Scenario -1:

Question 1. What are the strengths and weaknesses of each option?

Ans:

Option 1:

Strength: In this option since we are able to maintain our desired grain without modifying the fact table or without adding new fact table it will save the database space.

Weakness: As long as, one course is taught by two Instructor this option seems feasible but if number of instructors grows per course let's say to 3 in that case we would need to add extra column to instructor dimension to keep combination of 3 instructors.

Option 2:

Weakness: In this approach size of fact table will grow , since we are changing the grain of the fact table and it will split the single row into multiple rows based on the number of instructor per course. Also whenever number of instructor goes above two we would need to split the enrollment_count into 0.33 and so on and so forth. So this approach is not feasible.

Option 3:

Weakness: In this approach as well since we are introducing the new fact table it will increase the database size and also when we want to build any report that requires the data from both the tables we would need to join the same which adds the extra join complexity and extra maintenance of new fact table.

Question 2. Which option would you choose and why?

Ans:

I would choose option 1 because it has following benefits over other options:

1) Would not increase the database size
2) No extra over head of maintaining the new fact tables
3) Not much changes needed in the business logic
4) No complex joining of two fact tables needed.

Question 3. Would your answer to Question 2 be different if the majority of classes had multiple instructors? How about if only one or two classes had multiple instructors? (Explain your answer.)

Ans: No. My answer remains same irrespective of the quantum of the classes having multiple instructors. Because as mentioned earlier it has lot benefits and very less weaknesses:

Benefits:

1) Would not increase the database size
2) No extra over head of maintaining the new fact tables
3) Not much changes needed in the business logic
4) No complex joining of two fact tables needed.

Scenario -2:

Question 5. What are the strengths and weaknesses of each option?

Ans:

Option 1:

Strength: This approach is easier to implement as we are not preserving the history. Also it saves the database space since we are not adding new table or not preserving the older data.

Weakness: In this option since we are using scd-1 i.e. overwrite the .old score, we will not be able to see the trend of movement of customers from one segment to another segment. In short we wont be able to track the history of segments for customers.

Option 2:

Strength: We will be able to track and see the historical segment rather their trends for all customers.

Weakness: It's complex to implement as compared to option 1 and dimension table size will grow as we are preserving the history.

Option 3:

Strength: This option looks goods as we are introducing one new dimension which will be static in nature and our fact will contain the foreign key to this dimension, so from fact we will be able to track the segment changes across customer. Also using historical data we will be able to answer why and how one customer got moved from one to another segment.

Option 4:

Weakness**:** Since using this option we are not able to track the history so we wont be able to answer as to why and how activity level or profitability level of some customers changed over the period of time.

Question 6. Which option would you choose and why?

Ans:

I would choose option 3 because it has following benefits over other options:

1) We can answer both the business questions
2) Easier to implement as compared to scd-2
3) Will consume considerably less space than scd-2 in terms of no of rows.

Question 7.Would your answer to Question 6 be different if the number of customers and/or the time interval between score recalculations was much larger or much smaller? (Explain your answer.)

Ans: No. My answer remains same irrespective of the number of customers or time interval between score recalculation, because our solution is independent of that. Both of these cases will not impact our solution in terms of scaling/database size/history preservation: