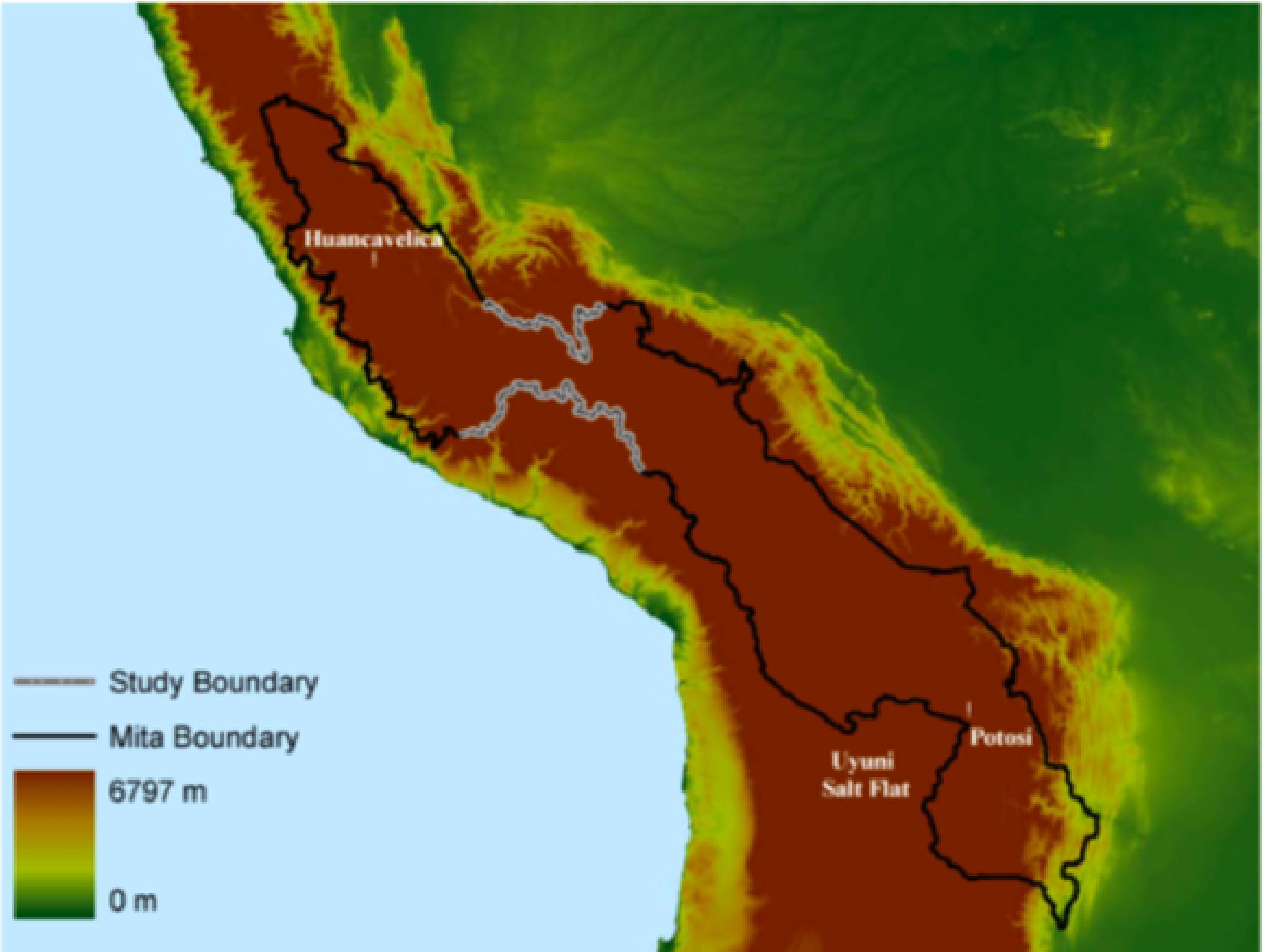
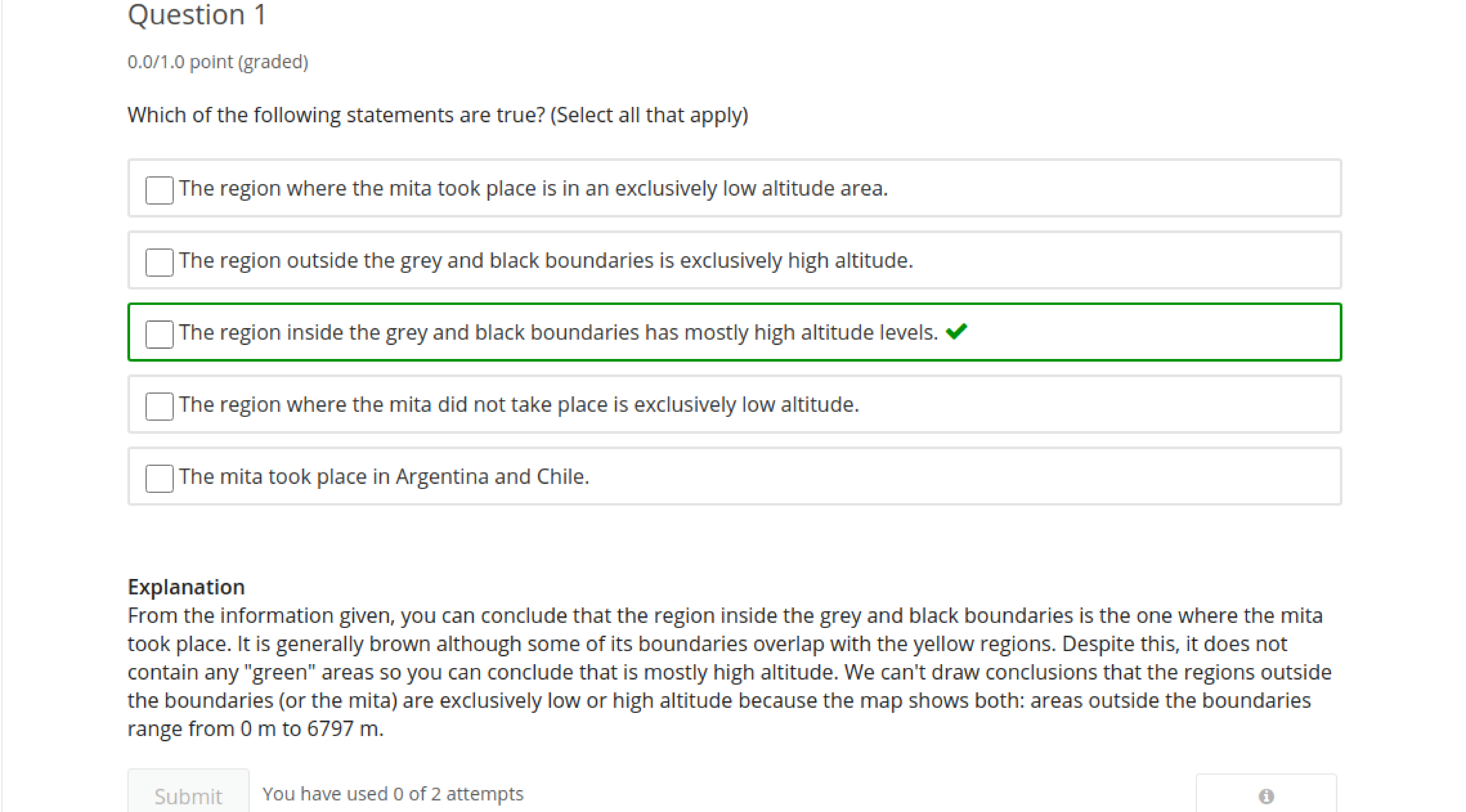
Dell (2010) studies the long-run impacts of the *mita*, an extensive forced mining labor system that was in effect in Peru and Bolivia between 1573 and 1812. The *mita* required over 200 indigenous communities to send one-seventh of their adult male population to work in silver and mercury mines. The *mita* took place within the boundary shown in the figure below (take a close look at the figure and be sure you understand it). It also graphs the altitude of the area with respect to the Earth's sea level (browner areas are at higher levels).

FIGURE 1. The *mita* boundary is in black and the study boundary in light gray. Districts falling inside the contiguous area formed by the *mita* boundary contributed to the *mita*. Elevation is shown in the background.





Question 1

0.0/1.0 point (graded)

Which of the following statements are true? (Select all that apply)

□ The region where the mita took place is in an exclusively low altitude area.

□ The region outside the grey and black boundaries is exclusively high altitude.

□ The region inside the grey and black boundaries has mostly high altitude levels. ✓

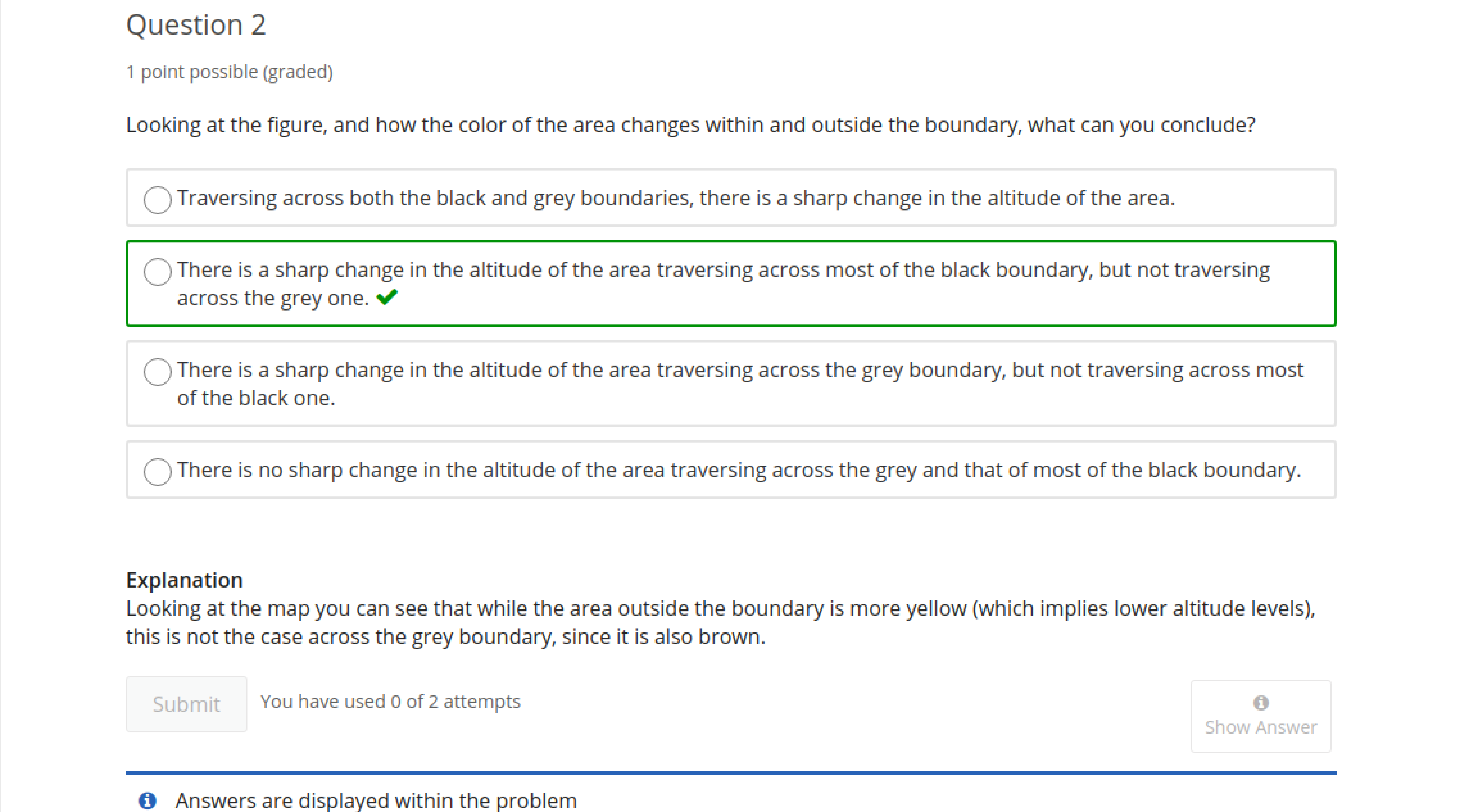
□ The region where the mita did not take place is exclusively low altitude.

□ The mita took place in Argentina and Chile.

Explanation

From the information given, you can conclude that the region inside the grey and black boundaries is the one where the mita took place. It is generally brown although some of its boundaries overlap with the yellow regions. Despite this, it does not contain any "green" areas so you can conclude that is mostly high altitude. We can't draw conclusions that the regions outside the boundaries (or the mita) are exclusively low or high altitude because the map shows both: areas outside the boundaries range from 0 m to 6797 m.

You have used 0 of 2 attempts



Question 2

1 point possible (graded)

Looking at the figure, and how the color of the area changes within and outside the boundary, what can you conclude?

○ Traversing across both the black and grey boundaries, there is a sharp change in the altitude of the area.

○ There is a sharp change in the altitude of the area traversing across most of the black boundary, but not traversing across the grey one. ✓

○ There is a sharp change in the altitude of the area traversing across the grey boundary, but not traversing across most of the black one.

○ There is no sharp change in the altitude of the area traversing across the grey and that of most of the black boundary.

