AP® STATISTICS
2009 SCORING GUIDELINES

Question 3

Intent of Question

The primary goals of this question were to assess a student's ability to (1) describe a randomization process required for comparing two groups in a randomized experiment; and (2) describe a potential consequence of using self-selection instead of randomization.

Solution

Part (a) (completely randomized design):

Each student will be assigned a unique random number using a random number generator on a calculator, statistical software, or a random number table. The assigned numbers will be listed in ascending order. The students with the lowest 12 numbers in the ordered list will receive the instructional program that requires physically dissecting frogs. The students with the highest 12 numbers will receive the instructional program that uses computer software to simulate the dissection of a frog.

Part (a) *alternative* (randomized block design):

Students will be paired or placed into blocks of size two, based on having similar pretest scores. So, the first block will contain the two students with the two lowest pretest scores, the second block will contain the two students with the third- and fourth-lowest pretest scores, and so on, with the last block containing the two students with the two highest pretest scores. In each block, the students will be assigned a unique random number using a random number generator on a calculator, statistical software, or a random number table. The student in each block with the lower random number will receive the instructional program that requires physically dissecting frogs, and the student with the higher random number will receive the instructional program that uses computer software to simulate the dissection of a frog.

Part (b):

By not randomizing and allowing the students to self-select, there is a potential for changes to occur in the differences between pretest and posttest scores for a particular group because of the characteristics of students who choose a particular instructional method, not because of the instructional method itself. For example, suppose frog-loving students already know a lot about frog anatomy; one would therefore expect these students to be less likely to show a large change between the pretest and posttest scores. Suppose the frog-loving students tend to select the computer simulation method (perhaps because they do not like the notion of dissecting the frogs they love). The possible low change between pretest and posttest scores for the computer simulation group might then be attributed to the students' already knowing a lot about frog anatomy beforehand, not to the instructional method itself. The frog dissection group might see a larger change in scores because the students entering this group are those with the lower pretest scores (less prior knowledge) and who are thus more likely to show greater improvement between pretest and posttest scores.

Scoring

Parts (a) and (b) are scored as essentially correct (E), partially correct (P), or incorrect (I).

Question 3 (continued)

Part (a) is scored as follows:

Essentially correct (E) if a proper method of randomization is described that (1) creates two groups of equal size; *AND* (2) assigns the named treatments to the groups in a manner that knowledgeable statistics users would employ to assign the students to the two instructional groups.

Partially correct (P) if only one of the two criteria above is met.

Incorrect (I) if neither criterion is met.

Notes:

- Coin tossing (or equivalent method) using a stopping rule to obtain equal sample sizes requires placing the students in the class in a random order. If this method does not include a random order, at best, part (a) is scored as partially correct.
- In using a random number table, if numbers are specified, the student must work with two-digit numbers. For example, if using the first 24 integers, the student must use 01–24, not 1–24. If the student uses numbers such as 1–24, a solution that would otherwise be essentially correct becomes partially correct, and a partially correct response becomes incorrect.

Part (a) *alternative* is scored as follows:

Essentially correct (E) if (1) blocks are formed based on students' having similar pretest scores; *AND* (2) the two students in each block are assigned to different treatments; *AND* (3) the method of randomization used to assign the students in each block to the treatments is correct and can be implemented after reading the student's response (in a manner that knowledgeable statistics users would employ to assign the students to the two instructional groups).

Partially correct (P) if two of the three components above are presented correctly.

Incorrect (I) if no more than one of the three components is presented correctly.

Part (b) is scored as follows:

Essentially correct (E) if (1) the example gives a reasonable characteristic of the self-selected students in the study; *AND* (2) explains how this characteristic could be associated with changes in the differences between the pretest and posttest scores.

Partially correct (P) if (1) the example gives a reasonable characteristic of the self-selected students in the study; *AND* (2) a weak explanation is provided of how this characteristic could be associated with changes in the differences between pretest and posttest scores.

