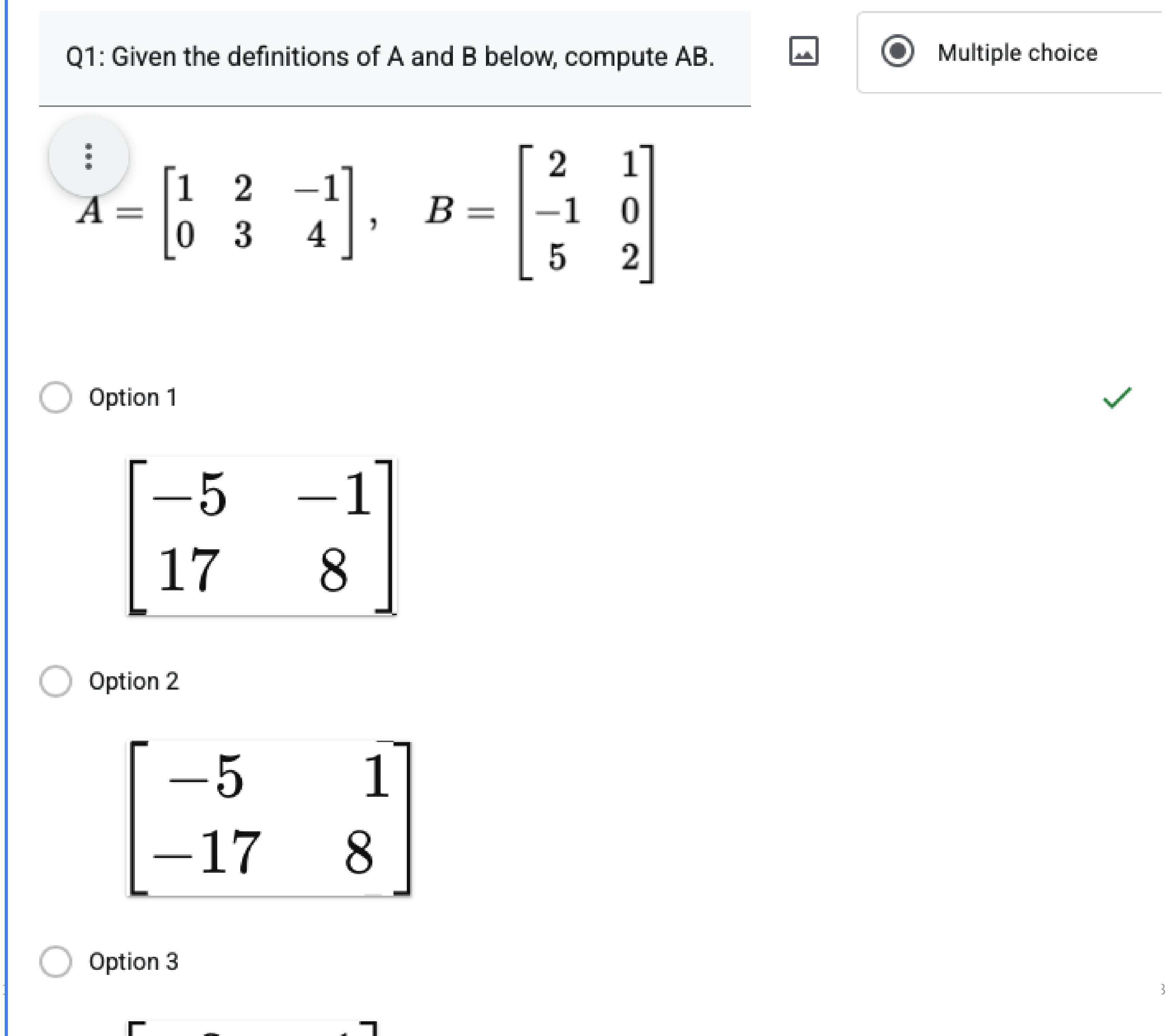
UVA CS 4774: Machine Learning

S6: Lecture 30: Quiz Reviews

Dr. Yanjun Qi

University of Virginia
Department Of Computer Science

Q1



Q1: Given the definitions of A and B below, compute AB.

● Multiple choice

$$A = \begin{bmatrix} 1 & 2 & -1 \\ 0 & 3 & 4 \end{bmatrix}, \quad B = \begin{bmatrix} 2 & 1 \\ -1 & 0 \\ 5 & 2 \end{bmatrix}$$

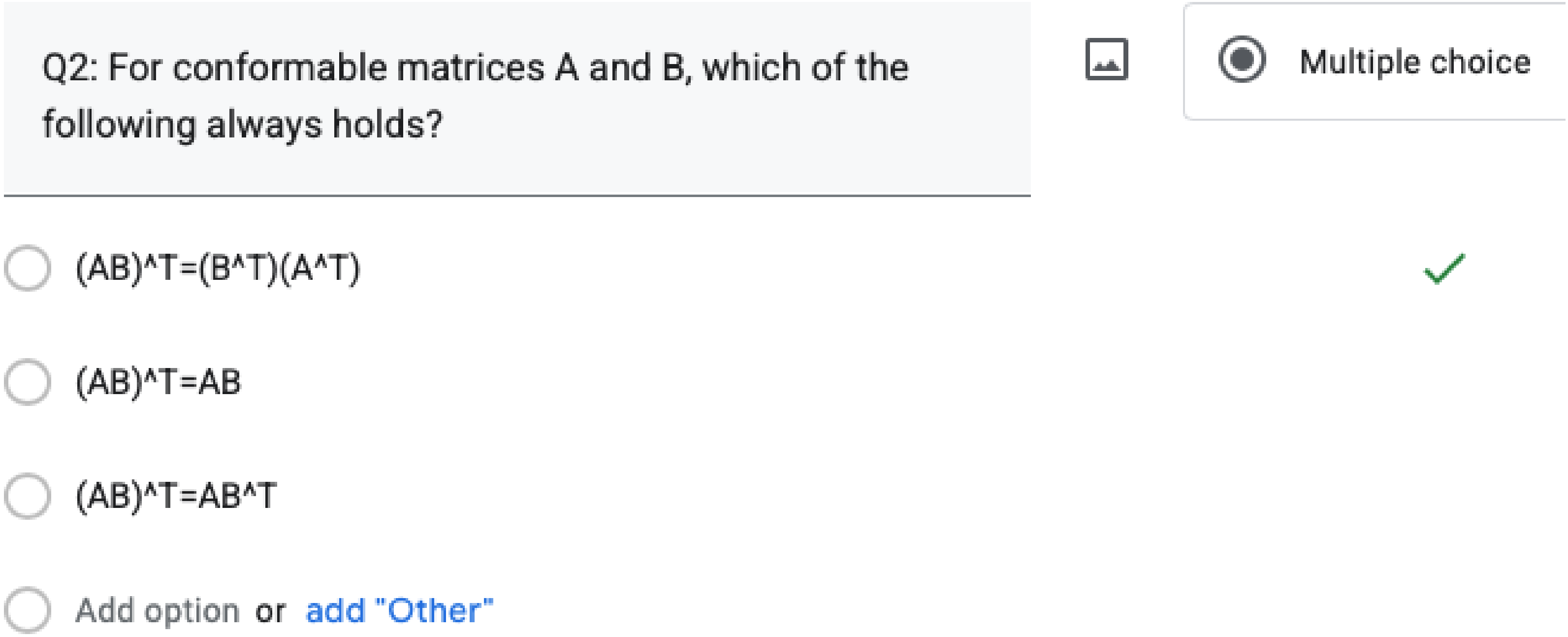
○ Option 1 ✓

$$\begin{bmatrix} -5 & -1 \\ 17 & 8 \end{bmatrix}$$

○ Option 2

$$\begin{bmatrix} -5 & 1 \\ -17 & 8 \end{bmatrix}$$

○ Option 3



Q2: For conformable matrices A and B, which of the following always holds?