Explanation

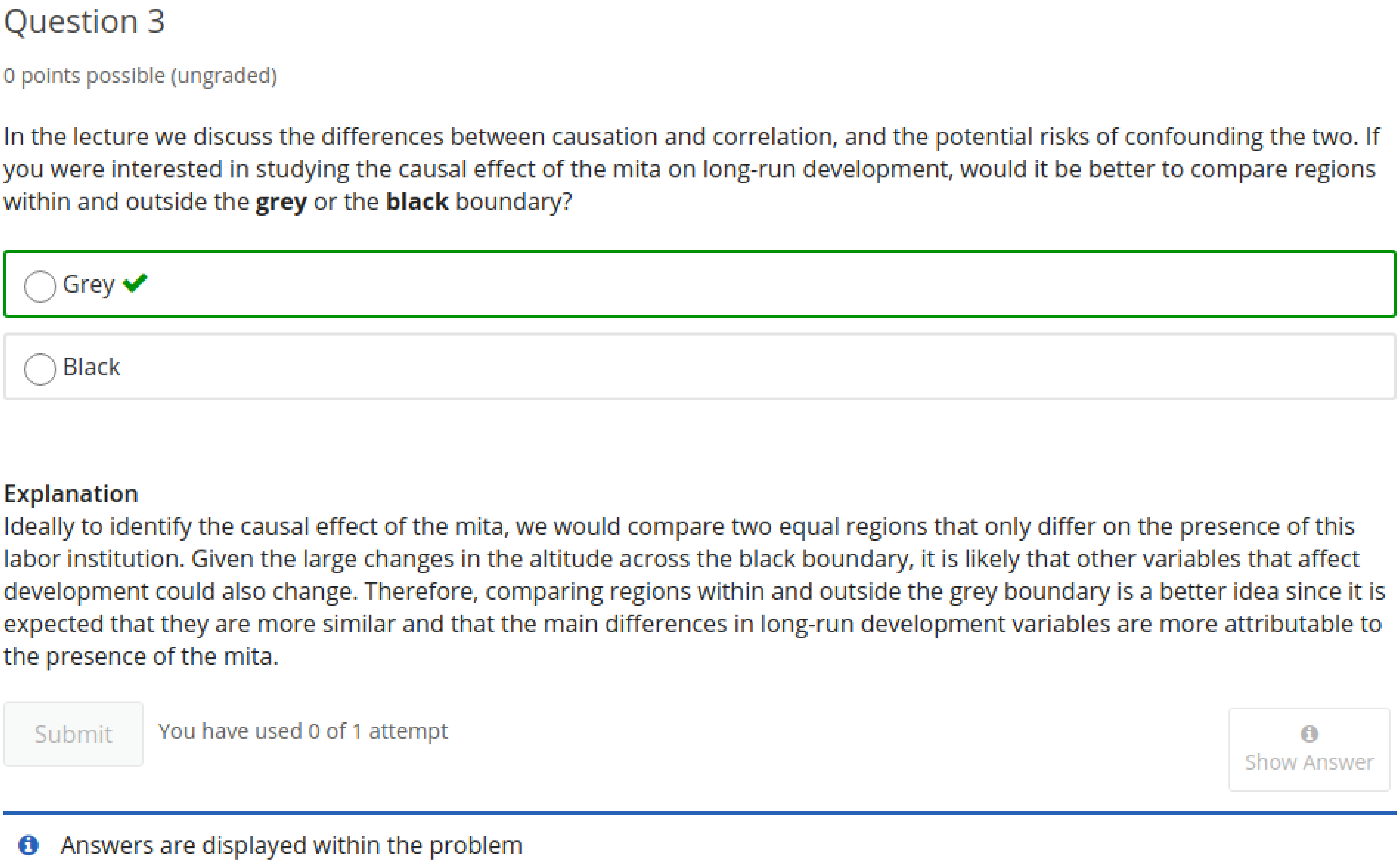
Looking at the map you can see that while the area outside the boundary is more yellow (which implies lower altitude levels), this is not the case across the grey boundary, since it is also brown.

Submit

You have used 0 of 2 attempts

Show Answer

Answers are displayed within the problem



Question 3

0 points possible (ungraded)

In the lecture we discuss the differences between causation and correlation, and the potential risks of confounding the two. If you were interested in studying the causal effect of the mita on long-run development, would it be better to compare regions within and outside the **grey** or the **black** boundary?

○ Grey ✓

○ Black

Explanation

Ideally to identify the causal effect of the mita, we would compare two equal regions that only differ on the presence of this labor institution. Given the large changes in the altitude across the black boundary, it is likely that other variables that affect development could also change. Therefore, comparing regions within and outside the grey boundary is a better idea since it is expected that they are more similar and that the main differences in long-run development variables are more attributable to the presence of the mita.

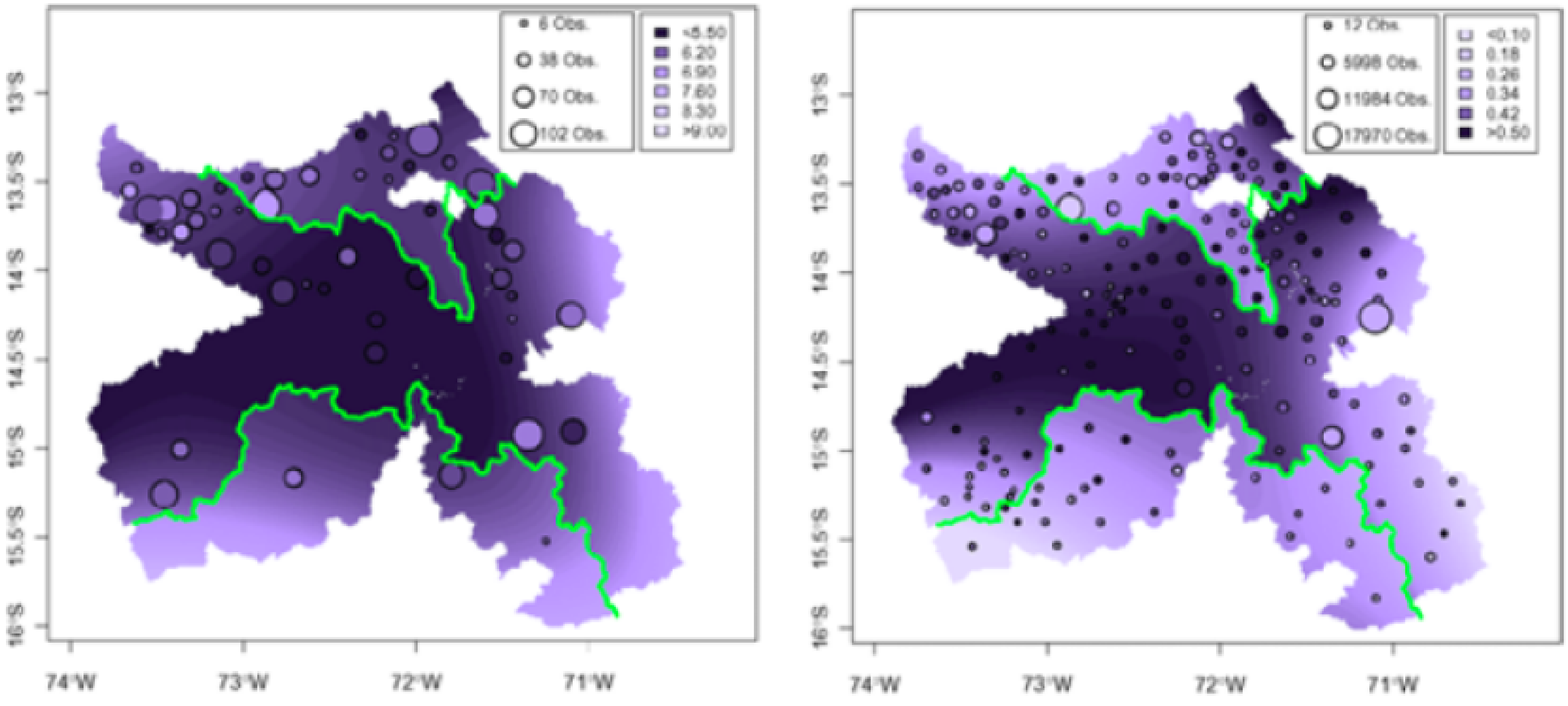
You have used 0 of 1 attempt

Answers are displayed within the problem

Homework due Sep 14, 2020 19:30 EDT **Past Due**

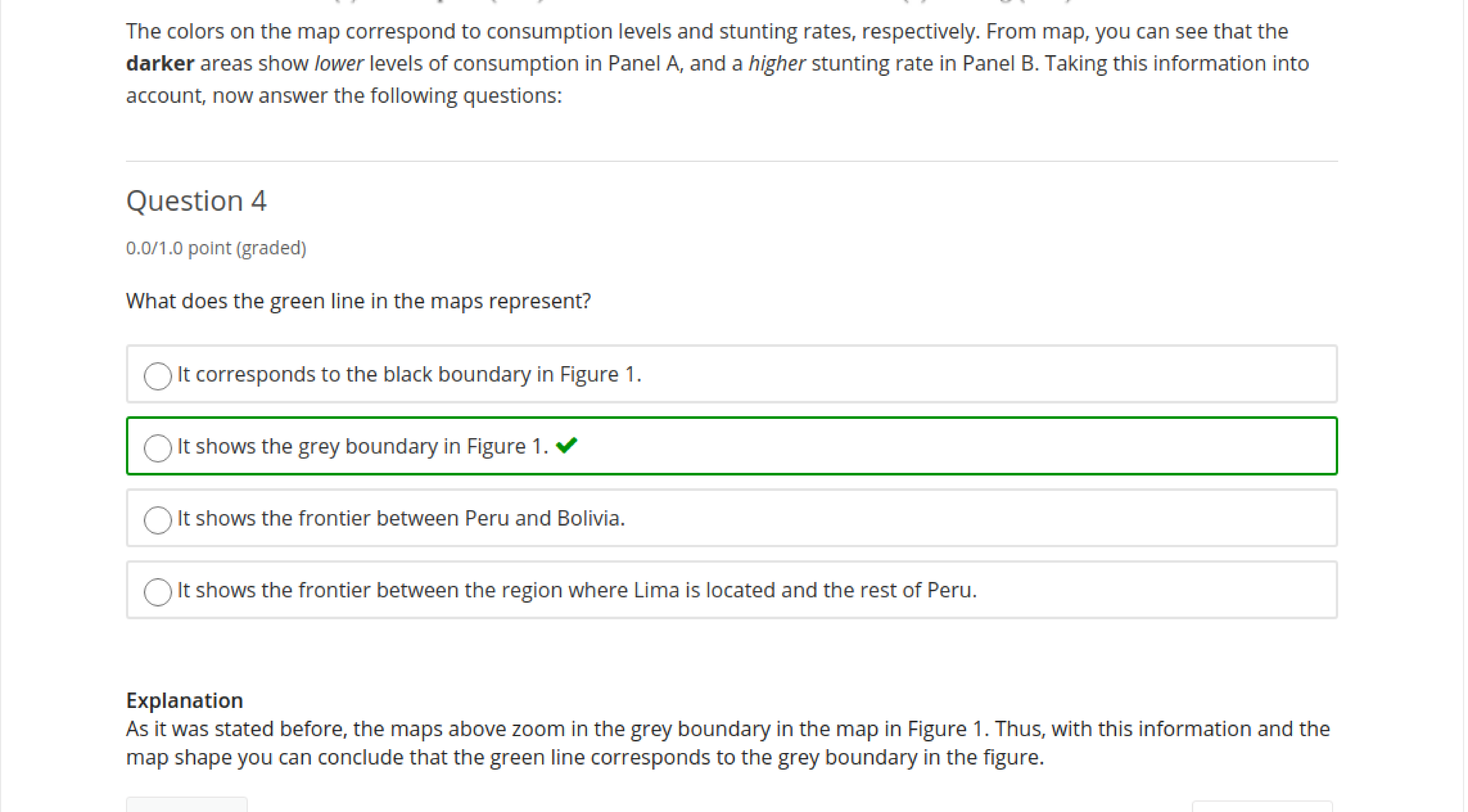
Continuing with Dell's research, she looks at the way in which more recent welfare variables look like in areas where the mita took place versus areas where it did not. The figure below shows a map zooming across the grey boundary: Panel A presents consumption levels in 2001, and Panel B the stunting rate in 2005. Take a look and some time to understand the maps and compare them to the one shown for Questions 1-3 (Figure 1).

Figure 2:



(a) Consumption (2001)

(b) Stunting (2005)



The colors on the map correspond to consumption levels and stunting rates, respectively. From map, you can see that the **darker** areas show *lower* levels of consumption in Panel A, and a *higher* stunting rate in Panel B. Taking this information into account, now answer the following questions:

Question 4

0.0/1.0 point (graded)

What does the green line in the maps represent?

○ It corresponds to the black boundary in Figure 1.

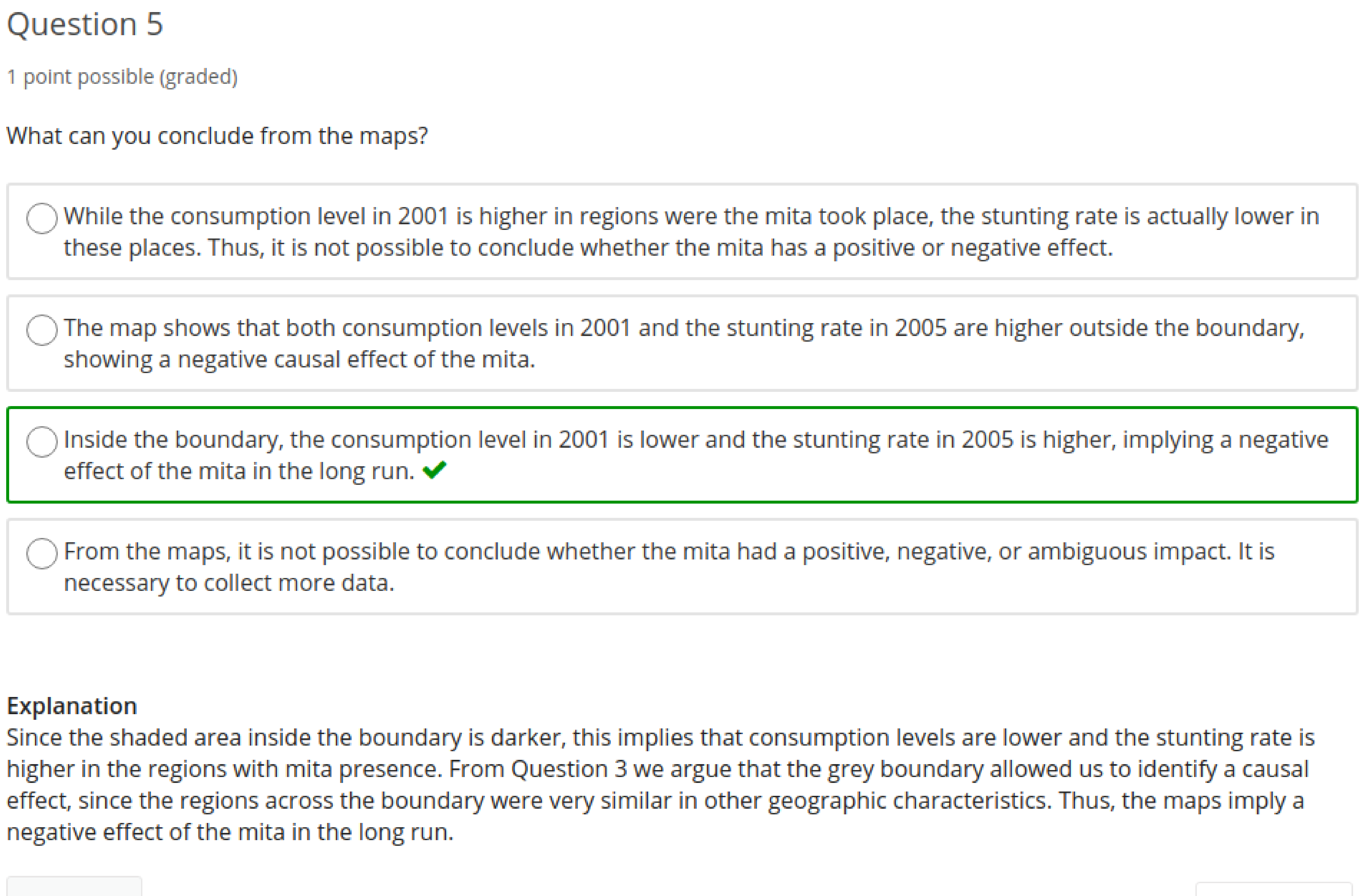
○ It shows the grey boundary in Figure 1. ✓

○ It shows the frontier between Peru and Bolivia.

