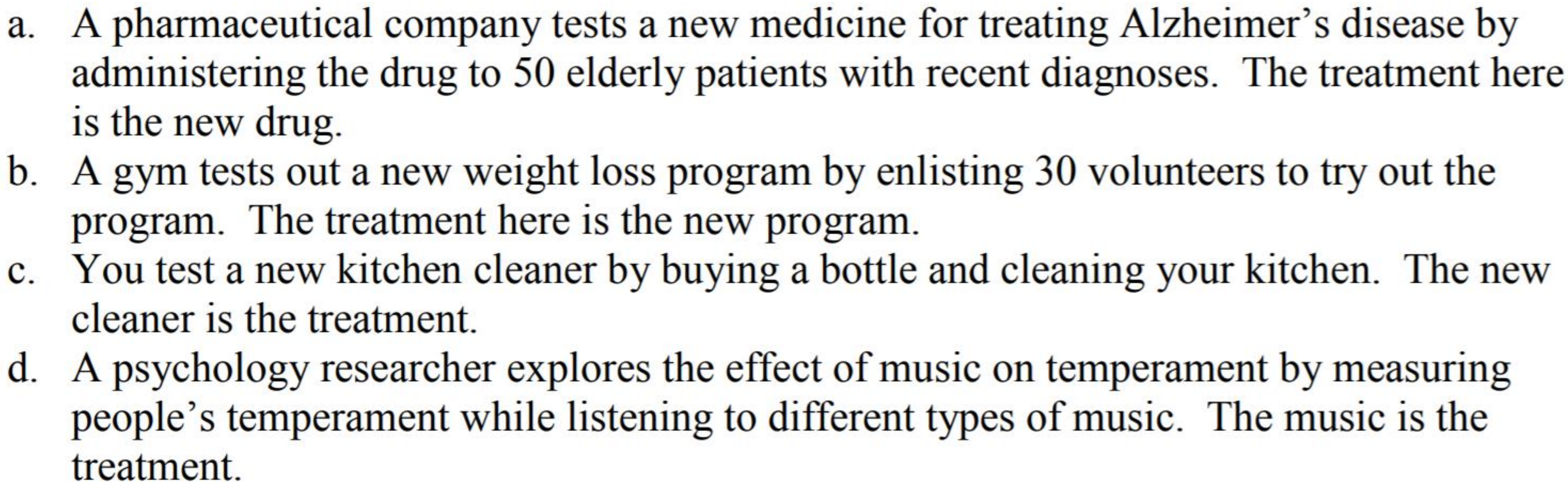
- **Sampling bias –** when the sample is not representative of the population.
- **Voluntary response bias –** the sampling bias that often occurs when the sample is volunteers.
- **Self-interest study –** bias that can occur when the researchers have an interest in the outcome.
- **Response bias –** when the responder gives inaccurate responses for any reason.
- **Perceived lack of anonymity –** when the responder fears giving an honest answer might negatively affect them.
- **Loaded questions –** when the question wording influences the responses.
- **Non-response bias –** when people refusing to participate in the study can influence the validity of the outcome.

Experiments

Observational studies and experiments

- An observational study is a study based on observations or measurements.
- An experiment is a study in which the effects of a **treatment** are measured.

- Here are some examples of experiments:



a. A pharmaceutical company tests a new medicine for treating Alzheimer's disease by administering the drug to 50 elderly patients with recent diagnoses. The treatment here is the new drug.

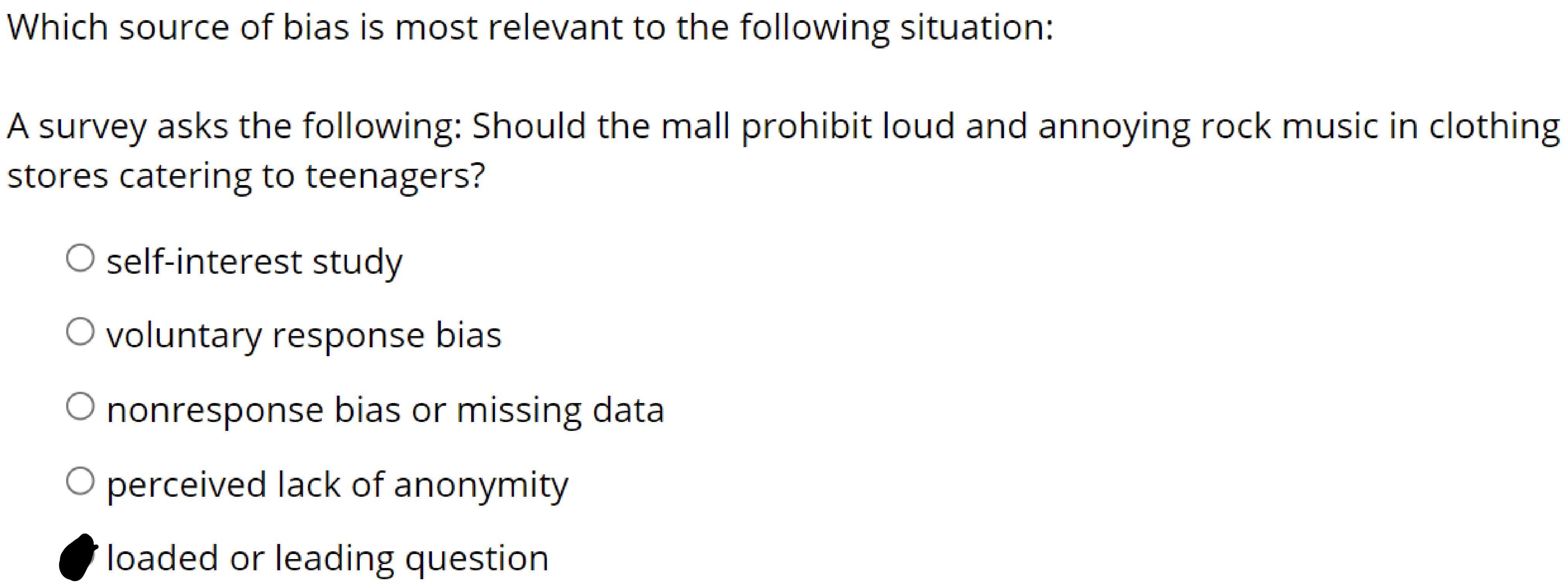
b. A gym tests out a new weight loss program by enlisting 30 volunteers to try out the program. The treatment here is the new program.