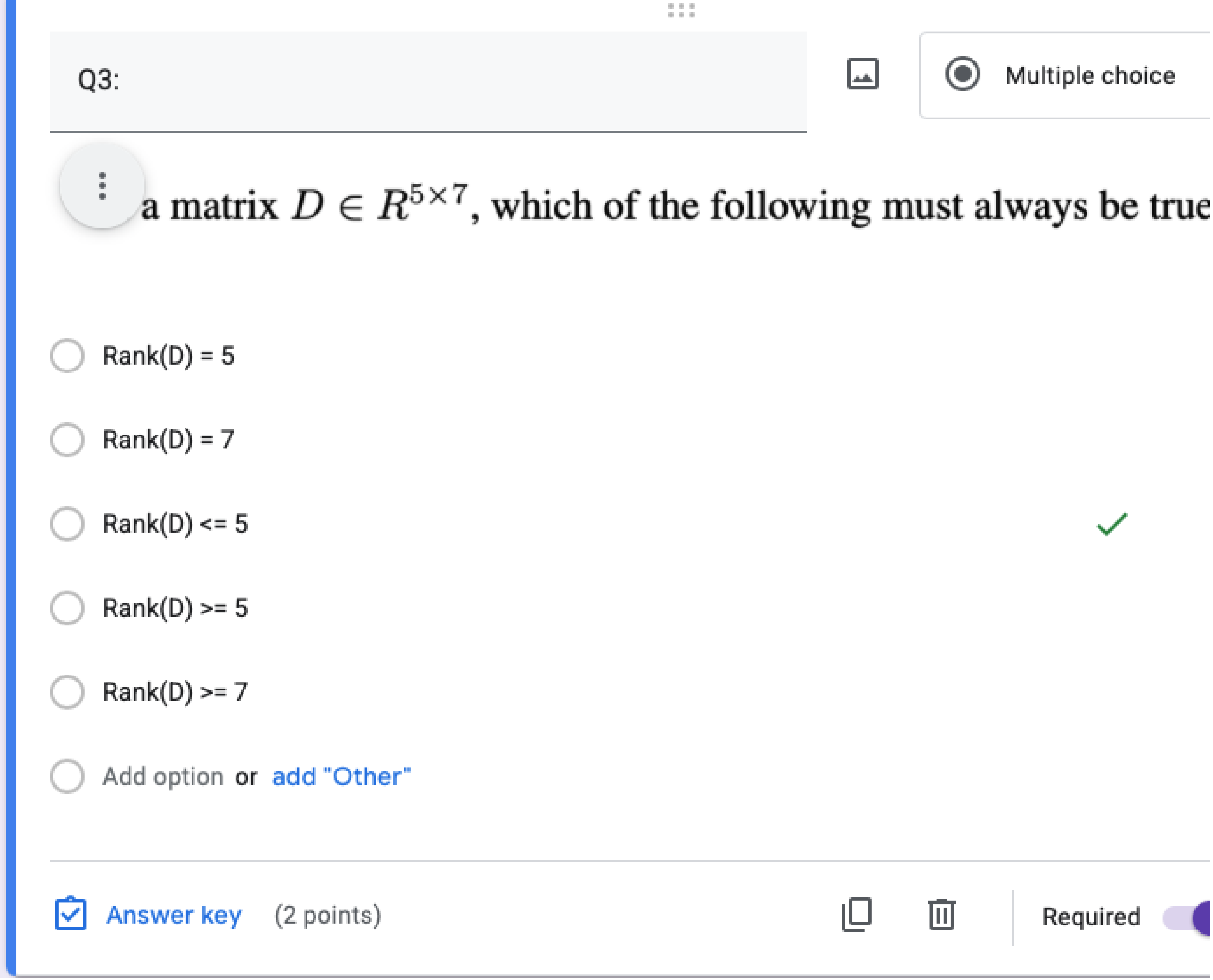
● Multiple choice

○ (AB)^T=(B^T)(A^T) ✓

○ (AB)^T=AB

○ (AB)^T=AB^T

○ Add option or add "Other"



Q3:

● Multiple choice

a matrix $D \in R^{5\times 7}$, which of the following must always be true

○ Rank(D) = 5

○ Rank(D) = 7

○ Rank(D) <= 5 ✓

○ Rank(D) >= 5

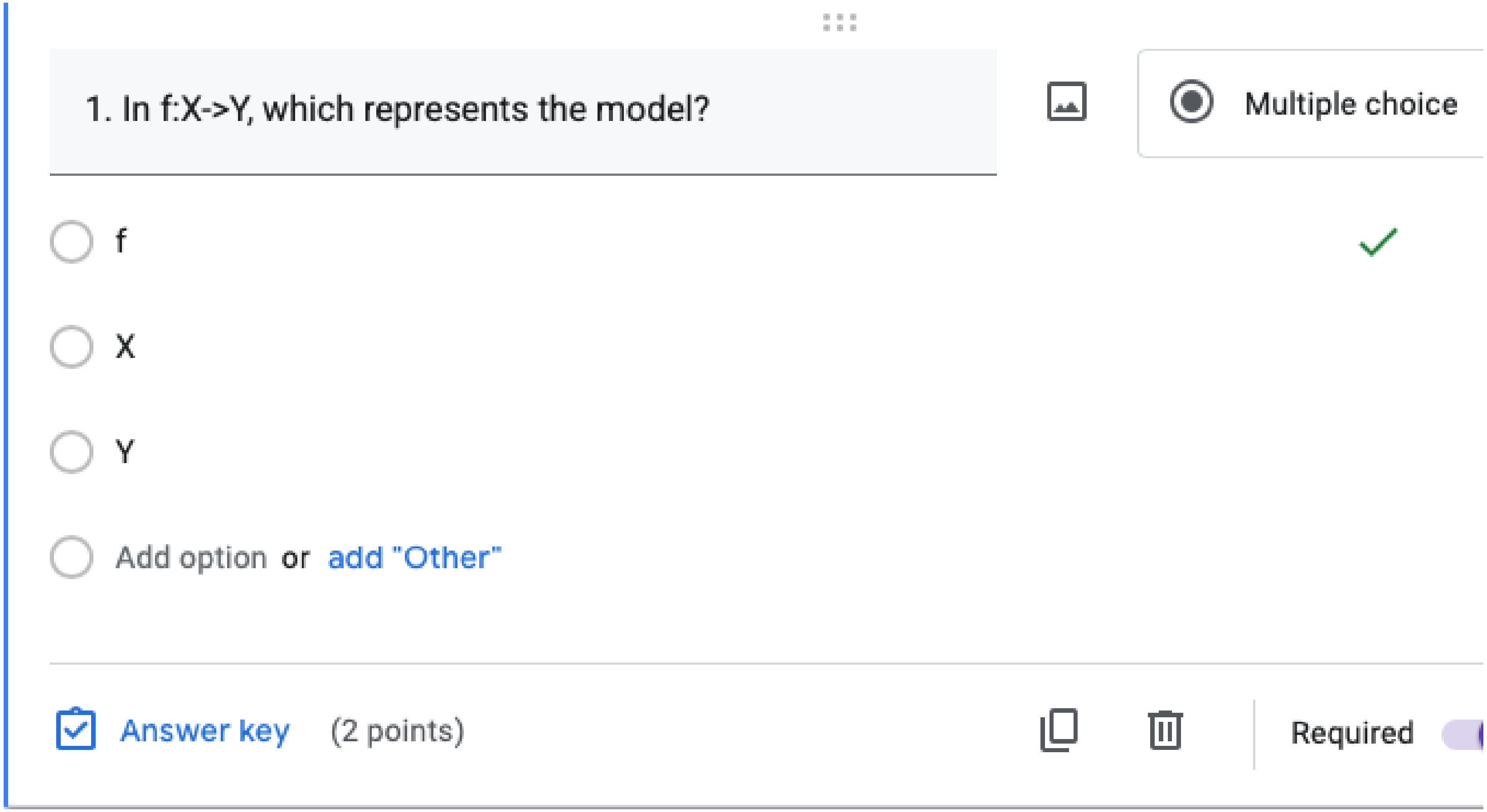
○ Rank(D) >= 7

○ Add option or add "Other"

☑ Answer key (2 points)

Required

Q2



1. In f:X->Y, which represents the model?

● Multiple choice

○ f ✓

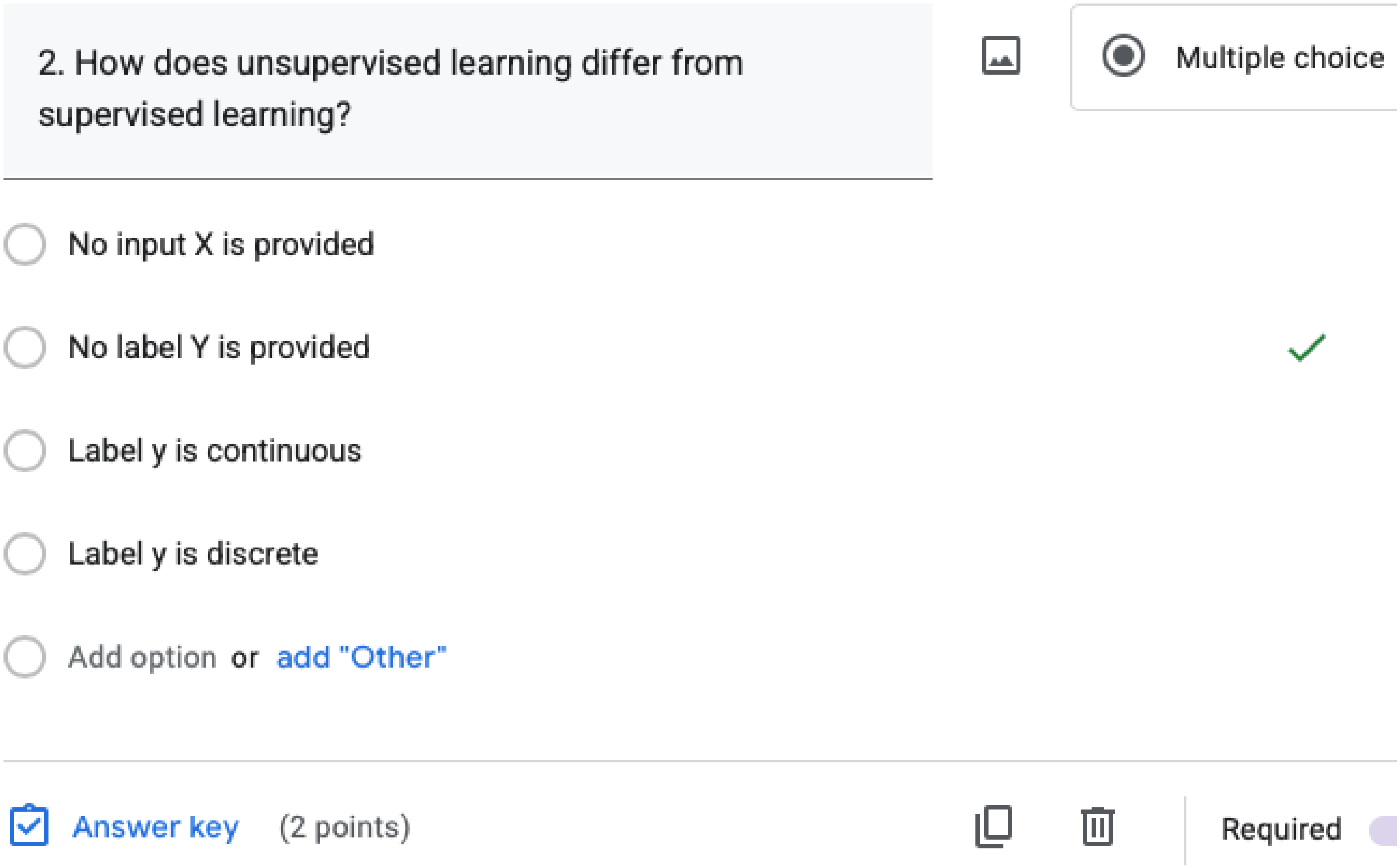
○ X

○ Y

○ Add option or add "Other"

☑ Answer key (2 points)

Required



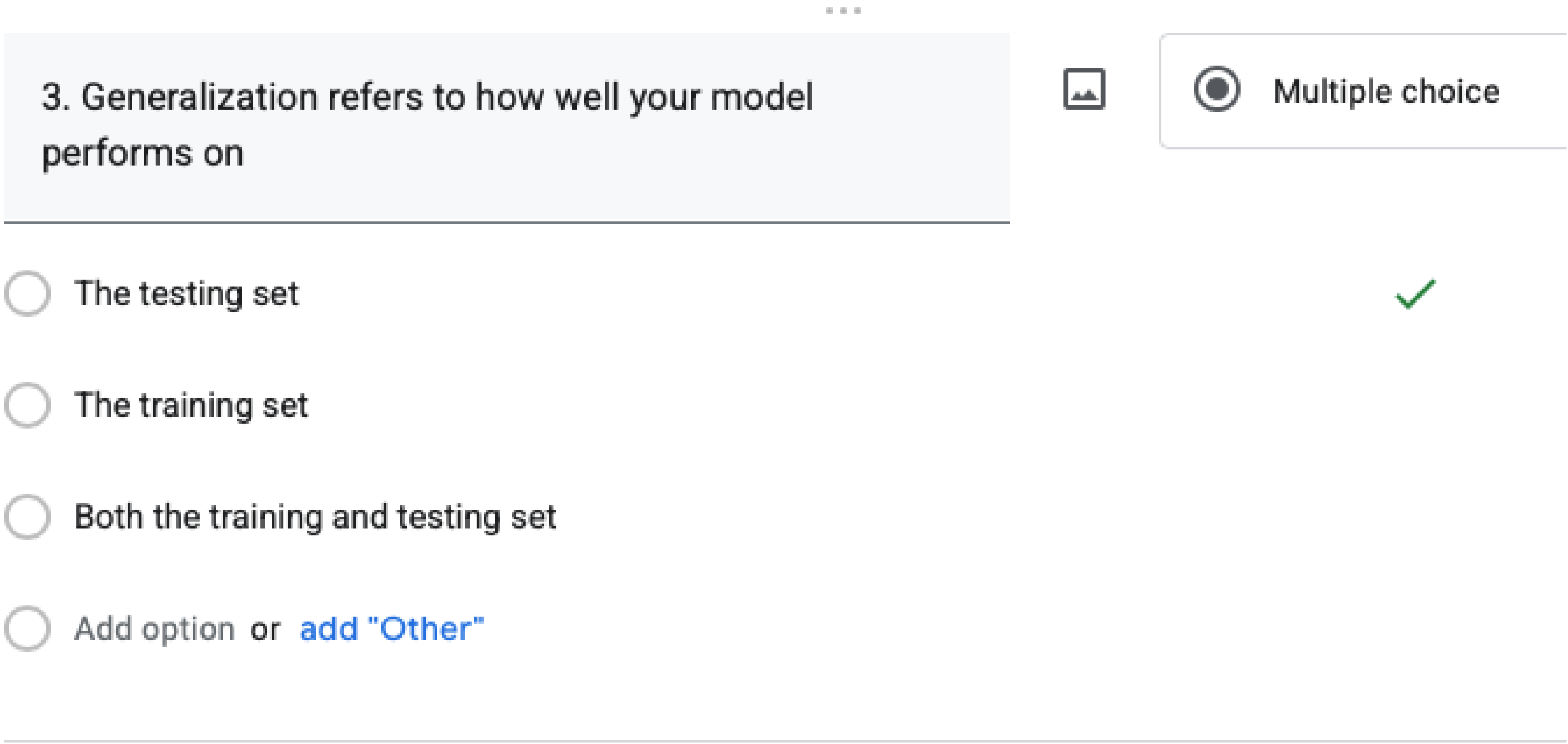
2. How does unsupervised learning differ from supervised learning?

○ No input X is provided

○ No label Y is provided ✓

○ Label y is continuous

○ Label y is discrete



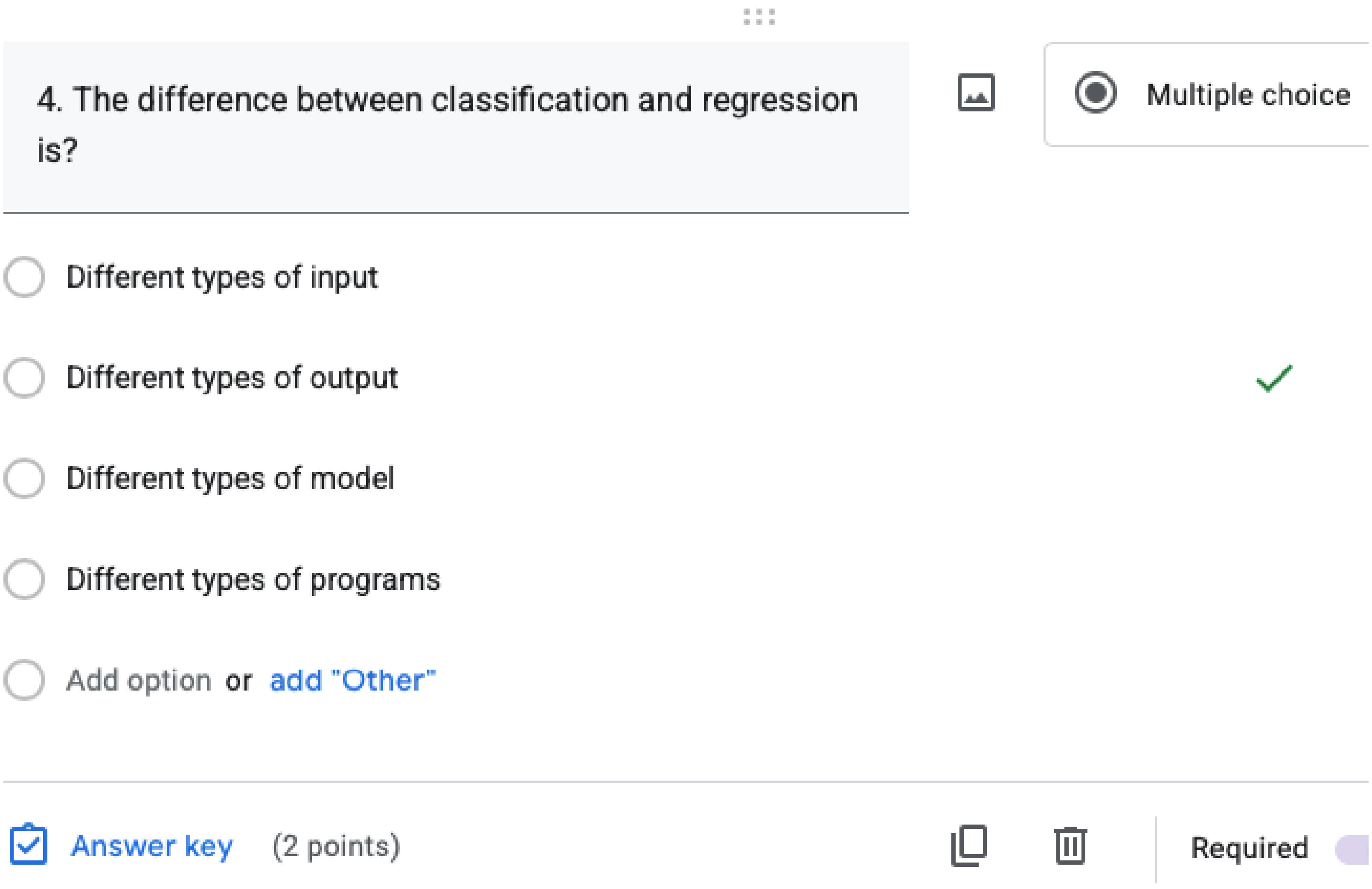
3. Generalization refers to how well your model performs on