○ It shows the frontier between the region where Lima is located and the rest of Peru.

Explanation

As it was stated before, the maps above zoom in the grey boundary in the map in Figure 1. Thus, with this information and the map shape you can conclude that the green line corresponds to the grey boundary in the figure.



Question 5

1 point possible (graded)

What can you conclude from the maps?

○ While the consumption level in 2001 is higher in regions were the mita took place, the stunting rate is actually lower in these places. Thus, it is not possible to conclude whether the mita has a positive or negative effect.

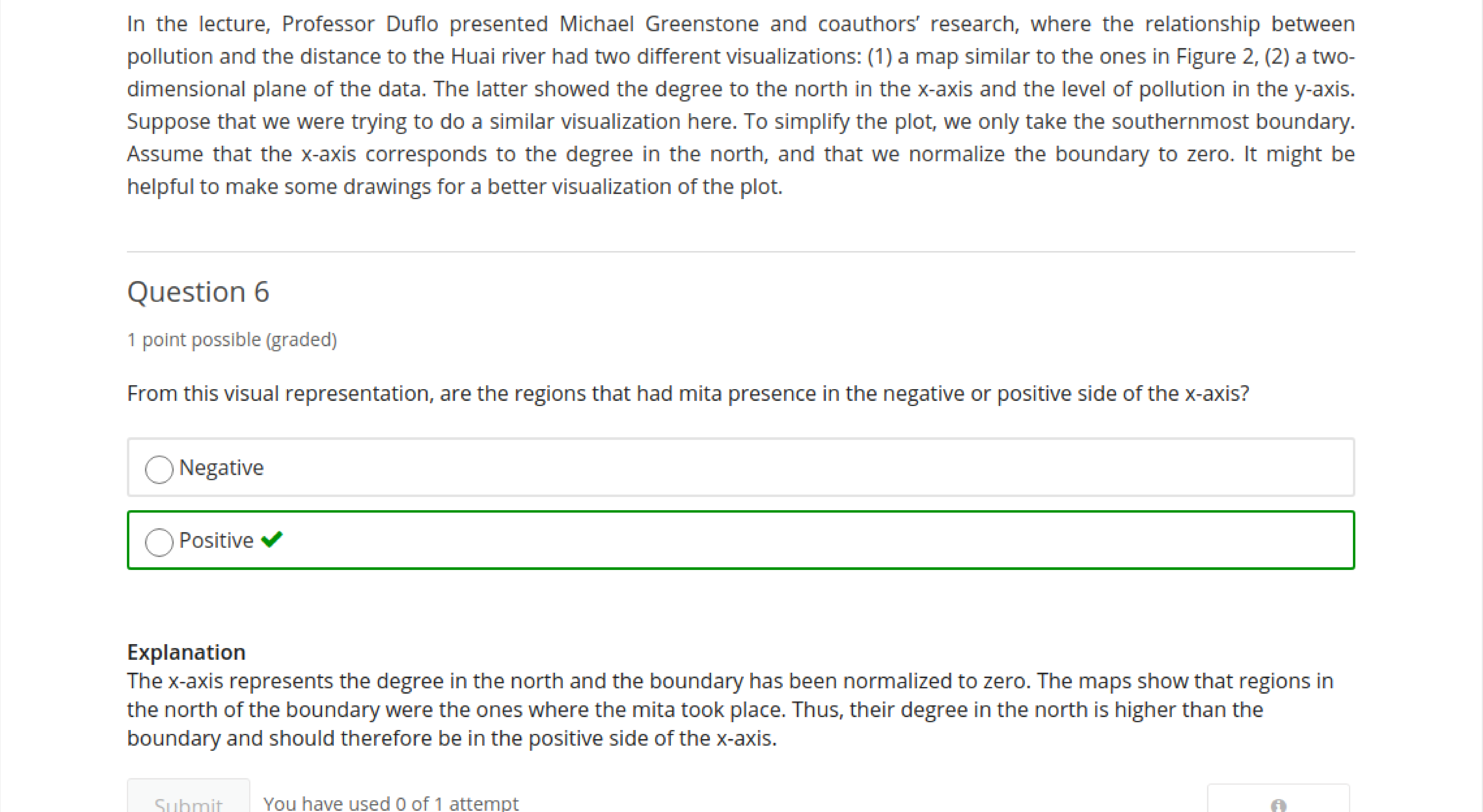
○ The map shows that both consumption levels in 2001 and the stunting rate in 2005 are higher outside the boundary, showing a negative causal effect of the mita.

○ Inside the boundary, the consumption level in 2001 is lower and the stunting rate in 2005 is higher, implying a negative effect of the mita in the long run. ✓

○ From the maps, it is not possible to conclude whether the mita had a positive, negative, or ambiguous impact. It is necessary to collect more data.

Explanation

Since the shaded area inside the boundary is darker, this implies that consumption levels are lower and the stunting rate is higher in the regions with mita presence. From Question 3 we argue that the grey boundary allowed us to identify a causal effect, since the regions across the boundary were very similar in other geographic characteristics. Thus, the maps imply a negative effect of the mita in the long run.



In the lecture, Professor Duflo presented Michael Greenstone and coauthors' research, where the relationship between pollution and the distance to the Huai river had two different visualizations: (1) a map similar to the ones in Figure 2, (2) a two-dimensional plane of the data. The latter showed the degree to the north in the x-axis and the level of pollution in the y-axis. Suppose that we were trying to do a similar visualization here. To simplify the plot, we only take the southernmost boundary. Assume that the x-axis corresponds to the degree in the north, and that we normalize the boundary to zero. It might be helpful to make some drawings for a better visualization of the plot.

Question 6

1 point possible (graded)

From this visual representation, are the regions that had mita presence in the negative or positive side of the x-axis?

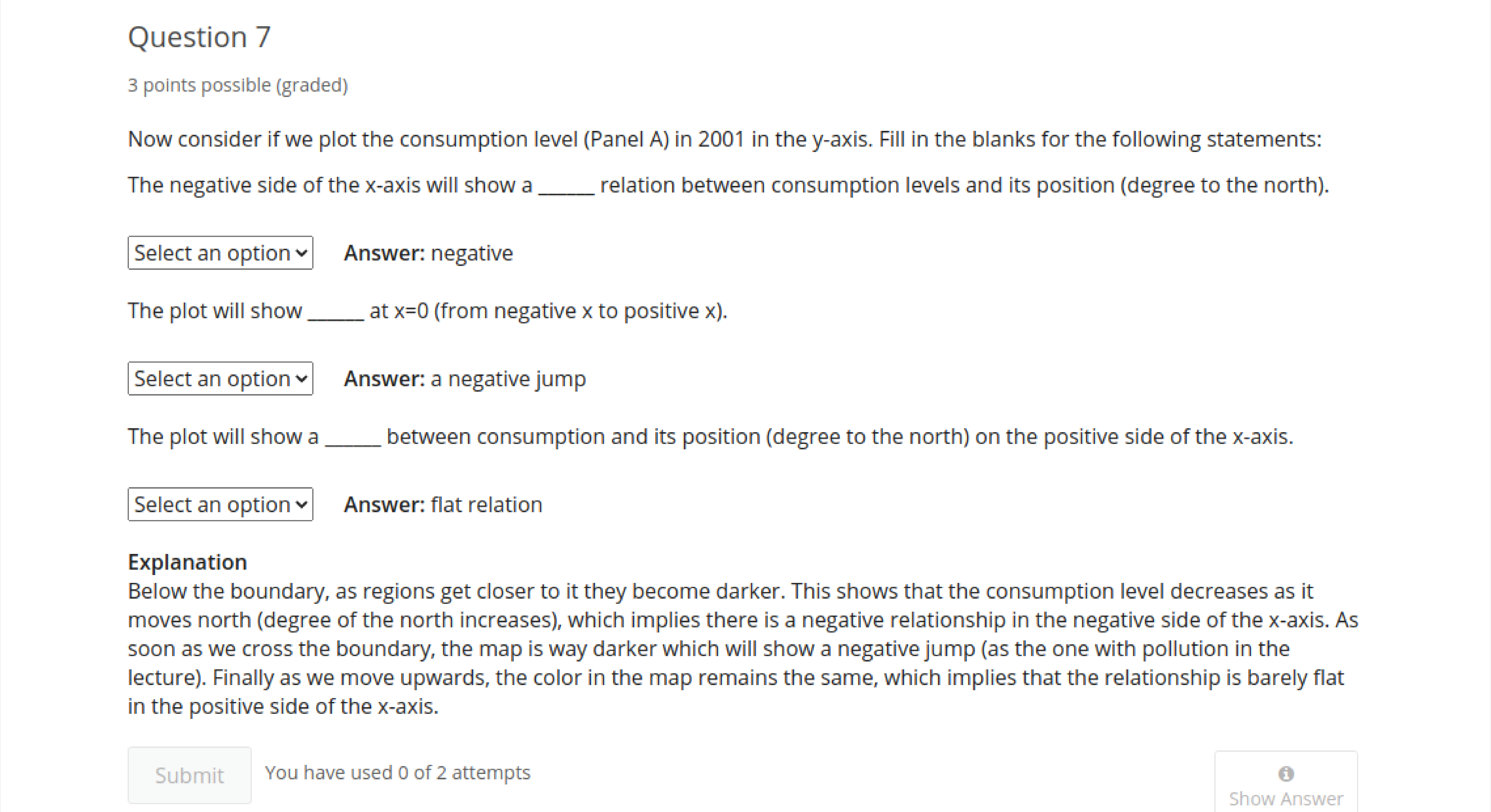
○ Negative

○ Positive ✓

Explanation

The x-axis represents the degree in the north and the boundary has been normalized to zero. The maps show that regions in the north of the boundary were the ones where the mita took place. Thus, their degree in the north is higher than the boundary and should therefore be in the positive side of the x-axis.

You have used 0 of 1 attempt



Question 7

3 points possible (graded)

Now consider if we plot the consumption level (Panel A) in 2001 in the y-axis. Fill in the blanks for the following statements:

The negative side of the x-axis will show a _____ relation between consumption levels and its position (degree to the north).

Answer: negative

The plot will show _____ at x=0 (from negative x to positive x).

Answer: a negative jump

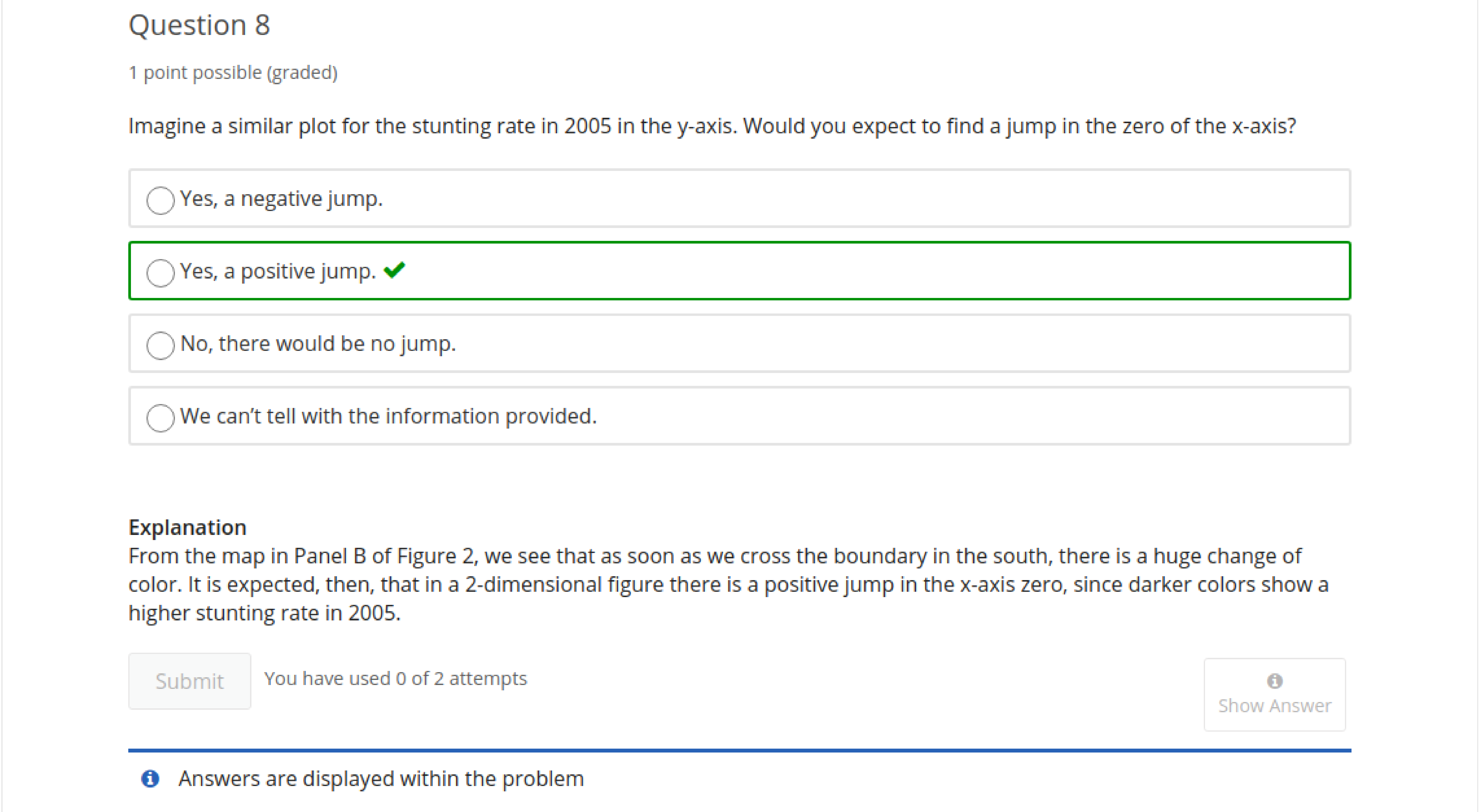
The plot will show a _____ between consumption and its position (degree to the north) on the positive side of the x-axis.

Answer: flat relation

Explanation

Below the boundary, as regions get closer to it they become darker. This shows that the consumption level decreases as it moves north (degree of the north increases), which implies there is a negative relationship in the negative side of the x-axis. As soon as we cross the boundary, the map is way darker which will show a negative jump (as the one with pollution in the lecture). Finally as we move upwards, the color in the map remains the same, which implies that the relationship is barely flat in the positive side of the x-axis.

You have used 0 of 2 attempts



Question 8

1 point possible (graded)

Imagine a similar plot for the stunting rate in 2005 in the y-axis. Would you expect to find a jump in the zero of the x-axis?

○ Yes, a negative jump.

○ Yes, a positive jump. ✓

○ No, there would be no jump.

○ We can't tell with the information provided.

Explanation

From the map in Panel B of Figure 2, we see that as soon as we cross the boundary in the south, there is a huge change of color. It is expected, then, that in a 2-dimensional figure there is a positive jump in the x-axis zero, since darker colors show a higher stunting rate in 2005.

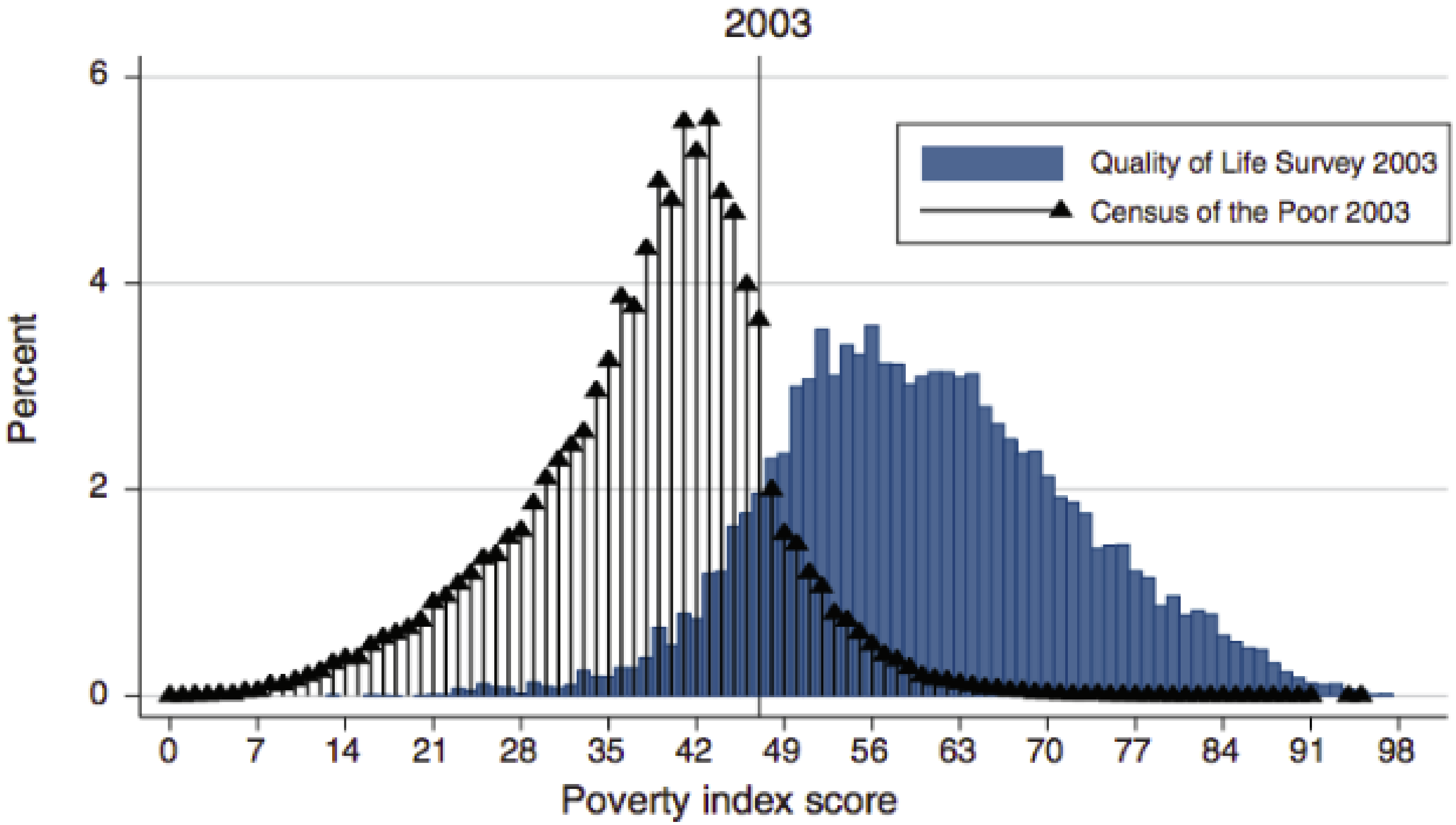
You have used 0 of 2 attempts

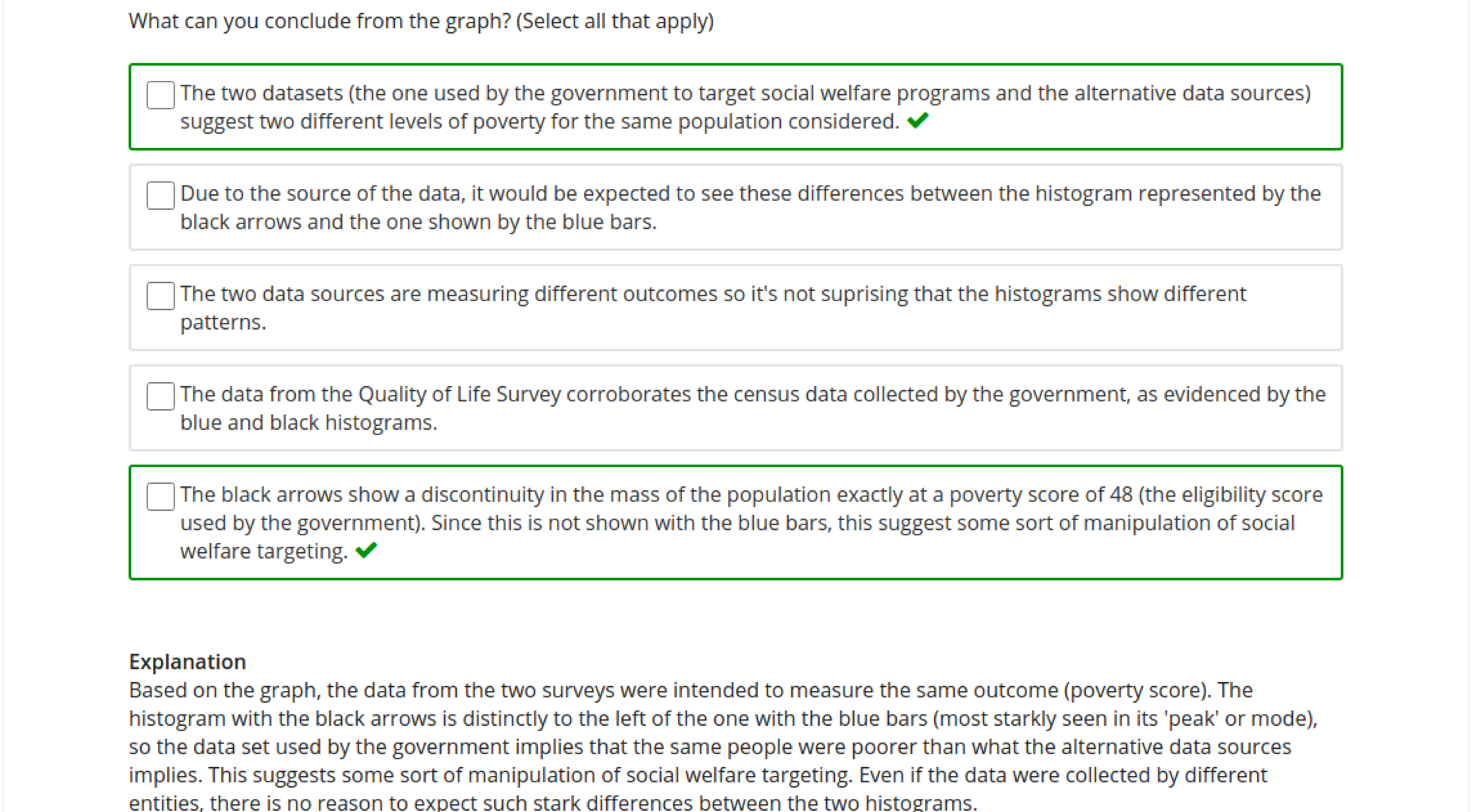
Answers are displayed within the problem

Homework due Sep 14, 2020 19:30 EDT **Past Due**

Camacho & Conover (2011) document manipulation of a targeting system for social welfare programs in Colombia. Take a look at the following figure, which shows two histograms: the black arrows present the histogram for a poverty score (lower numbers mean being poorer) that was calculated using the same data the government collected to target social welfare programs – where only individuals with a poverty score below 48 were eligible to receive most of these programs. The blue bars correspond to the histogram reconstructing this poverty score using other data sources that were not used by the government for this purpose.

Figure 3





What can you conclude from the graph? (Select all that apply)

□ The two datasets (the one used by the government to target social welfare programs and the alternative data sources) suggest two different levels of poverty for the same population considered. ✓

□ Due to the source of the data, it would be expected to see these differences between the histogram represented by the black arrows and the one shown by the blue bars.

□ The two data sources are measuring different outcomes so it's not suprising that the histograms show different patterns.

□ The data from the Quality of Life Survey corroborates the census data collected by the government, as evidenced by the blue and black histograms.

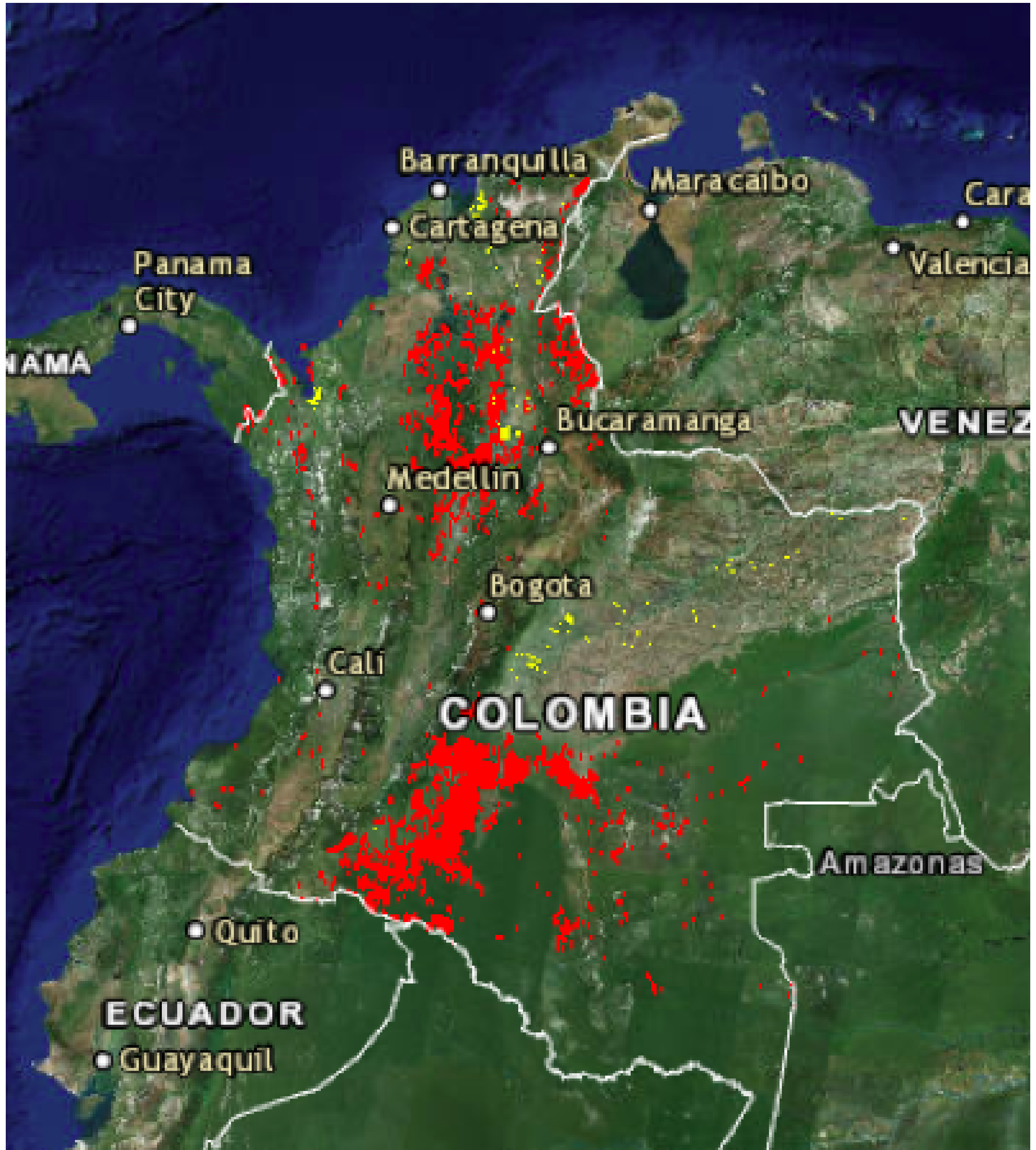
□ The black arrows show a discontinuity in the mass of the population exactly at a poverty score of 48 (the eligibility score used by the government). Since this is not shown with the blue bars, this suggest some sort of manipulation of social welfare targeting. ✓