c. You test a new kitchen cleaner by buying a bottle and cleaning your kitchen. The new cleaner is the treatment.

d. A psychology researcher explores the effect of music on temperament by measuring people's temperament while listening to different types of music. The music is the treatment.

Sampling Methods

Q10:



Which source of bias is most relevant to the following situation:

A survey asks the following: Should the mall prohibit loud and annoying rock music in clothing stores catering to teenagers?

○ self-interest study

○ voluntary response bias

○ nonresponse bias or missing data

○ perceived lack of anonymity

● loaded or leading question

Experiments

- When conducting experiments, it is essential to isolate the treatment being tested. This is called **confounding**.
- Confounding is the downfall of many experiments, though sometimes it is hidden.

Confounding

- Confounding occurs when there are two potential variables that could have caused the outcome and it is not possible to determine which one actually caused the result.

- There are several measures that can be introduced to help reduce the likelihood of confounding.
  - The primary measure is to use a control group.

Control group

- When using a control group, the participants are divided into two or more groups, typically a control group and a treatment group.
  - The treatment group receives the treatment being tested; the control group does not receive the treatment.

Experiments

Placebo and Placebo controlled experiments

- A placebo is a dummy treatment given to control for the placebo effect.
  - An experiment that gives the control group a placebo is called a **placebo-controlled experiment**.

Placebo effect

- The placebo effect is when the effectiveness of a treatment is influenced by the patient’s perception of how effective they think the treatment will be, so a result might be seen even if the treatment is ineffectual.

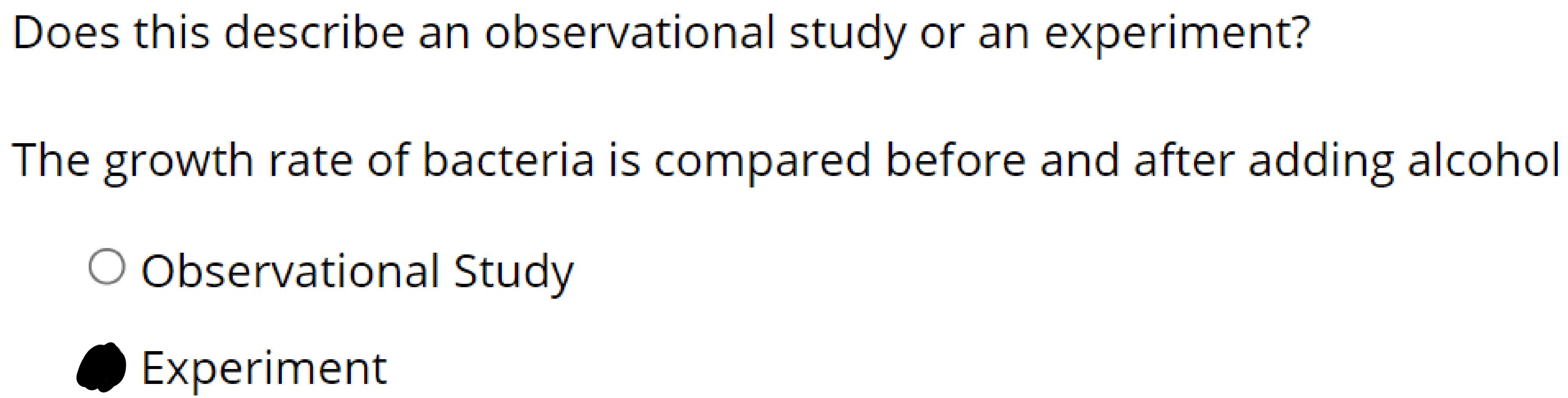
Experiments

Blind studies

- A blind study is one in which the participant does not know whether or not they are receiving the treatment or a placebo.
- A double-blind study is one in which those interacting with the participants don’t know who is in the treatment group and who is in the control group.