● Multiple choice

○ The testing set ✓

○ The training set

○ Both the training and testing set



4. The difference between classification and regression is?

● Multiple choice

○ Different types of input

○ Different types of output ✓

○ Different types of model

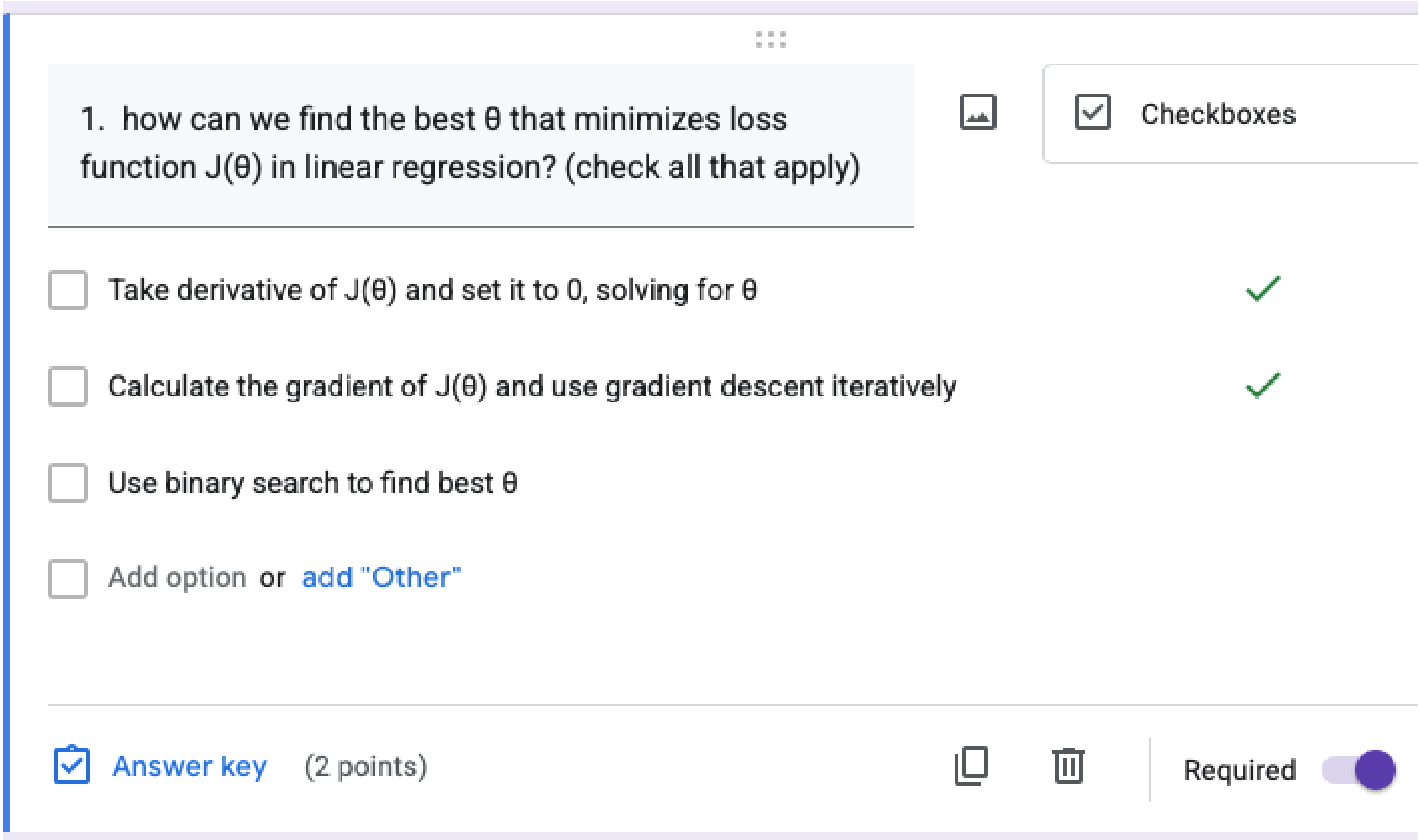
○ Different types of programs

○ Add option or add "Other"

(2 points)

Required

Q3



1. how can we find the best θ that minimizes loss function J(θ) in linear regression? (check all that apply)

☑ Checkboxes

□ Take derivative of J(θ) and set it to 0, solving for θ ✓

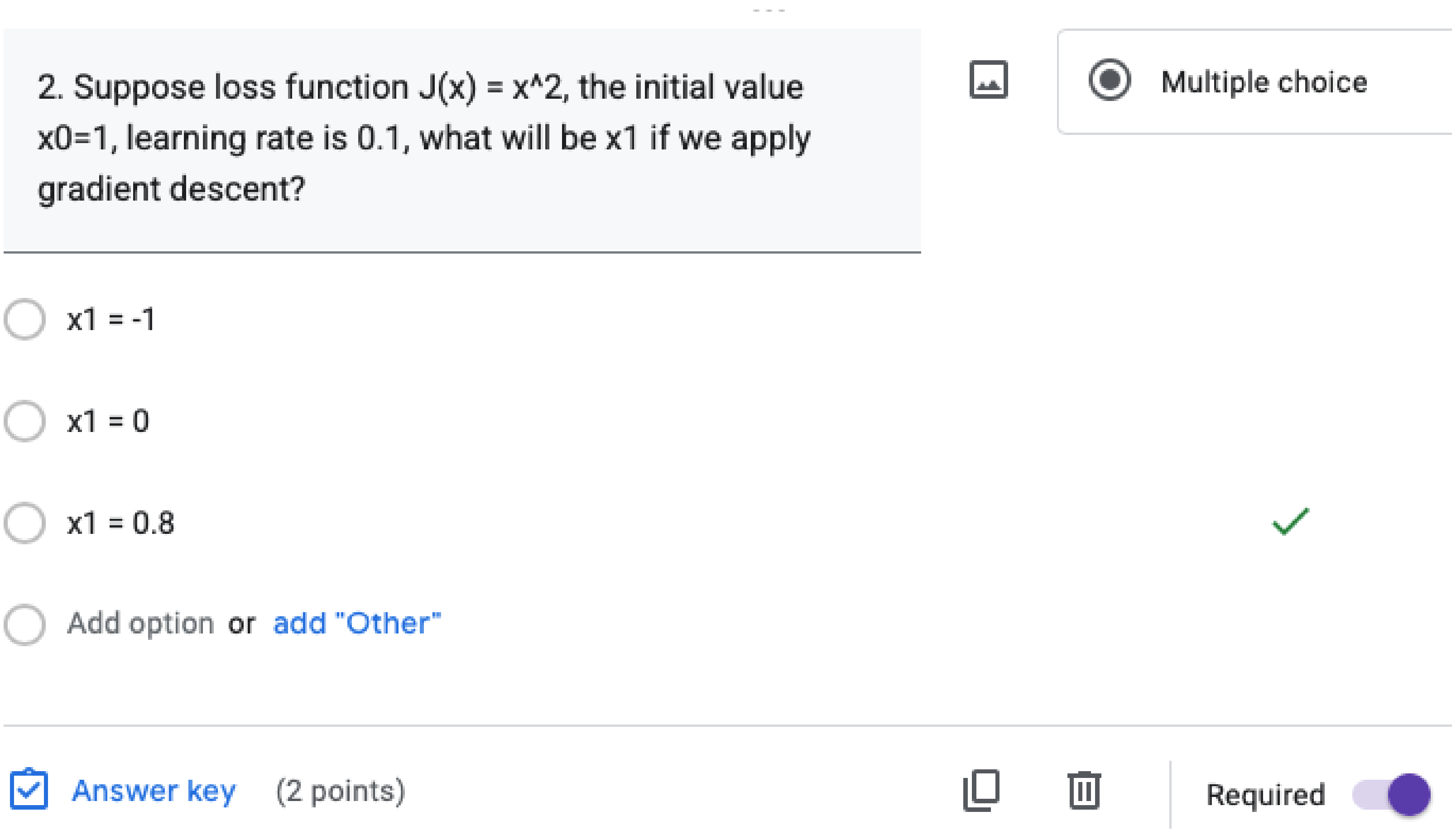
□ Calculate the gradient of J(θ) and use gradient descent iteratively ✓

□ Use binary search to find best θ

□ Add option or add "Other"

☑ Answer key (2 points)

Required



2. Suppose loss function J(x) = x^2, the initial value x0=1, learning rate is 0.1, what will be x1 if we apply gradient descent?