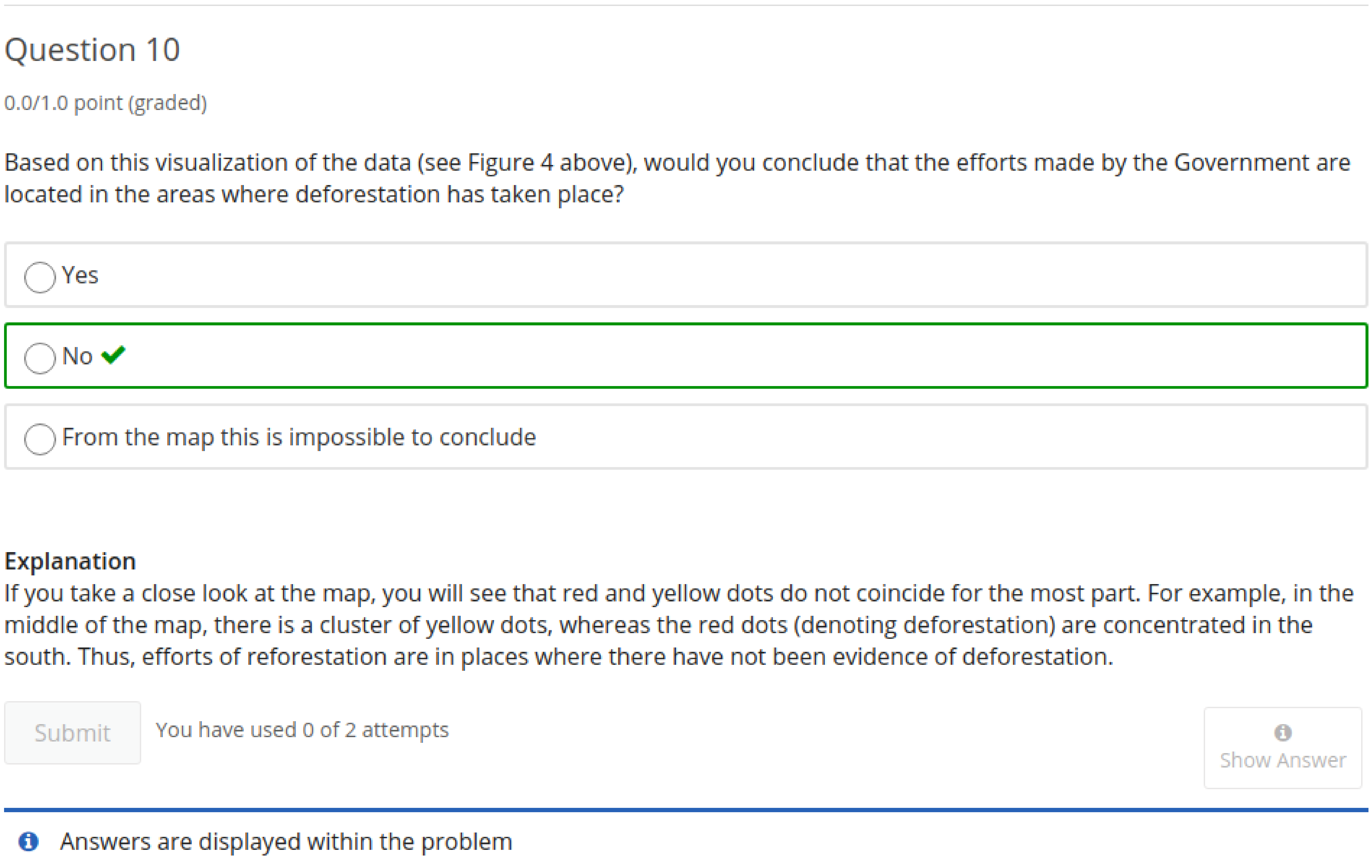
Explanation

Based on the graph, the data from the two surveys were intended to measure the same outcome (poverty score). The histogram with the black arrows is distinctly to the left of the one with the blue bars (most starkly seen in its 'peak' or mode), so the data set used by the government implies that the same people were poorer than what the alternative data sources implies. This suggests some sort of manipulation of social welfare targeting. Even if the data were collected by different entities, there is no reason to expect such stark differences between the two histograms.

Continuing with Colombia, *www.laramaciudadana.com* is a blog that publishes quantitative information about different topics of national interest. Their objective is to inform public policy debate by collecting data on these controversial topics and displaying it to a general audience. Their most recent project uses satellite photos to map deforestation and evaluate industrial reforestation efforts in the country. The map is presented in Figure 4: the red dots show the locations where satellites previously detected deforestation activities, and the yellow dots give an overview of the Government's responding industrial reforestation efforts. Take a close look at the map.

Figure 4





Question 10

0.0/1.0 point (graded)

Based on this visualization of the data (see Figure 4 above), would you conclude that the efforts made by the Government are located in the areas where deforestation has taken place?

○ Yes

○ No ✓

○ From the map this is impossible to conclude

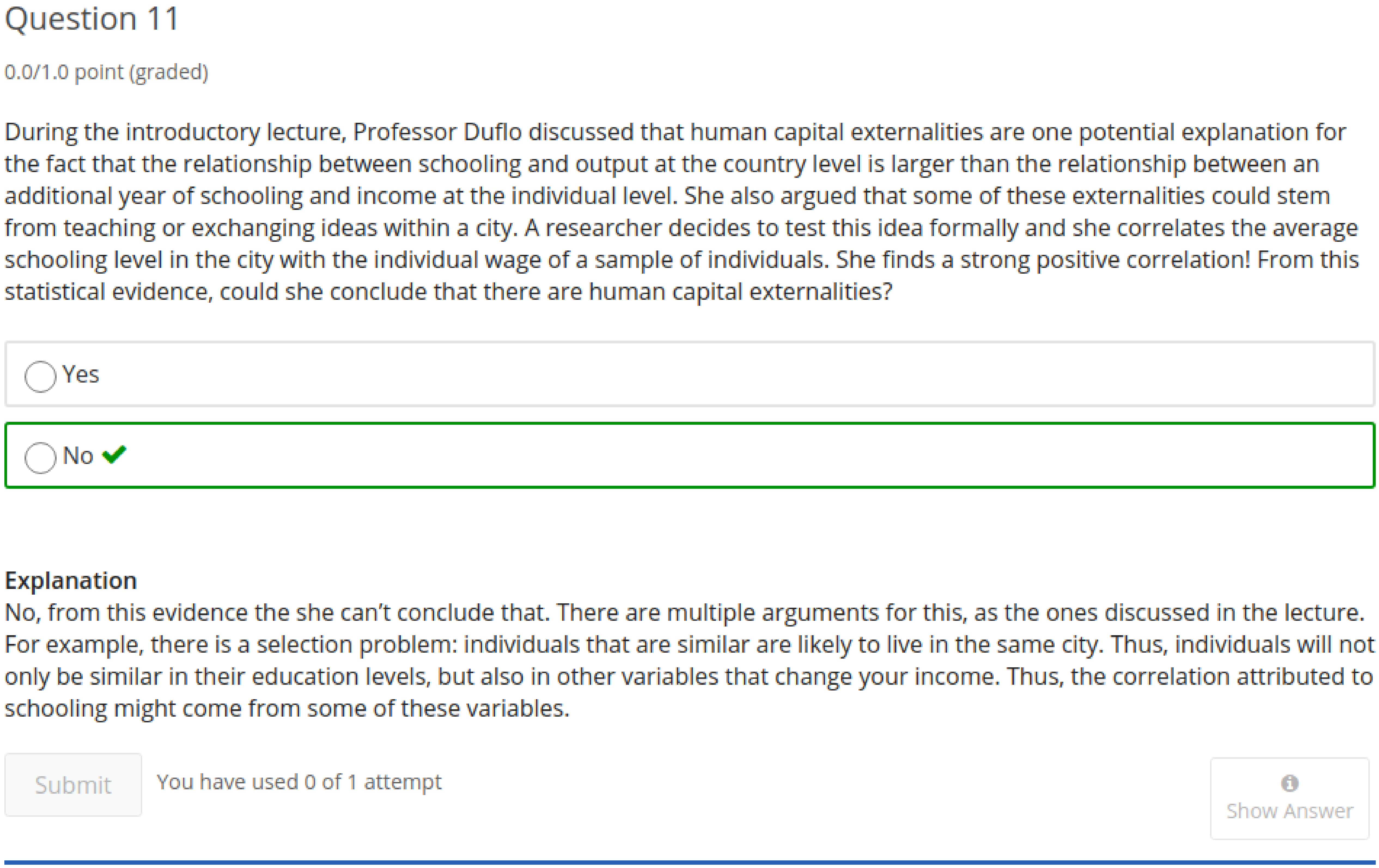
Explanation

If you take a close look at the map, you will see that red and yellow dots do not coincide for the most part. For example, in the middle of the map, there is a cluster of yellow dots, whereas the red dots (denoting deforestation) are concentrated in the south. Thus, efforts of reforestation are in places where there have not been evidence of deforestation.

Submit You have used 0 of 2 attempts

Show Answer

Answers are displayed within the problem



Question 11

0.0/1.0 point (graded)

During the introductory lecture, Professor Duflo discussed that human capital externalities are one potential explanation for the fact that the relationship between schooling and output at the country level is larger than the relationship between an additional year of schooling and income at the individual level. She also argued that some of these externalities could stem from teaching or exchanging ideas within a city. A researcher decides to test this idea formally and she correlates the average schooling level in the city with the individual wage of a sample of individuals. She finds a strong positive correlation! From this statistical evidence, could she conclude that there are human capital externalities?

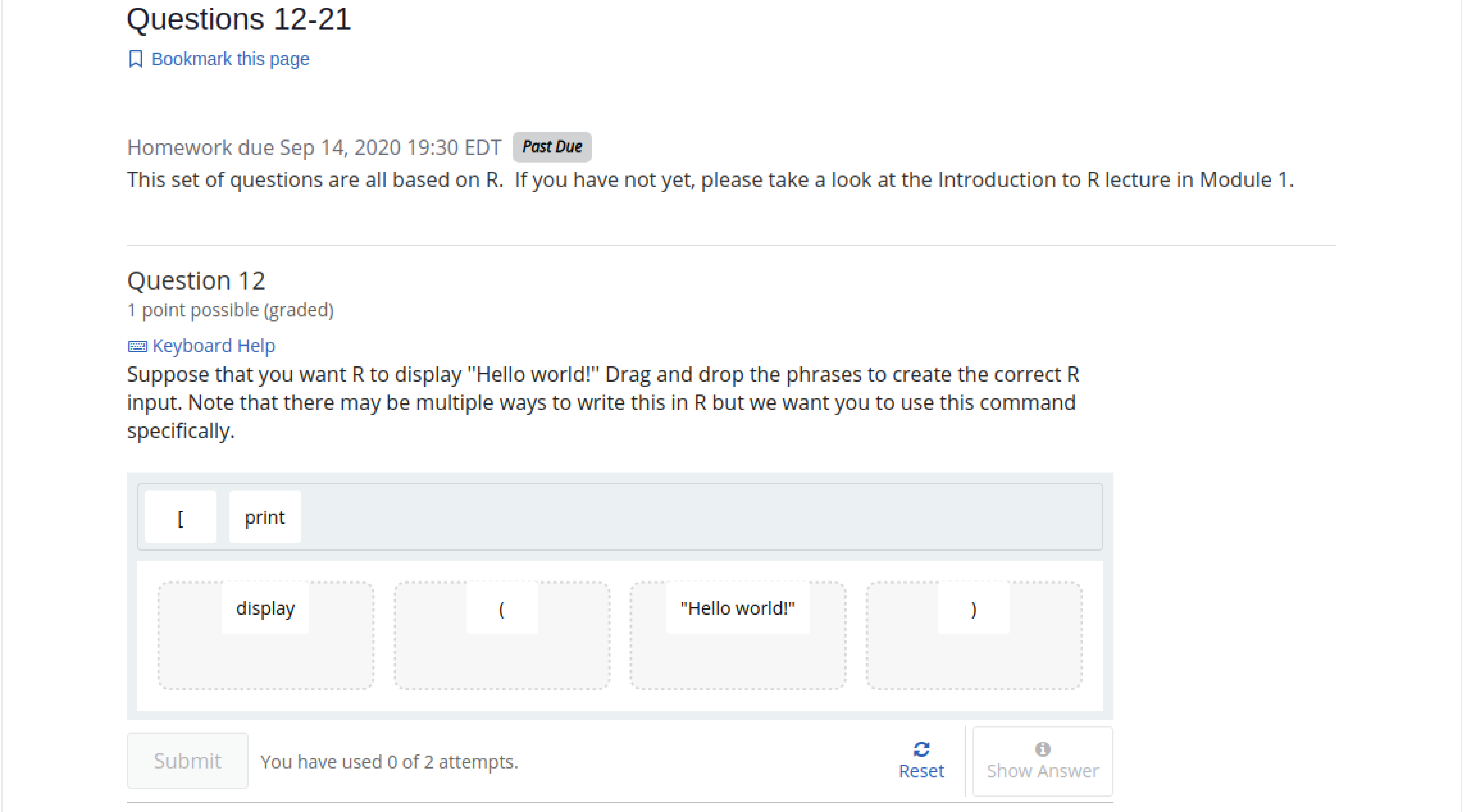
○ Yes

○ No ✓

Explanation

No, from this evidence the she can’t conclude that. There are multiple arguments for this, as the ones discussed in the lecture. For example, there is a selection problem: individuals that are similar are likely to live in the same city. Thus, individuals will not only be similar in their education levels, but also in other variables that change your income. Thus, the correlation attributed to schooling might come from some of these variables.

You have used 0 of 1 attempt



Questions 12-21

Bookmark this page

Homework due Sep 14, 2020 19:30 EDT *Past Due*

This set of questions are all based on R. If you have not yet, please take a look at the Introduction to R lecture in Module 1.