Experiments

Q11:



Does this describe an observational study or an experiment?

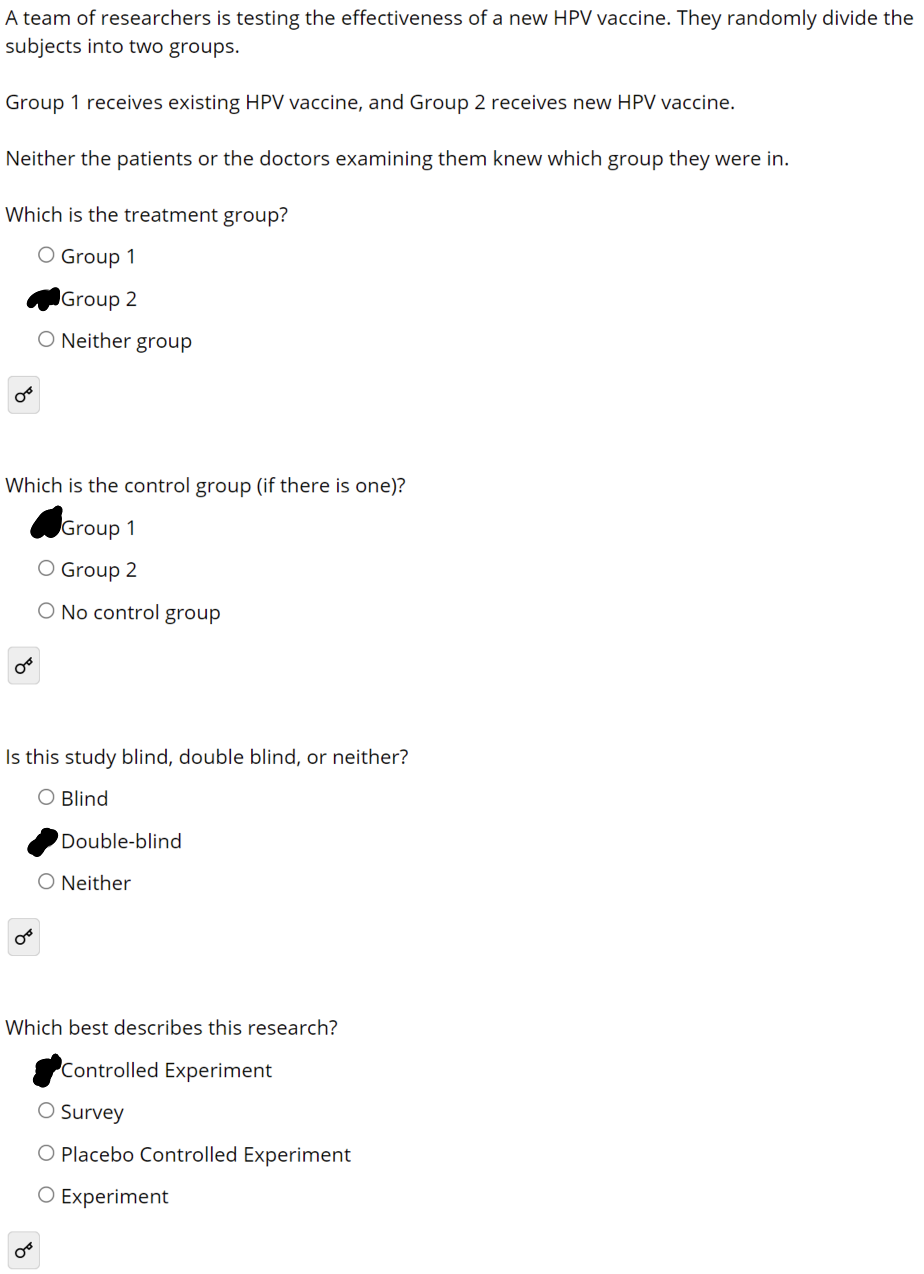
The growth rate of bacteria is compared before and after adding alcohol

○ Observational Study

● Experiment

Experiments

Q12:



A team of researchers is testing the effectiveness of a new HPV vaccine. They randomly divide the subjects into two groups.

Group 1 receives existing HPV vaccine, and Group 2 receives new HPV vaccine.

Neither the patients or the doctors examining them knew which group they were in.

Which is the treatment group?

○ Group 1

● Group 2

○ Neither group

Which is the control group (if there is one)?

● Group 1

○ Group 2

○ No control group

Is this study blind, double blind, or neither?

○ Blind

● Double-blind

○ Neither

Which best describes this research?

● Controlled Experiment

○ Survey

○ Placebo Controlled Experiment

○ Experiment

Make sure to review all the materials and videos on this topic I have posed on canvas.