● Multiple choice

○ x1 = -1

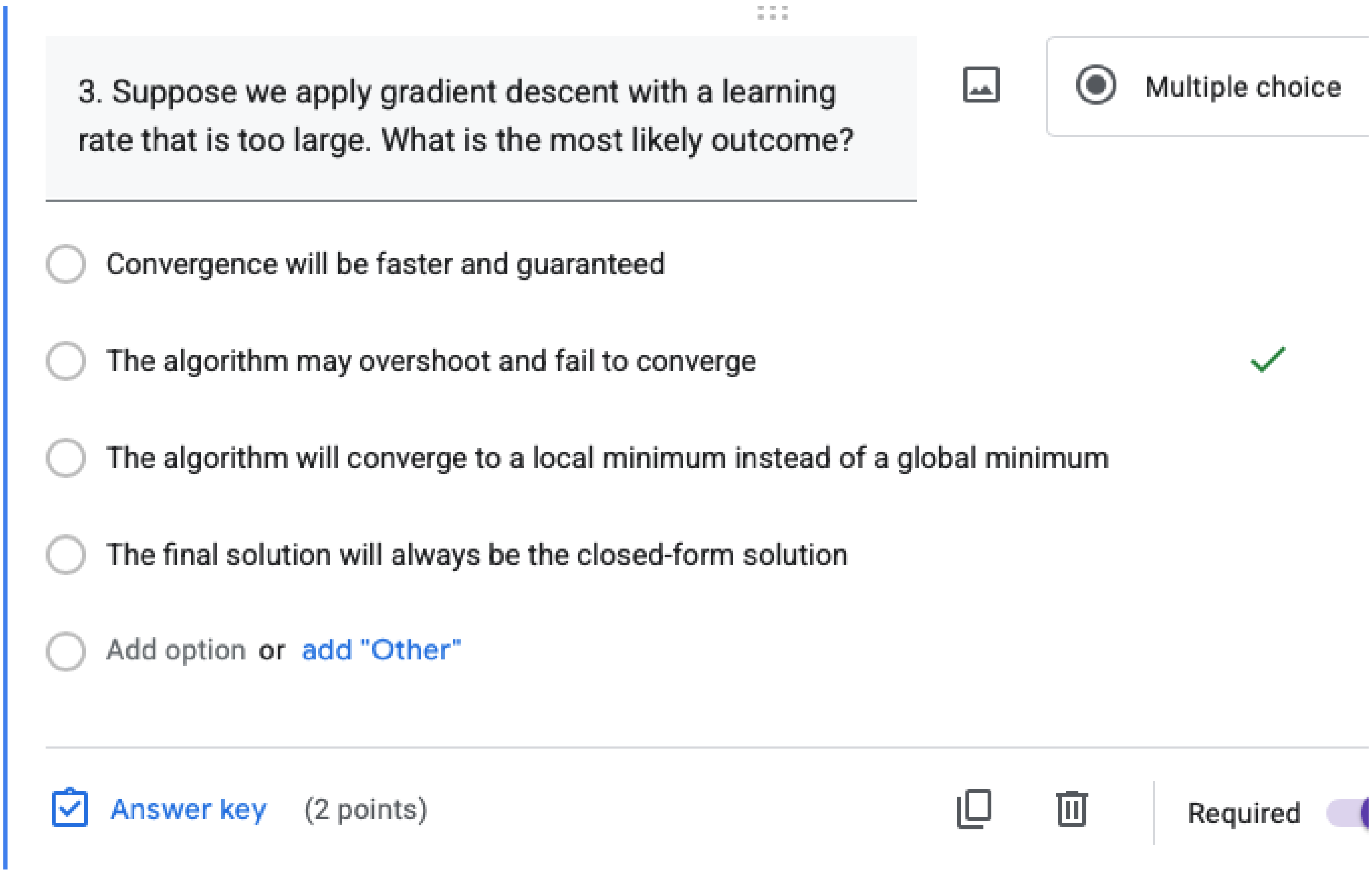
○ x1 = 0

○ x1 = 0.8 ✓

○ Add option or add "Other"

☑ Answer key (2 points)

Required ●



3. Suppose we apply gradient descent with a learning rate that is too large. What is the most likely outcome?

● Multiple choice

○ Convergence will be faster and guaranteed

○ The algorithm may overshoot and fail to converge ✓

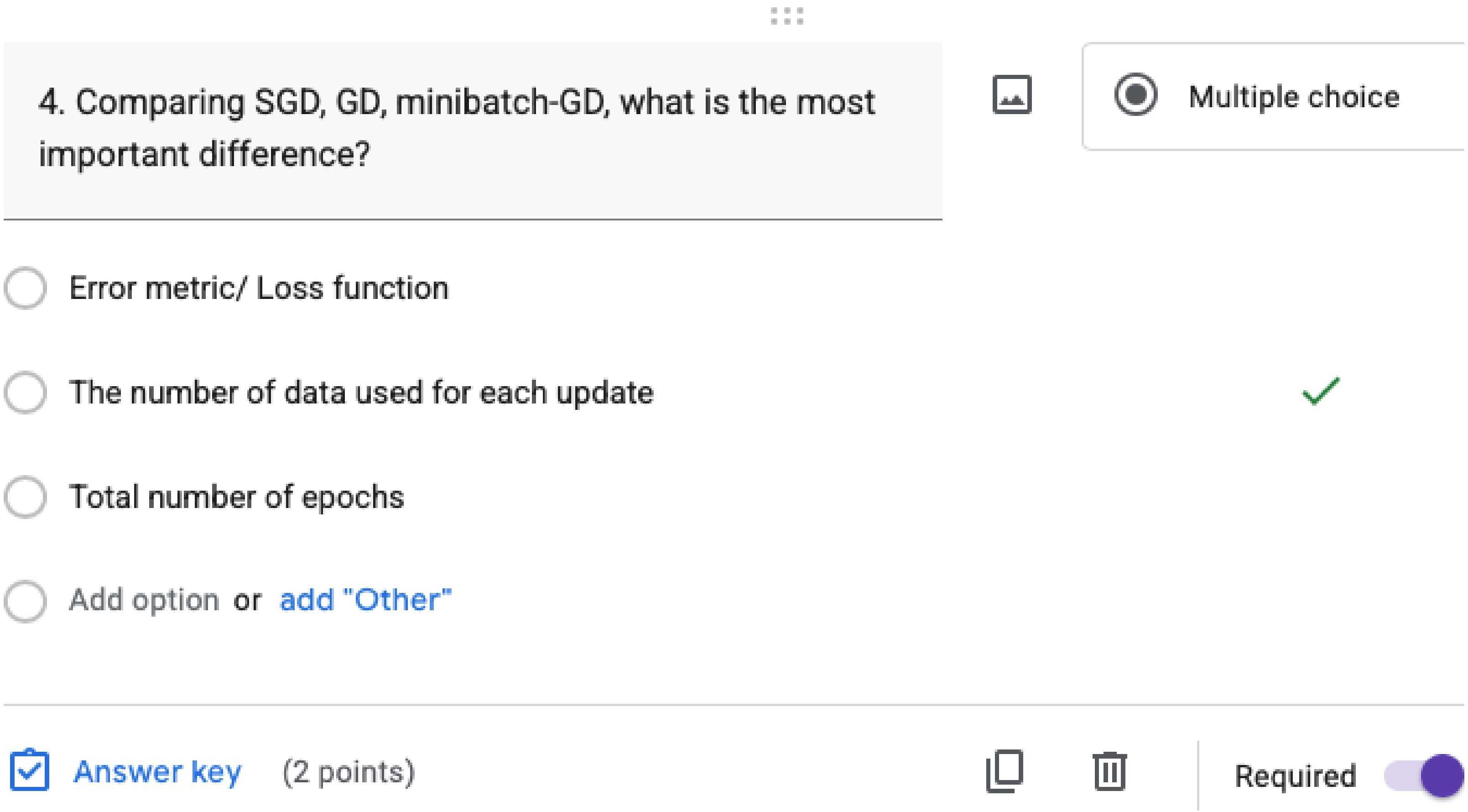
○ The algorithm will converge to a local minimum instead of a global minimum

○ The final solution will always be the closed-form solution

○ Add option or add "Other"

☑ Answer key (2 points)