Question 12

1 point possible (graded)

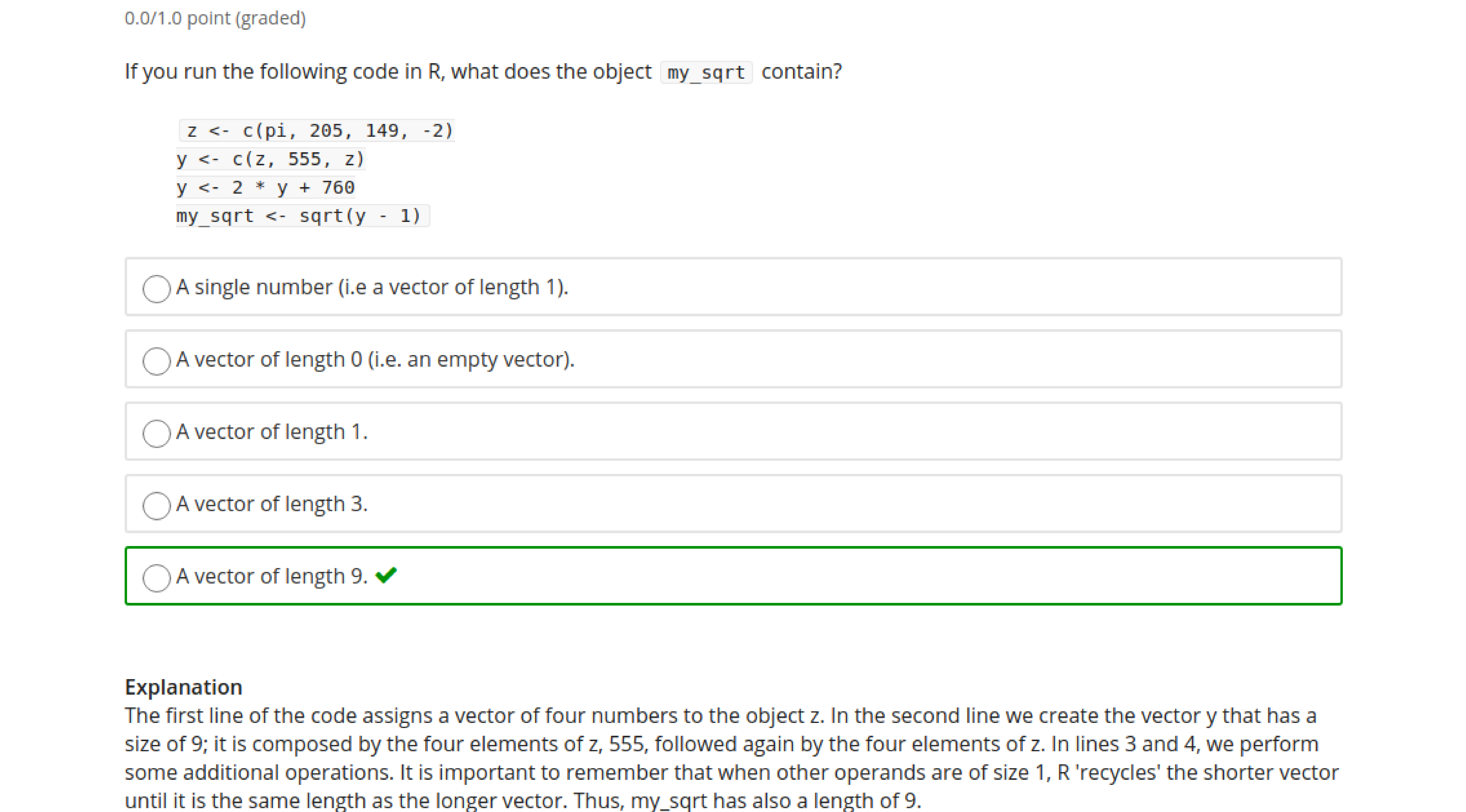
Keyboard Help

Suppose that you want R to display "Hello world!" Drag and drop the phrases to create the correct R input. Note that there may be multiple ways to write this in R but we want you to use this command specifically.

[print

display ("Hello world!")

Submit You have used 0 of 2 attempts.



0.0/1.0 point (graded)

If you run the following code in R, what does the object `my_sqrt` contain?

```
 z <- c(pi, 205, 149, -2)
y <- c(z, 555, z)
y <- 2 * y + 760
my_sqrt <- sqrt(y - 1)
```

○ A single number (i.e a vector of length 1).

○ A vector of length 0 (i.e. an empty vector).

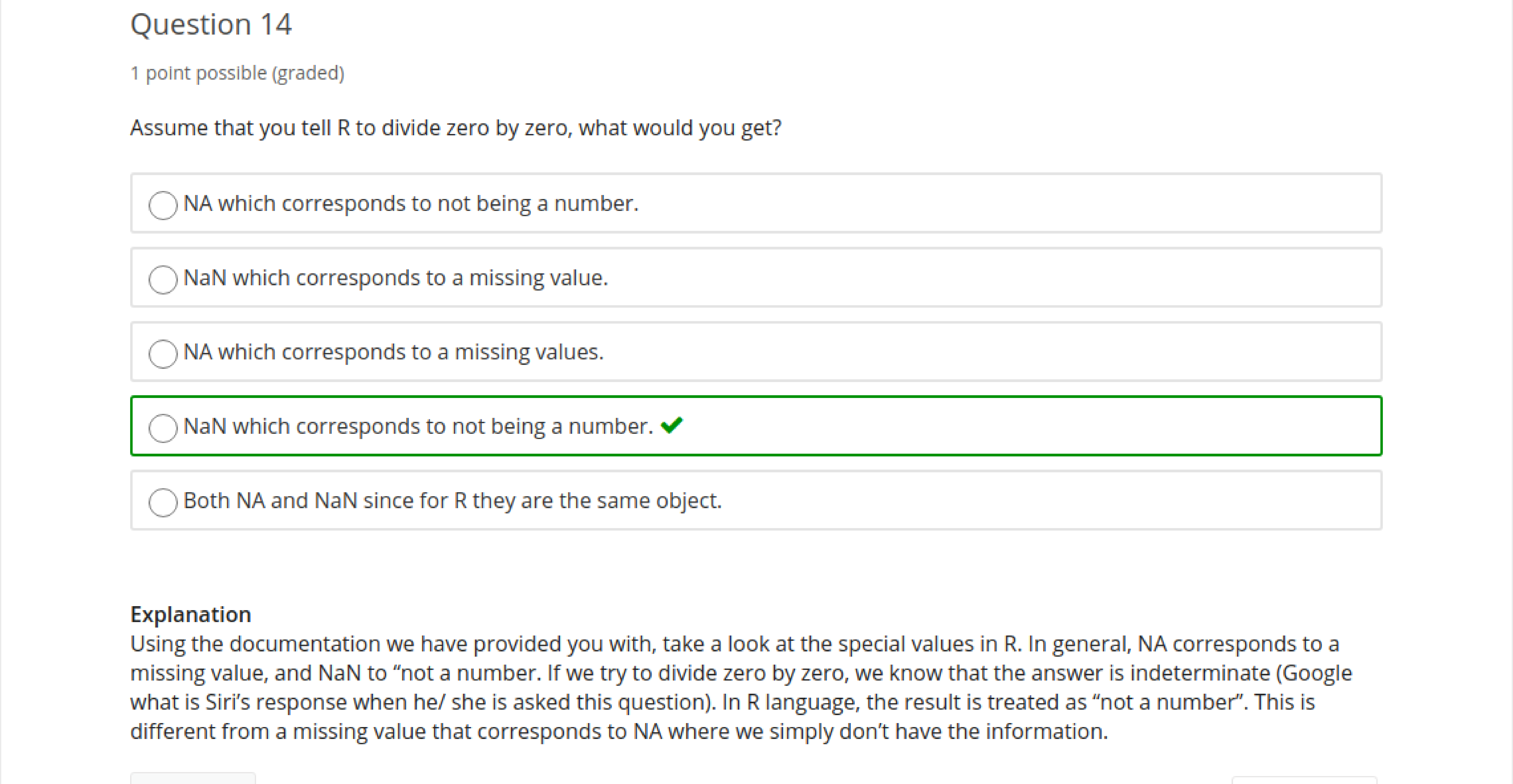
○ A vector of length 1.

○ A vector of length 3.

○ A vector of length 9. ✓

Explanation

The first line of the code assigns a vector of four numbers to the object z. In the second line we create the vector y that has a size of 9; it is composed by the four elements of z, 555, followed again by the four elements of z. In lines 3 and 4, we perform some additional operations. It is important to remember that when other operands are of size 1, R 'recycles' the shorter vector until it is the same length as the longer vector. Thus, my_sqrt has also a length of 9.



Question 14

1 point possible (graded)

Assume that you tell R to divide zero by zero, what would you get?

○ NA which corresponds to not being a number.

○ NaN which corresponds to a missing value.

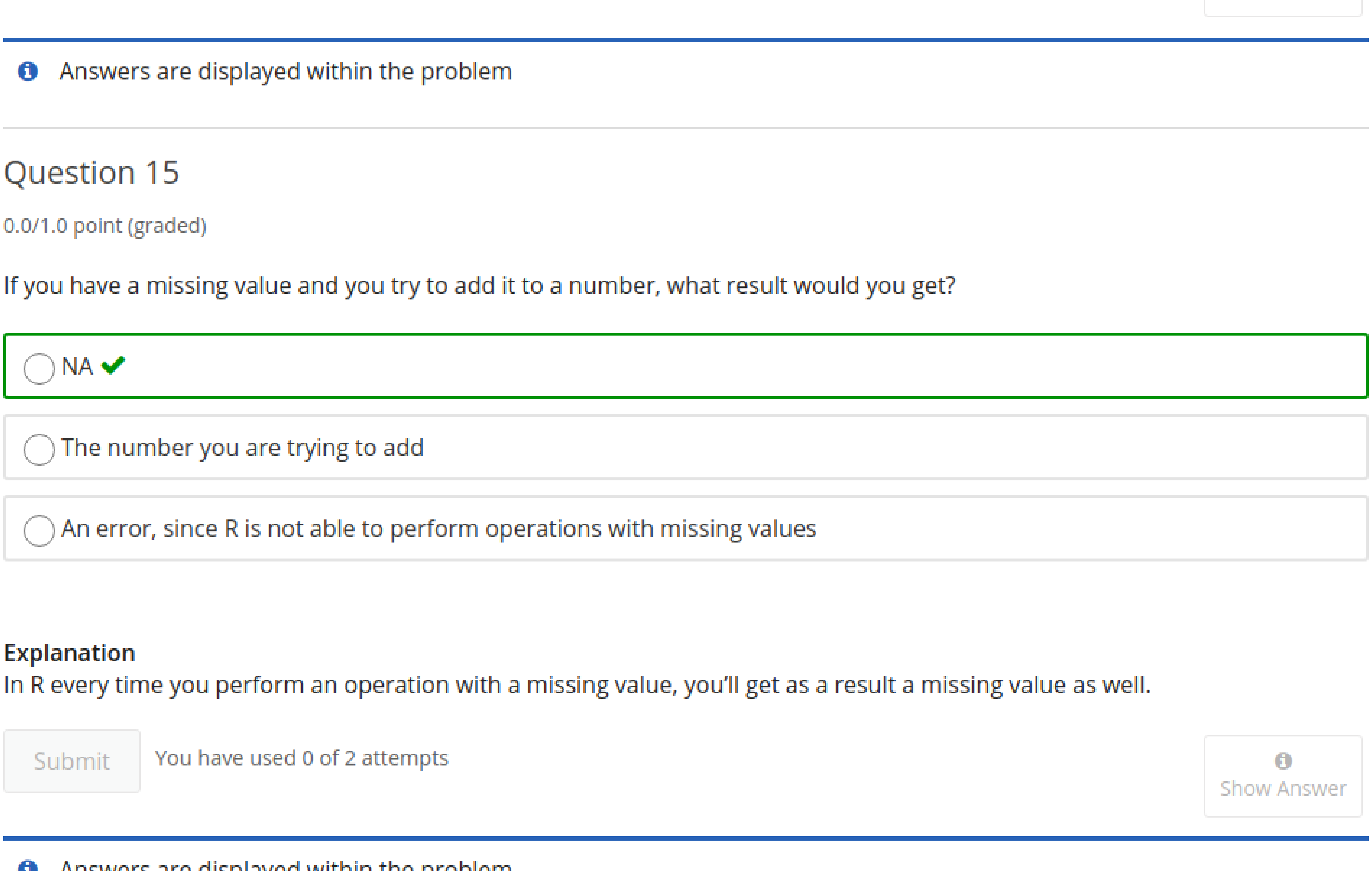
○ NA which corresponds to a missing values.

○ NaN which corresponds to not being a number. ✓

○ Both NA and NaN since for R they are the same object.

Explanation

Using the documentation we have provided you with, take a look at the special values in R. In general, NA corresponds to a missing value, and NaN to “not a number. If we try to divide zero by zero, we know that the answer is indeterminate (Google what is Siri’s response when he/ she is asked this question). In R language, the result is treated as “not a number”. This is different from a missing value that corresponds to NA where we simply don’t have the information.



Answers are displayed within the problem

Question 15

0.0/1.0 point (graded)

If you have a missing value and you try to add it to a number, what result would you get?

○ NA ✓

○ The number you are trying to add

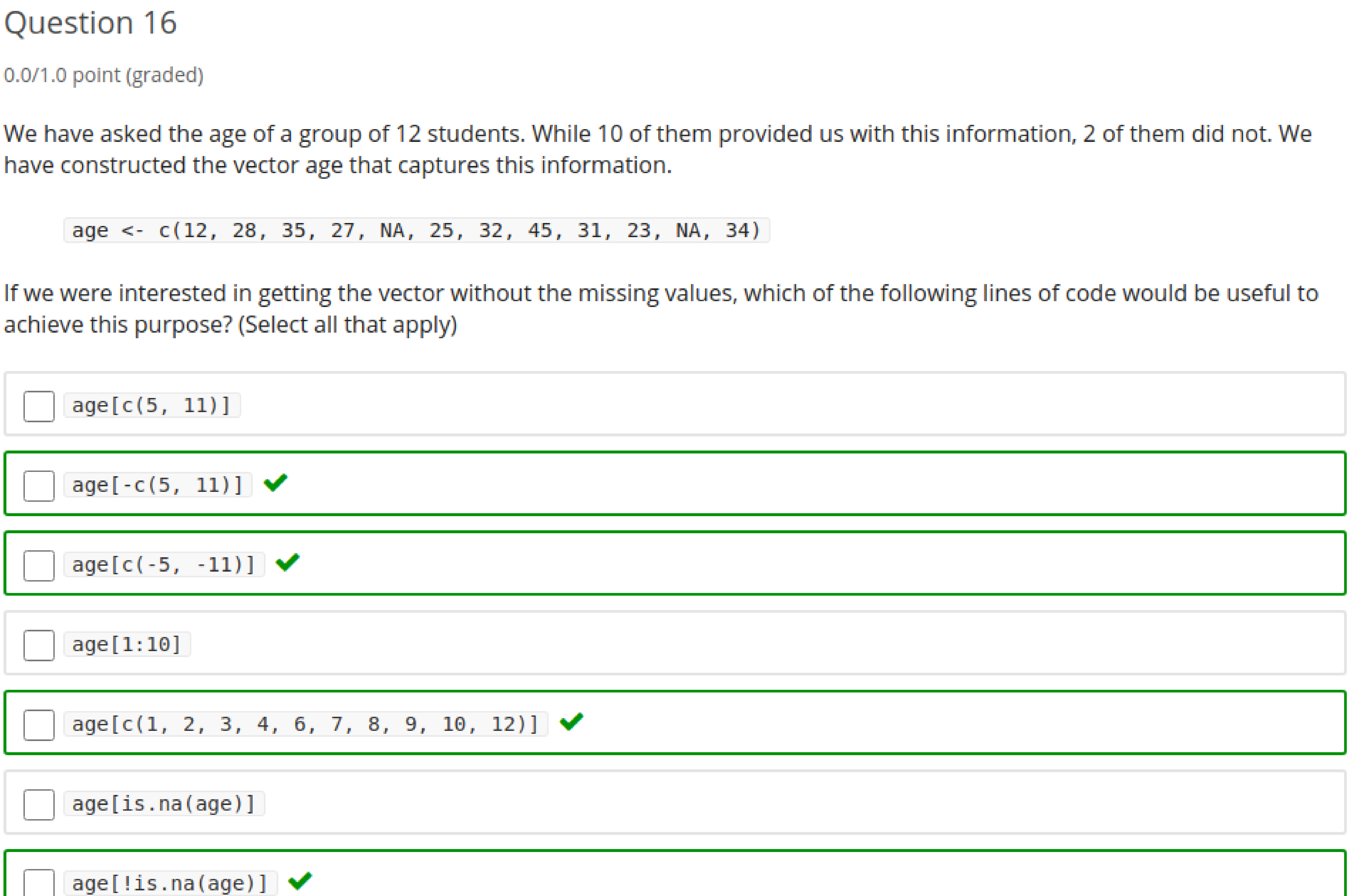
○ An error, since R is not able to perform operations with missing values

Explanation

In R every time you perform an operation with a missing value, you'll get as a result a missing value as well.