Note: A *weak* explanation of how a characteristic could be associated with changes in the differences between pretest and posttest scores must at least mention test scores or state that one group will perform better than the other. (Simply mentioning a behavioral difference is not sufficient.)

Question 3 (continued)

Incorrect (I) if an incorrect or no explanation is provided of how a characteristic could be associated with changes in the differences between pretest and posttest scores

OR

the example does not give a reasonable characteristic of the self-selected students in the study

OR

a student says that there must be an equal number of students in the class assigned to each treatment.

4 Complete Response

Both parts essentially correct

3 Substantial Response

One part essentially correct and the other part partially correct

2 Developing Response

One part essentially correct and the other part incorrect

OR

Both parts partially correct

1 Minimal Response

No part essentially correct and only one part partially correct

3. Before beginning a unit on frog anatomy, a seventh-grade biology teacher gives each of the 24 students in the class a pretest to assess their knowledge of frog anatomy. The teacher wants to compare the effectiveness of an instructional program in which students physically dissect frogs with the effectiveness of a different program in which students use computer software that only simulates the dissection of a frog. After completing one of the two programs, students will be given a posttest to assess their knowledge of frog anatomy. The teacher will then analyze the changes in the test scores (score on posttest minus score on pretest).

(a) Describe a method for assigning the 24 students to two groups of equal size that allows for a statistically valid comparison of the two instructional programs.

Give each student a number, 01, 02, 03 ... 24

Then use a random digit table or a random digit generator to generate digits. The first 12 digits chosen will be assigned to dissect the frog, and the 12 leftover will use the computer program.

(b) Suppose the teacher decided to allow the students in the class to select which instructional program on frog anatomy (physical dissection or computer simulation) they prefer to take, and 11 students choose actual dissection and 13 students choose computer simulation. How might that self-selection process jeopardize a statistically valid comparison of the changes in the test scores (score on posttest minus score on pretest) for the two instructional programs? Provide a specific example to support your answer.

Perhaps all of the children who know a lot about frogs already really like frogs, so they want to dissect hands on. Because they have more previous knowledge, they have less room for improvement on their post-test The kids who don't like frogs, and don't know much about them will use the computer. Because they will most likely score lower on the pre-test, they have more room for improvement. If children who use the computer program are improving more because they knew less in the first place, there is no way to prove that either of one of the programs is more effective than the other.

GO ON TO THE NEXT PAGE.

3B 3B

3. Before beginning a unit on frog anatomy, a seventh-grade biology teacher gives each of the 24 students in the class a pretest to assess their knowledge of frog anatomy. The teacher wants to compare the effectiveness of an instructional program in which students physically dissect frogs with the effectiveness of a different program in which students use computer software that only simulates the dissection of a frog. After completing one of the two programs, students will be given a posttest to assess their knowledge of frog anatomy. The teacher will then analyze the changes in the test scores (score on posttest minus score on pretest).

(a) Describe a method for assigning the 24 students to two groups of equal size that allows for a statistically valid comparison of the two instructional programs.

What the teacher should do is follow a Matched Pairs Design. The teacher should pair similar units together, which in this case are the students. The teacher, to prevent some confounding, should match the students with their closest classmate in terms of pretest score. Then, the teacher can randomly assign the physical dissection program to one and the simulated program to the other. After they are completed their programs, the teacher can then give them their posttests and compare those scores between their original pretest partners.

(b) Suppose the teacher decided to allow the students in the class to select which instructional program on frog anatomy (physical dissection or computer simulation) they prefer to take, and 11 students choose actual dissection and 13 students choose computer simulation. How might that self-selection process jeopardize a statistically valid comparison of the changes in the test scores (score on posttest minus score on pretest) for the two instructional programs? Provide a specific example to support your answer.

The self-selection process would turn the experiment into an observation, in which confounding would pose a threat. For example, better test takers might favor the computer, and therefore the simulated dissection program, thereby resulting in different changes in testscores and differences between testscore changes that could be more or less extreme.

GO ON TO THE NEXT PAGE.