Required



4. Comparing SGD, GD, minibatch-GD, what is the most important difference?

Multiple choice

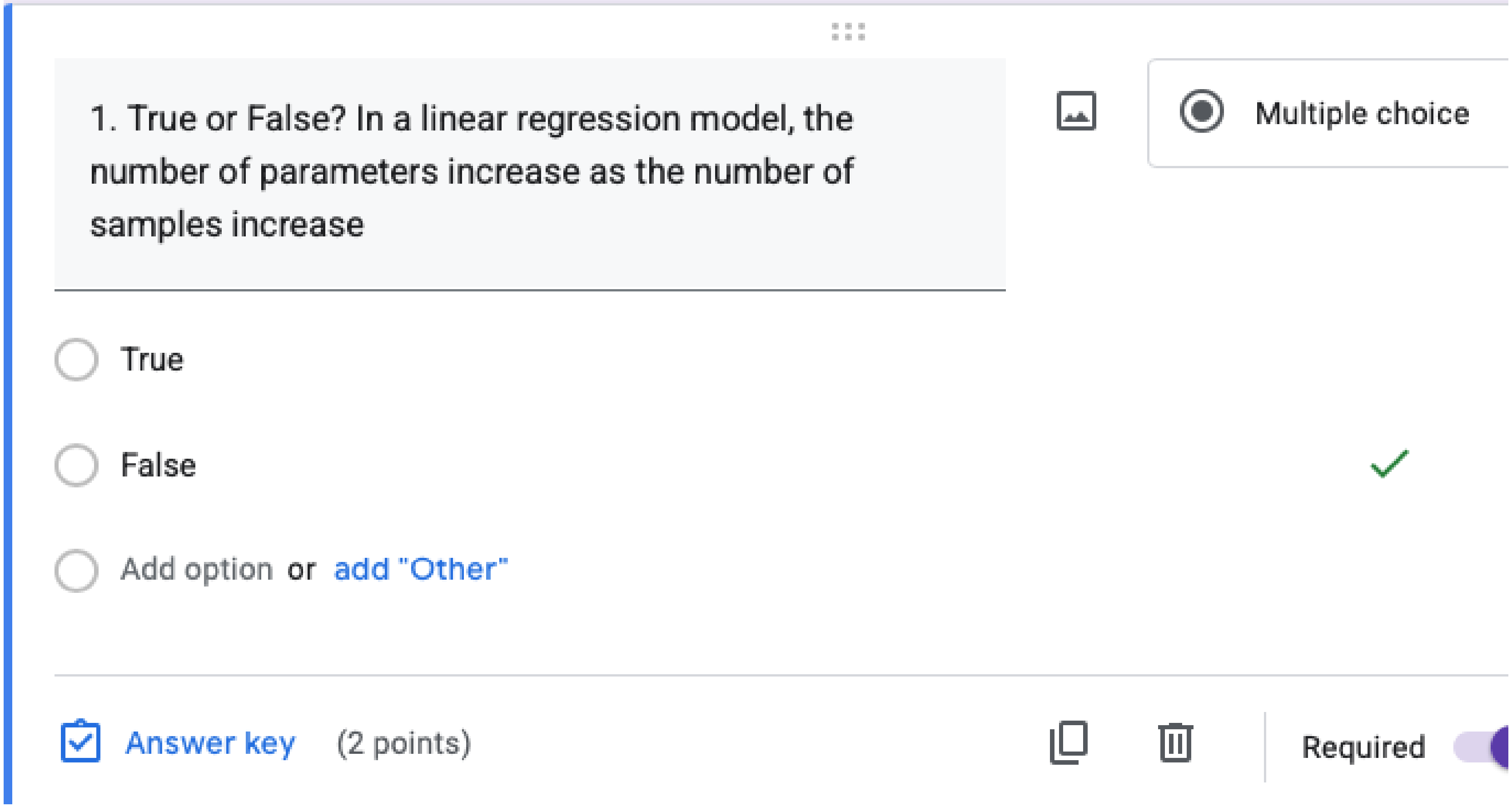
○ Error metric/ Loss function

○ The number of data used for each update ✓

○ Total number of epochs

Q4

Q4 review



1. True or False? In a linear regression model, the number of parameters increase as the number of samples increase

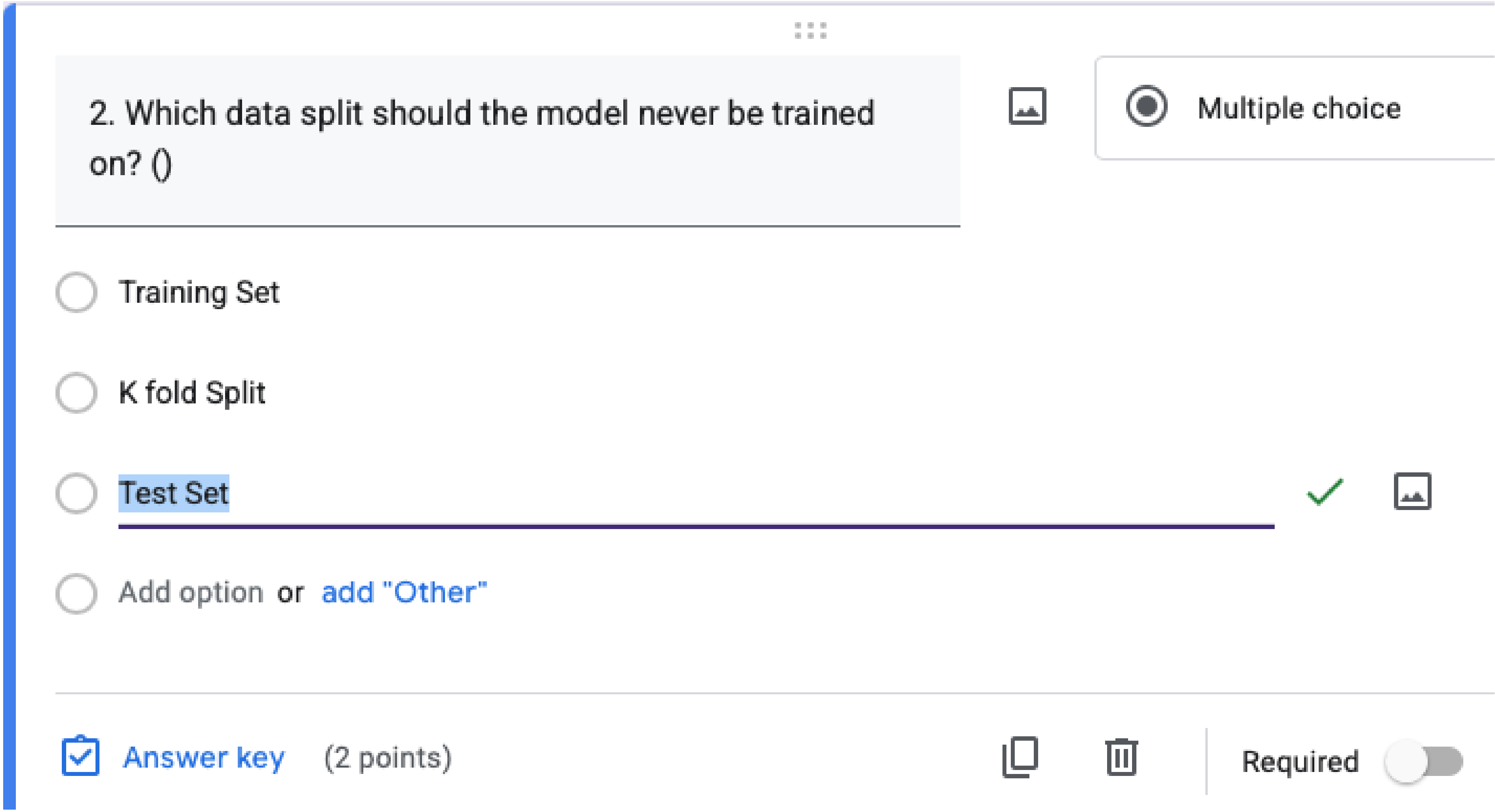
○ True

○ False ✓

○ Add option or add "Other"

(2 points)