You have used 0 of 2 attempts

Answers are displayed within the problem



Question 16

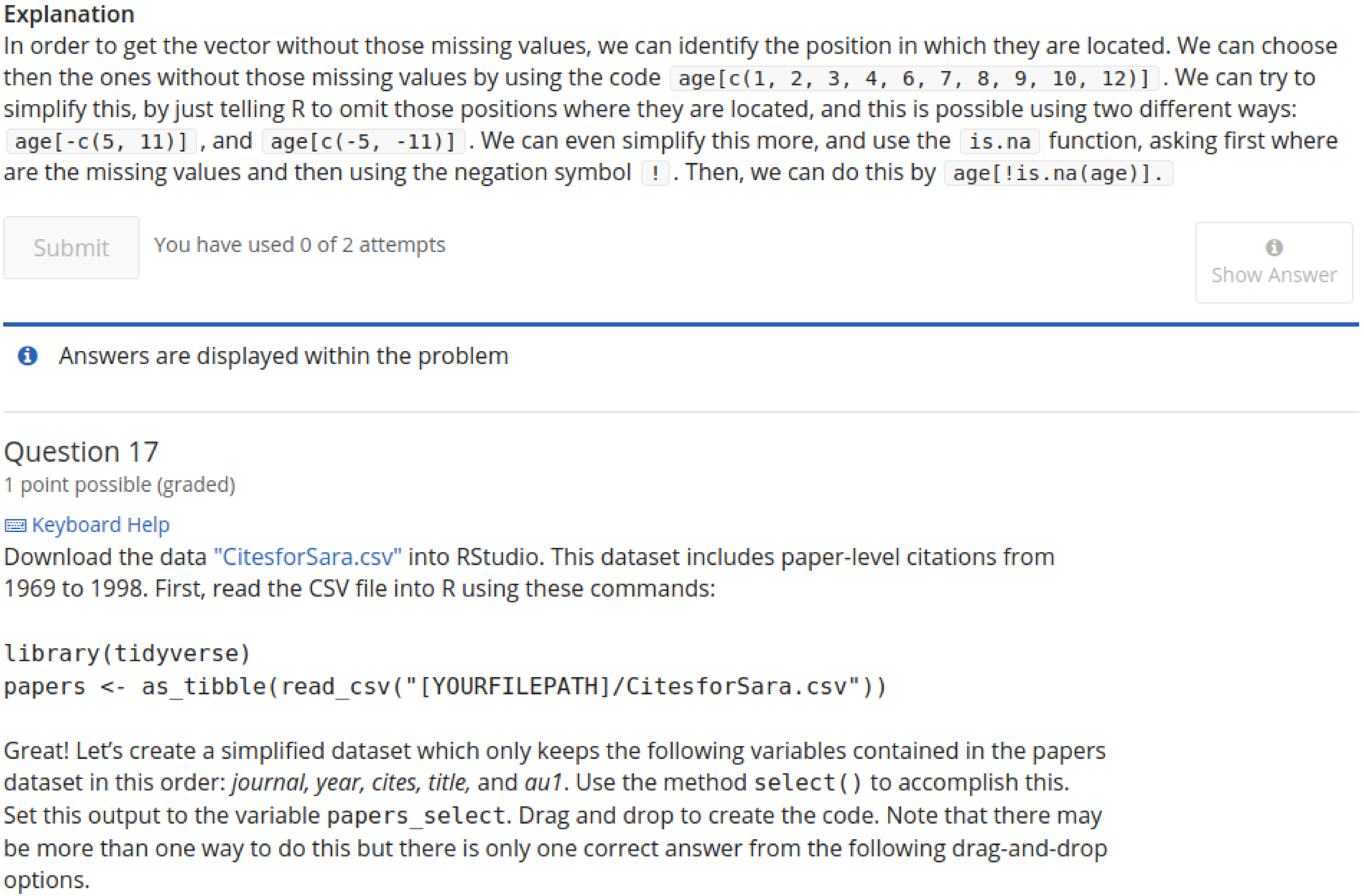
0.0/1.0 point (graded)

We have asked the age of a group of 12 students. While 10 of them provided us with this information, 2 of them did not. We have constructed the vector age that captures this information.

```
age <- c(12, 28, 35, 27, NA, 25, 32, 45, 31, 23, NA, 34)
```

If we were interested in getting the vector without the missing values, which of the following lines of code would be useful to achieve this purpose? (Select all that apply)

- ☐ `age[c(5, 11)]`
- ☐ `age[-c(5, 11)]` ✓
- ☐ `age[c(-5, -11)]` ✓
- ☐ `age[1:10]`
- ☐ `age[c(1, 2, 3, 4, 6, 7, 8, 9, 10, 12)]` ✓
- ☐ `age[is.na(age)]`
- ☐ `age[!is.na(age)]` ✓



Explanation

In order to get the vector without those missing values, we can identify the position in which they are located. We can choose then the ones without those missing values by using the code `age[c(1, 2, 3, 4, 6, 7, 8, 9, 10, 12)]` . We can try to simplify this, by just telling R to omit those positions where they are located, and this is possible using two different ways: `age[-c(5, 11)]` , and `age[c(-5, -11)]` . We can even simplify this more, and use the `is.na` function, asking first where are the missing values and then using the negation symbol `!` . Then, we can do this by `age[!is.na(age)].`

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

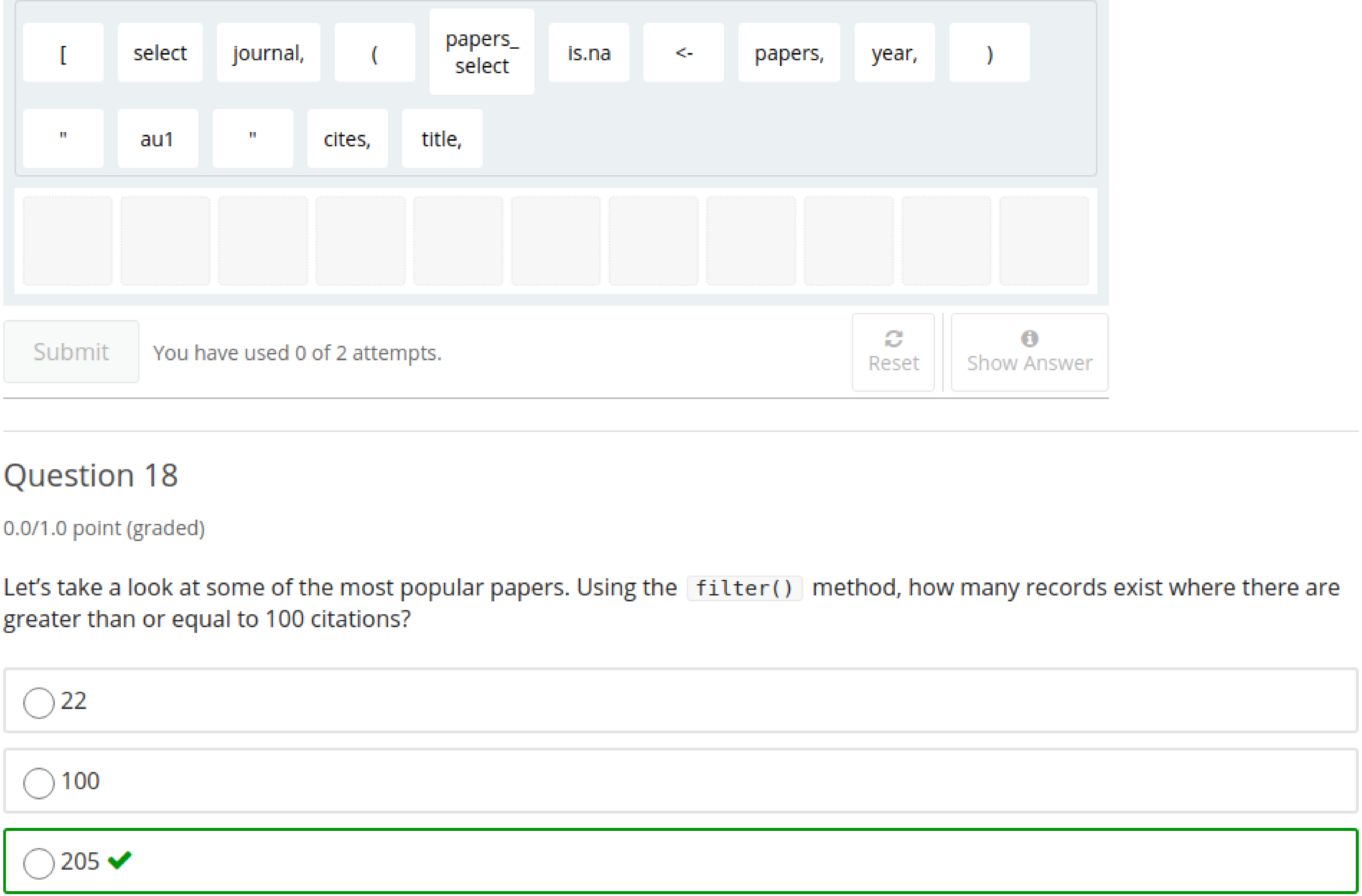
Question 17

1 point possible (graded)

Download the data "CitesforSara.csv" into RStudio. This dataset includes paper-level citations from 1969 to 1998. First, read the CSV file into R using these commands:

```
library(tidyverse)
papers <- as_tibble(read_csv("[YOURFILEPATH]/CitesforSara.csv"))
```

Great! Let's create a simplified dataset which only keeps the following variables contained in the papers dataset in this order: *journal, year, cites, title,* and *au1*. Use the method `select()` to accomplish this. Set this output to the variable `papers_select`. Drag and drop to create the code. Note that there may be more than one way to do this but there is only one correct answer from the following drag-and-drop options.



You have used 0 of 2 attempts.

Question 18

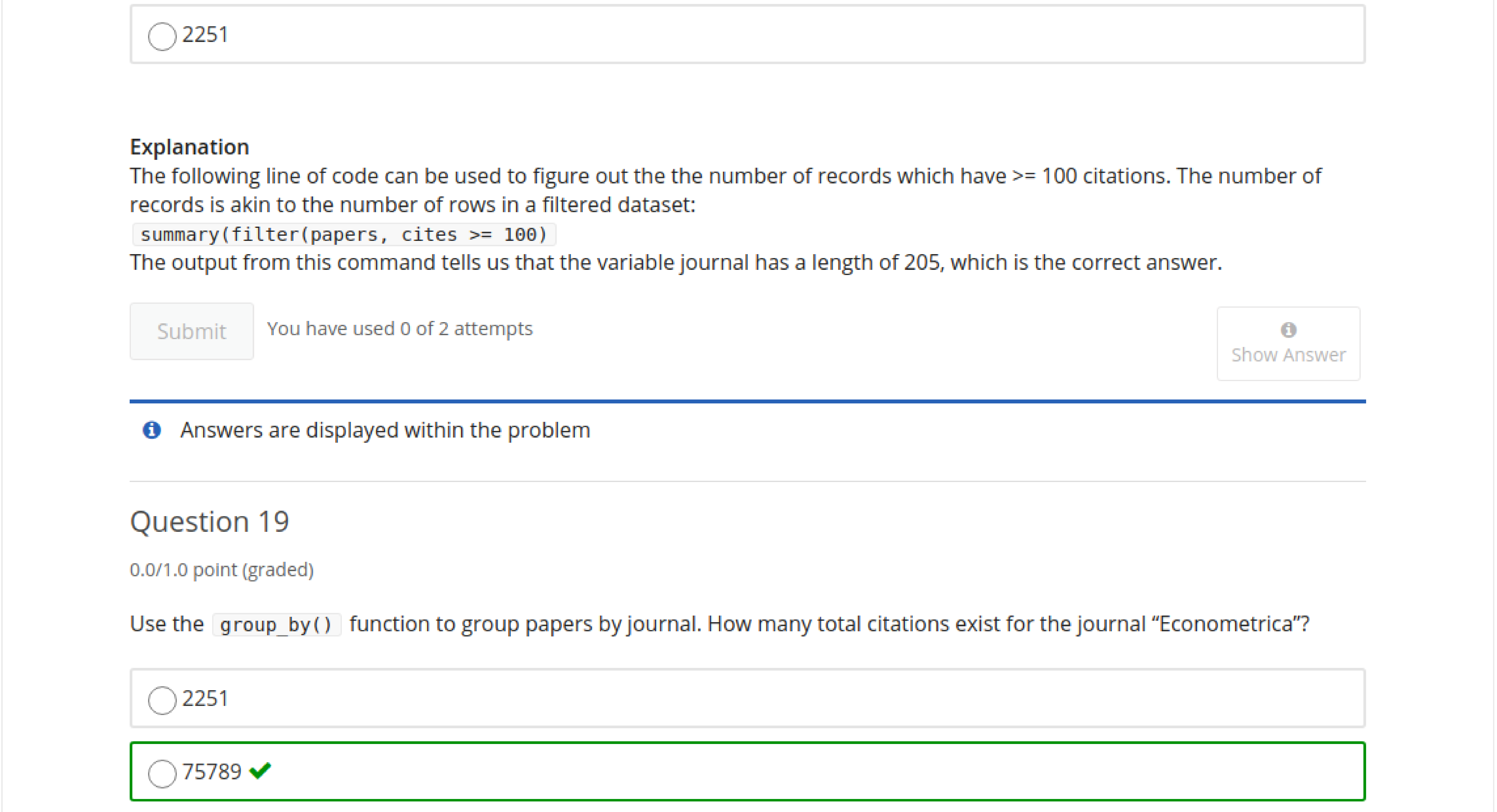
0.0/1.0 point (graded)

Let's take a look at some of the most popular papers. Using the `filter()` method, how many records exist where there are greater than or equal to 100 citations?

○ 22

○ 100

○ 205 ✓



○ 2251

Explanation

The following line of code can be used to figure out the the number of records which have >= 100 citations. The number of records is akin to the number of rows in a filtered dataset:

```
summary(filter(papers, cites >= 100)
```

The output from this command tells us that the variable journal has a length of 205, which is the correct answer.

You have used 0 of 2 attempts

Answers are displayed within the problem

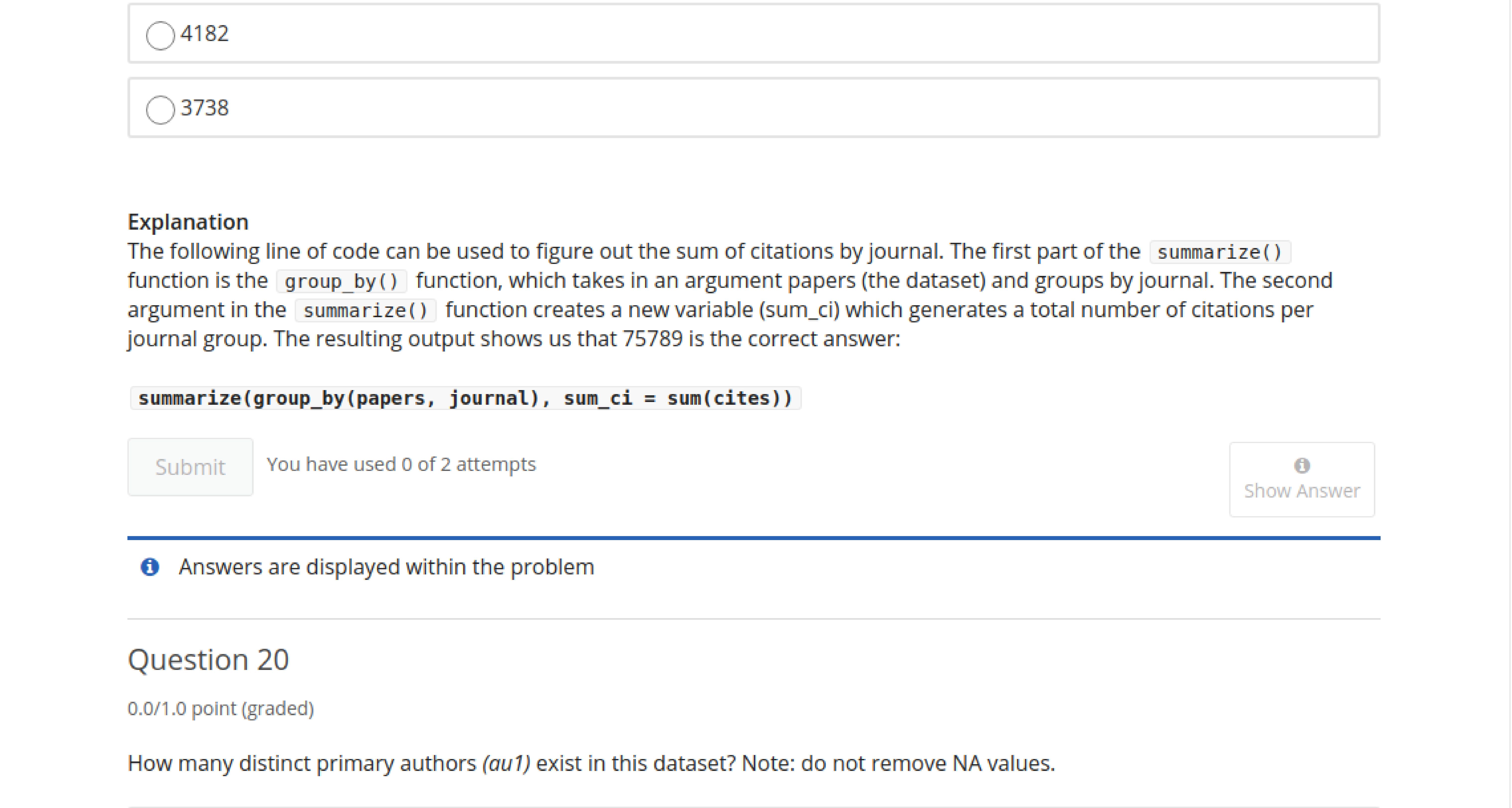
Question 19

0.0/1.0 point (graded)

Use the `group_by()` function to group papers by journal. How many total citations exist for the journal “Econometrica”?

○ 2251

○ 75789 ✓



○ 4182

○ 3738

Explanation

The following line of code can be used to figure out the sum of citations by journal. The first part of the `summarize()` function is the `group_by()` function, which takes in an argument papers (the dataset) and groups by journal. The second argument in the `summarize()` function creates a new variable (sum_ci) which generates a total number of citations per journal group. The resulting output shows us that 75789 is the correct answer:

```
summarize(group_by(papers, journal), sum_ci = sum(cites))
```

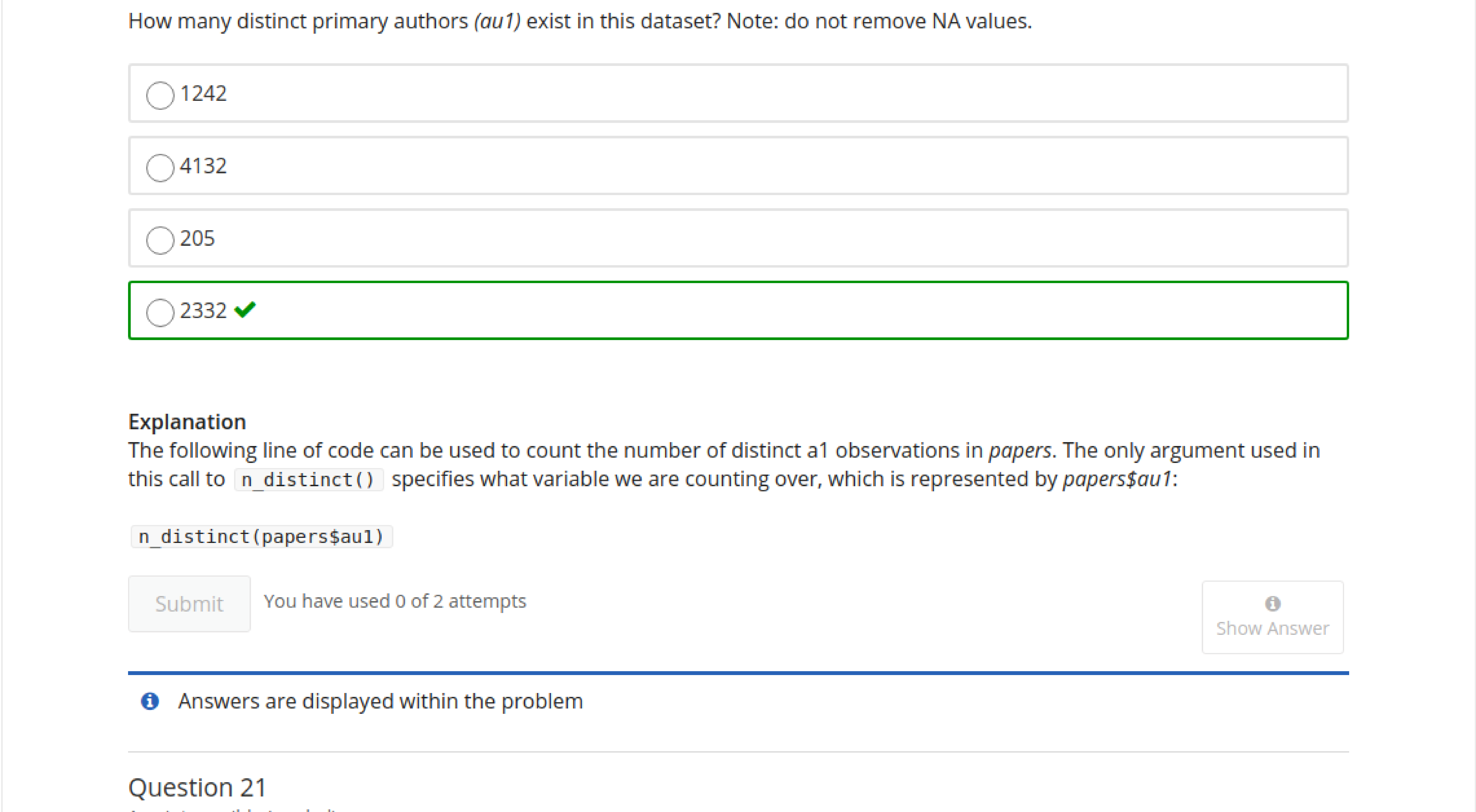
You have used 0 of 2 attempts

Answers are displayed within the problem

Question 20

0.0/1.0 point (graded)

How many distinct primary authors *(au1)* exist in this dataset? Note: do not remove NA values.



How many distinct primary authors *(au1)* exist in this dataset? Note: do not remove NA values.

○ 1242

○ 4132

○ 205

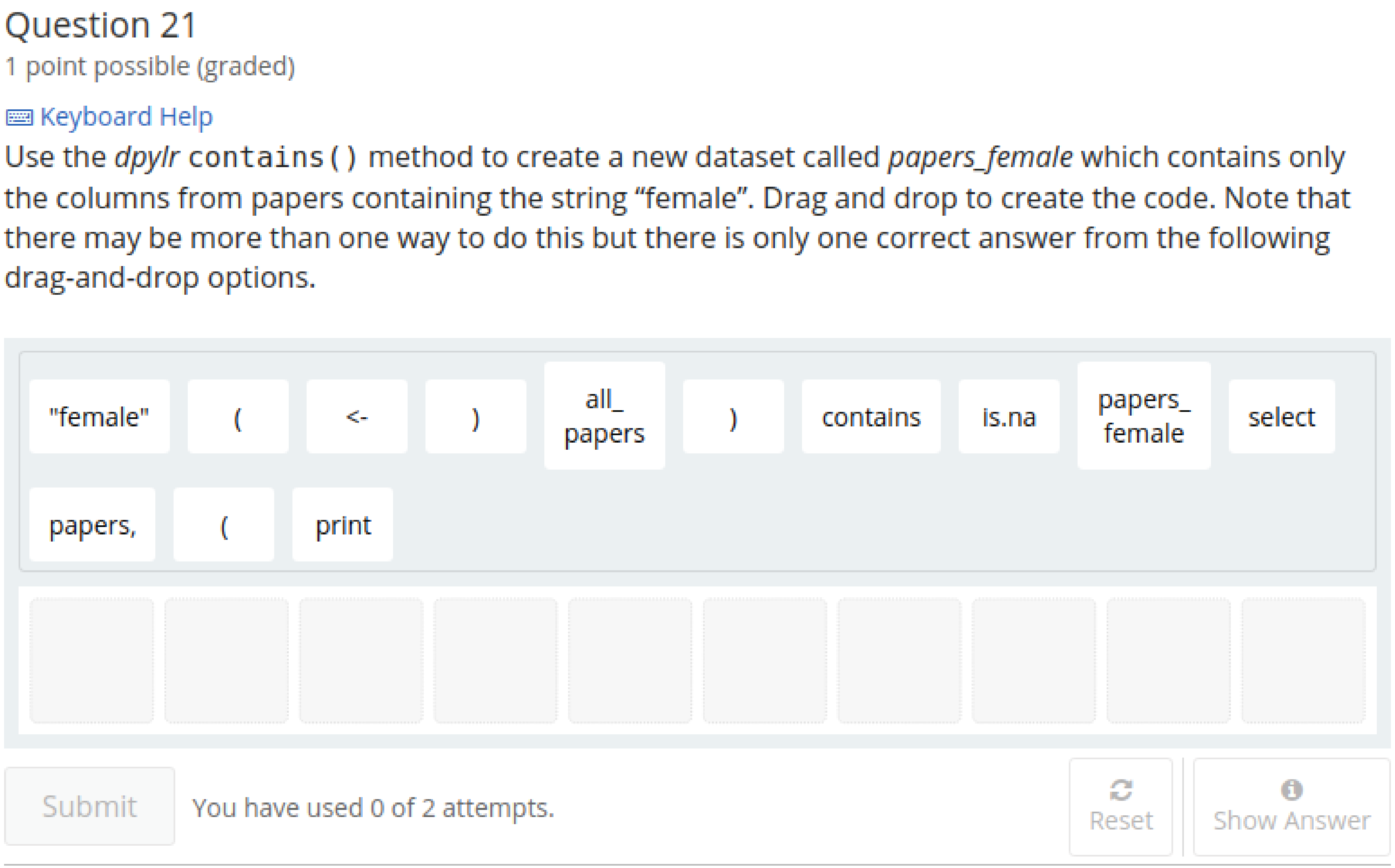
○ 2332 ✓

Explanation
The following line of code can be used to count the number of distinct a1 observations in *papers*. The only argument used in this call to `n_distinct()` specifies what variable we are counting over, which is represented by *papers$au1*:

```
n_distinct(papers$au1)
```

Submit
You have used 0 of 2 attempts
Show Answer
Answers are displayed within the problem

Question 21



Question 21

1 point possible (graded)

Keyboard Help

Use the *dpylr* `contains()` method to create a new dataset called *papers_female* which contains only the columns from papers containing the string “female”. Drag and drop to create the code. Note that there may be more than one way to do this but there is only one correct answer from the following drag-and-drop options.

You have used 0 of 2 attempts.