3C 3. Before beginning a unit on frog anatomy, a seventh-grade biology teacher gives each of the 24 students in the class a pretest to assess their knowledge of frog anatomy. The teacher wants to compare the effectiveness of an instructional program in which students physically dissect frogs with the effectiveness of a different program in which students use computer software that only simulates the dissection of a frog. After completing one of the two programs, students will be given a posttest to assess their knowledge of frog anatomy. The teacher will then analyze the changes in the test scores (score on posttest minus score on pretest). 3C

(a) Describe a method for assigning the 24 students to two groups of equal size that allows for a statistically valid comparison of the two instructional programs.

Number students according to an alphabetical list. 00-23; skip 24-99. Use table of random digits to assign the first 12 chosen to go to test 1 (actual disection) and whoever is left over goes to test 2 (computer)

NO REPEATS!

(b) Suppose the teacher decided to allow the students in the class to select which instructional program on frog anatomy (physical dissection or computer simulation) they prefer to take, and 11 students choose actual dissection and 13 students choose computer simulation. How might that self-selection process jeopardize a statistically valid comparison of the changes in the test scores (score on posttest minus score on pretest) for the two instructional programs? Provide a specific example to support your answer.

Students may have chosen the method based on where their FRIENDS went. When they are with their friends they are less likely to focus on learning and more on socializing so we cannot accurately deduce whether the program was innefective because it wasn't good or because the child wasn't paying attention

GO ON TO THE NEXT PAGE.

Question 3

Overview

The primary goals of this question were to assess a student's ability to (1) describe a randomization process required for comparing two groups in a randomized experiment and (2) describe a potential consequence of using self-selection instead of randomization.

Sample: 3A
Score: 4

In part (a) the student assigns each student in the biology class a unique number from 01 to 24 and uses a random number generator correctly to form two groups of size 12. The student indicates which group will dissect the frog and which will use the computer program, giving the context of the problem. Part (a) was scored as essentially correct. In part (b) the student clearly explains that the self-selection to programs could be based on the amount children know about frogs: the "children who know a lot about frogs already" choose the dissection program, while the children who "don't know much about" frogs choose the computer program. The student then argues that the students in the dissection program will tend to have a relatively small improvement, while the students in the computer program have a greater opportunity for improvement. In the last sentence the student provides a clear summary of the problem caused by self-selection. The strong response in this part was scored as essentially correct. The entire answer, based on both parts, was judged a complete response and earned a score of 4 points.

Sample: 3B
Score: 3

In part (a) the student tries to describe a "Matched Pairs Design." Blocks are reasonably formed, consisting of "students with their closest classmate in terms of pretest score." It should be noted that matching students based on pretest without explicitly saying they would be students with *similar* pretests would have been an insufficient description of the blocks. There is also a clear indication that the student knows that the two students in each block are to be assigned different treatments. However, the student makes no attempt to describe a randomization process that would assign the students to the treatments. Because the student provides two of the three required components, part (a) was scored as partially correct. In part (b) the student indicates that "better test takers might favor the computer, and therefore the simulated dissection program"; by implication, the worse test-takers must be in the actual dissection program. The student argues that the differing test-taking abilities of the students in the two programs would result in "different changes in testscores and differences between testscore changes that could be more or less extreme." While it would have been better for the student to indicate that the change for the simulation group might be larger than with a random sample and smaller for the dissection program, the student demonstrates a reasonable understanding of the problem, and part (b) was scored as essentially correct. With one part essentially correct and one part partially correct, the entire answer was judged a substantial response and earned a score of 3 points.

Sample: 3C
Score: 2

In part (a) the student uses a table of random digits to assign the first 12 students to "actual disection [*sic*]" and the remaining students to "computer." Part (a) was scored as essentially correct. In part (b) the student

Question 3 (continued)

defines the self-selection criterion as students choosing to join the group “where their FRIENDS went.” But then the student describes the consequence of the self-selection as a change in behavior, not a change in test performance, so this part was scored as incorrect. With one part essentially correct and one part incorrect, the entire answer was judged a developing response and earned a score of 2 points.