Q4 review



2. Which data split should the model never be trained on? ()

● Multiple choice

○ Training Set

○ K fold Split

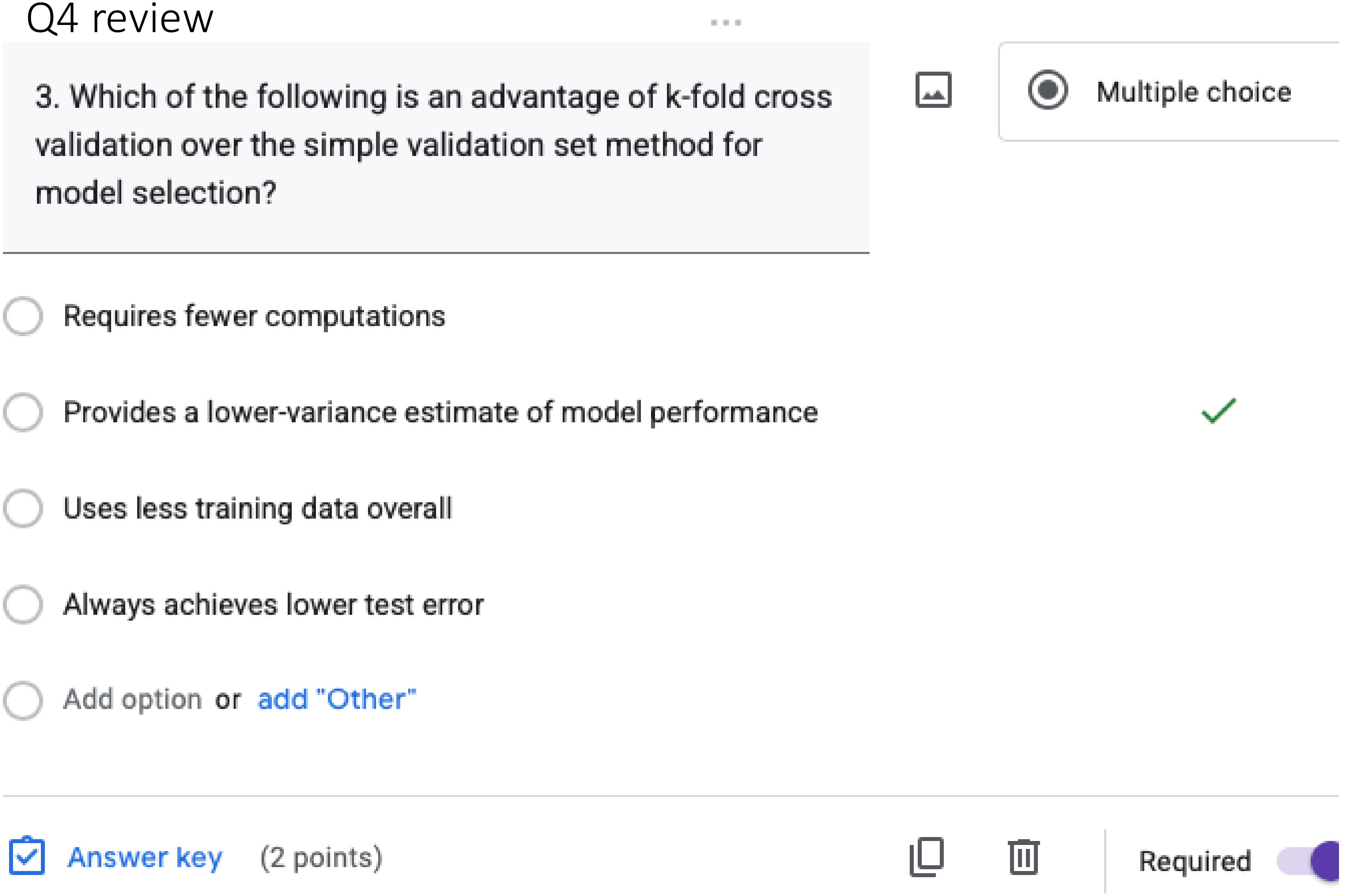
○ Test Set ✓

○ Add option or add "Other"

☑ Answer key (2 points)

Required

Q4 review



● Multiple choice

3. Which of the following is an advantage of k-fold cross validation over the simple validation set method for model selection?

○ Requires fewer computations

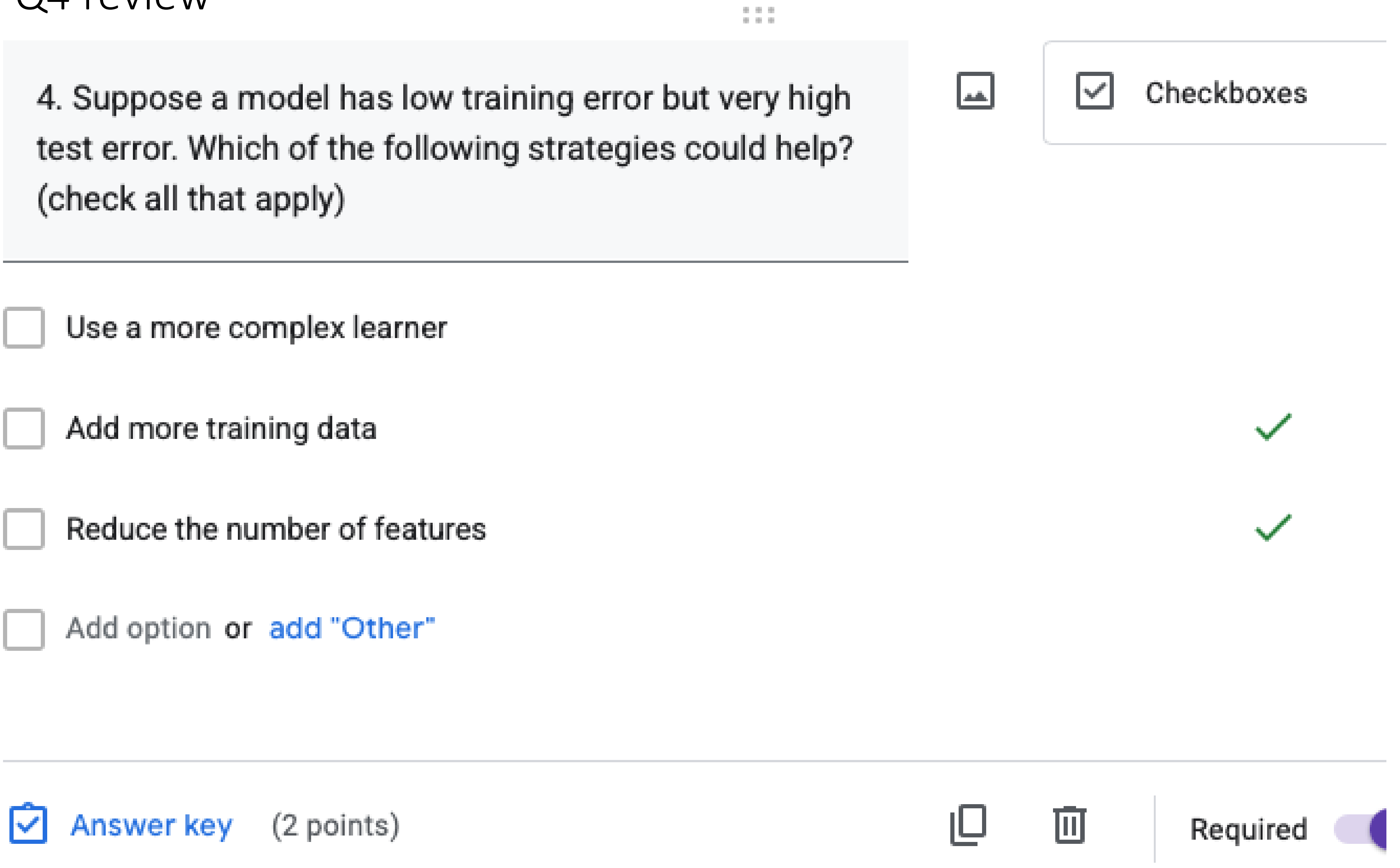
○ Provides a lower-variance estimate of model performance ✓

○ Uses less training data overall

○ Always achieves lower test error

○ Add option or add "Other"

Q4 review



4. Suppose a model has low training error but very high test error. Which of the following strategies could help? (check all that apply)

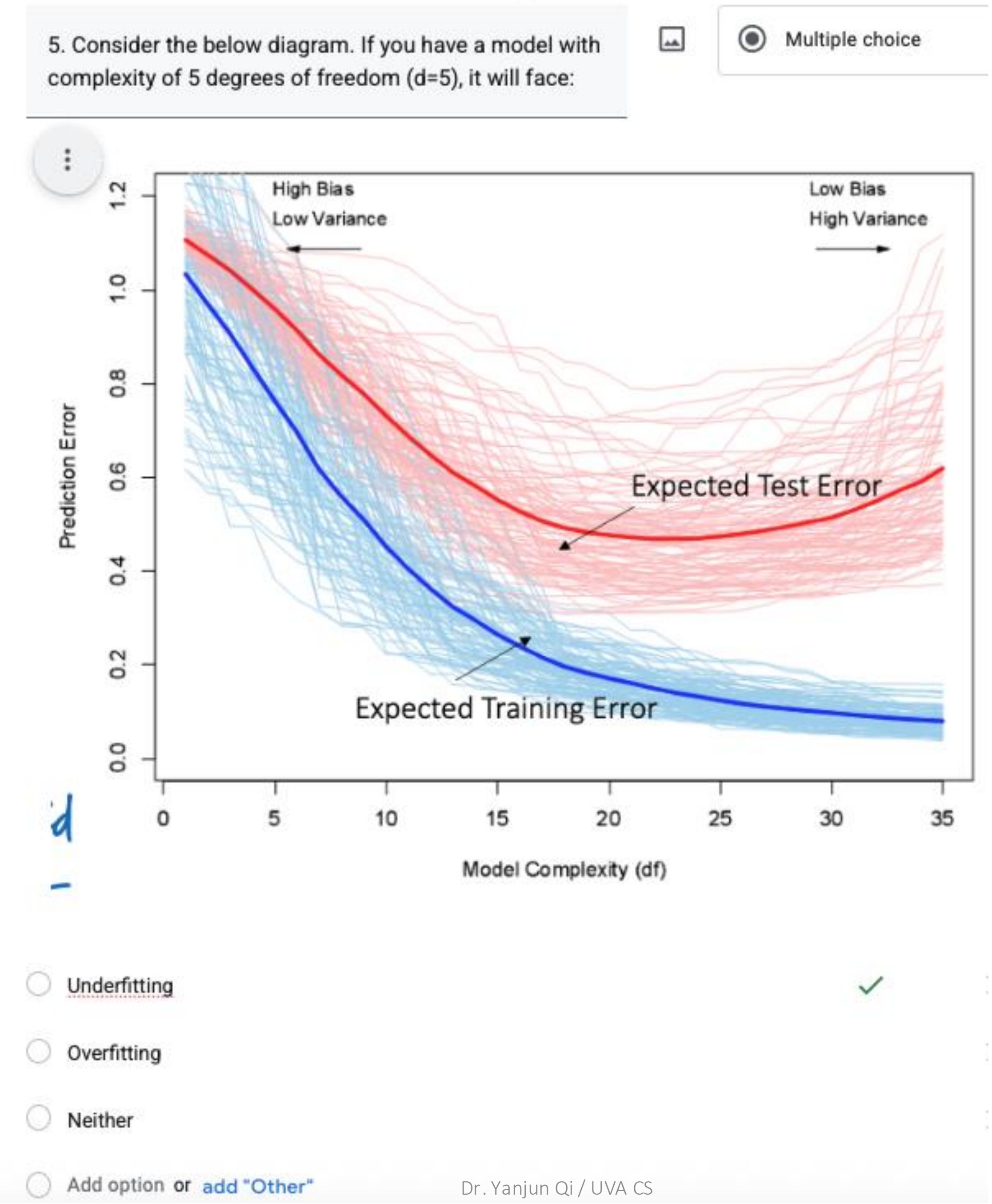
□ Use a more complex learner

□ Add more training data ✓

□ Reduce the number of features ✓

□ Add option or add "Other"

Answer key (2 points)



5. Consider the below diagram. If you have a model with complexity of 5 degrees of freedom (d=5), it will face:

● Multiple choice

d

○ Underfitting ✓

○ Overfitting

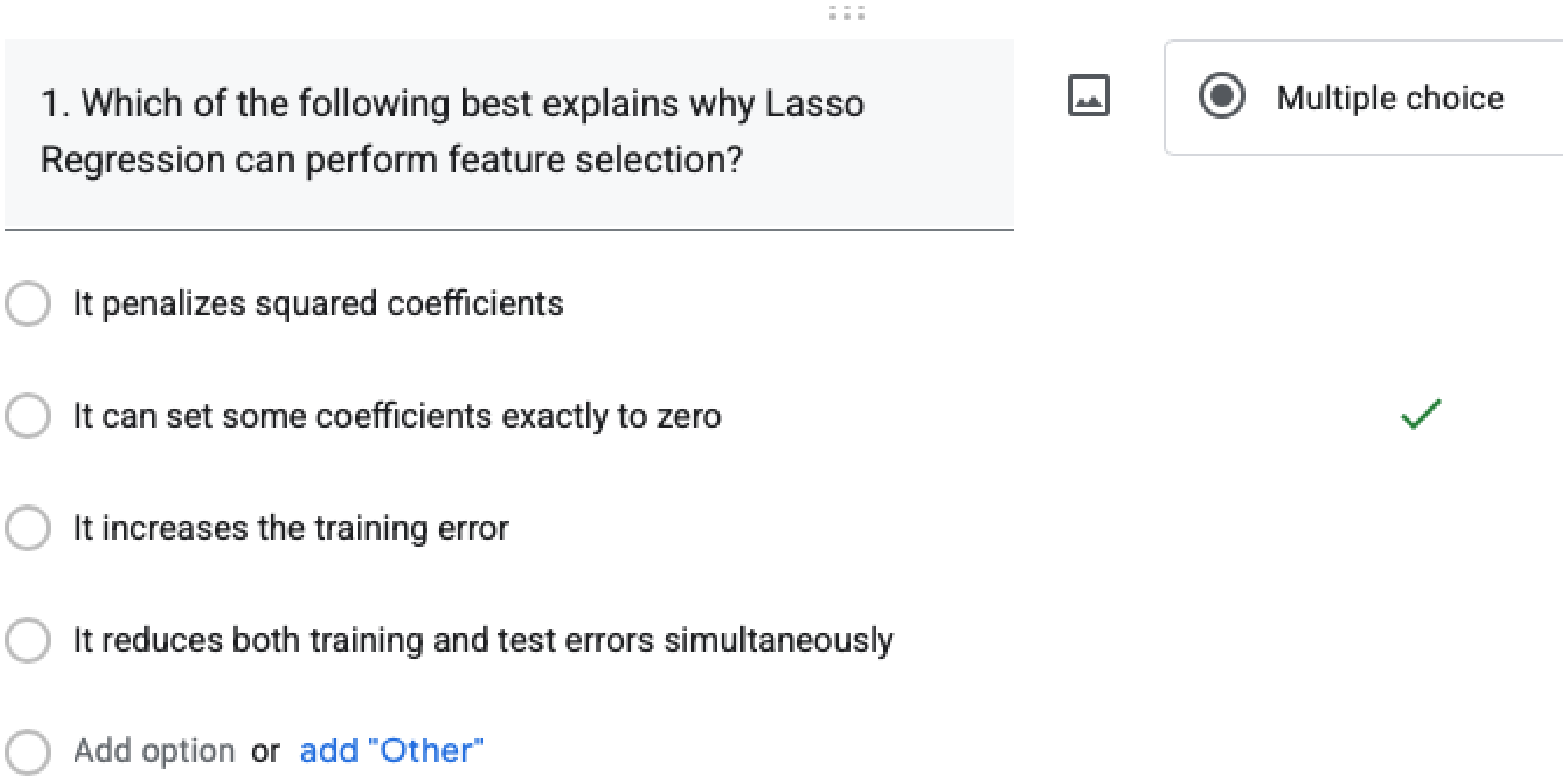
○ Neither

○ Add option or add "Other"

Q5

Q5 Review

[redacted]



1. Which of the following best explains why Lasso Regression can perform feature selection?

● Multiple choice

○ It penalizes squared coefficients

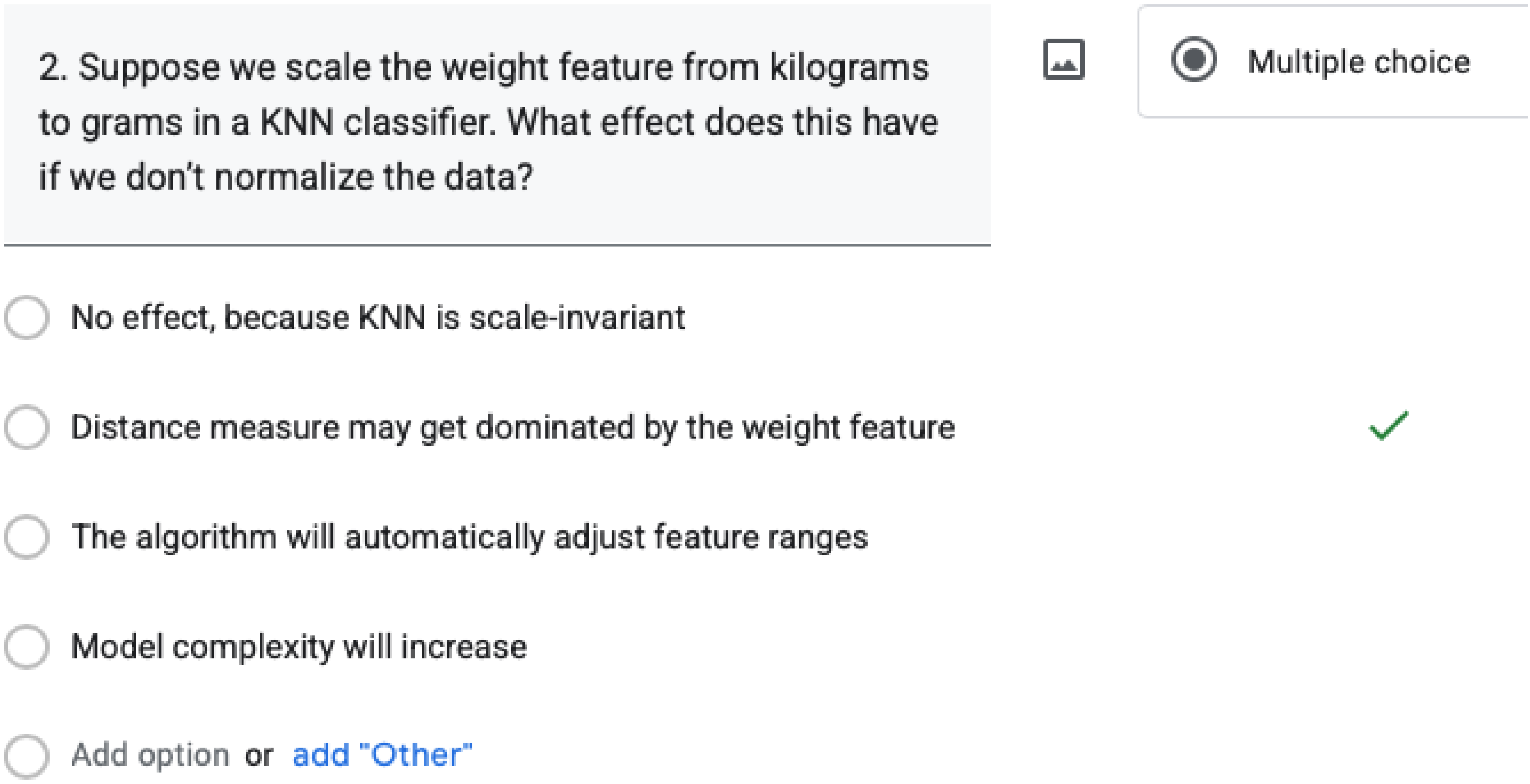
○ It can set some coefficients exactly to zero ✓

○ It increases the training error

○ It reduces both training and test errors simultaneously

○ Add option or add "Other"

Q5



2. Suppose we scale the weight feature from kilograms to grams in a KNN classifier. What effect does this have if we don't normalize the data?

● Multiple choice

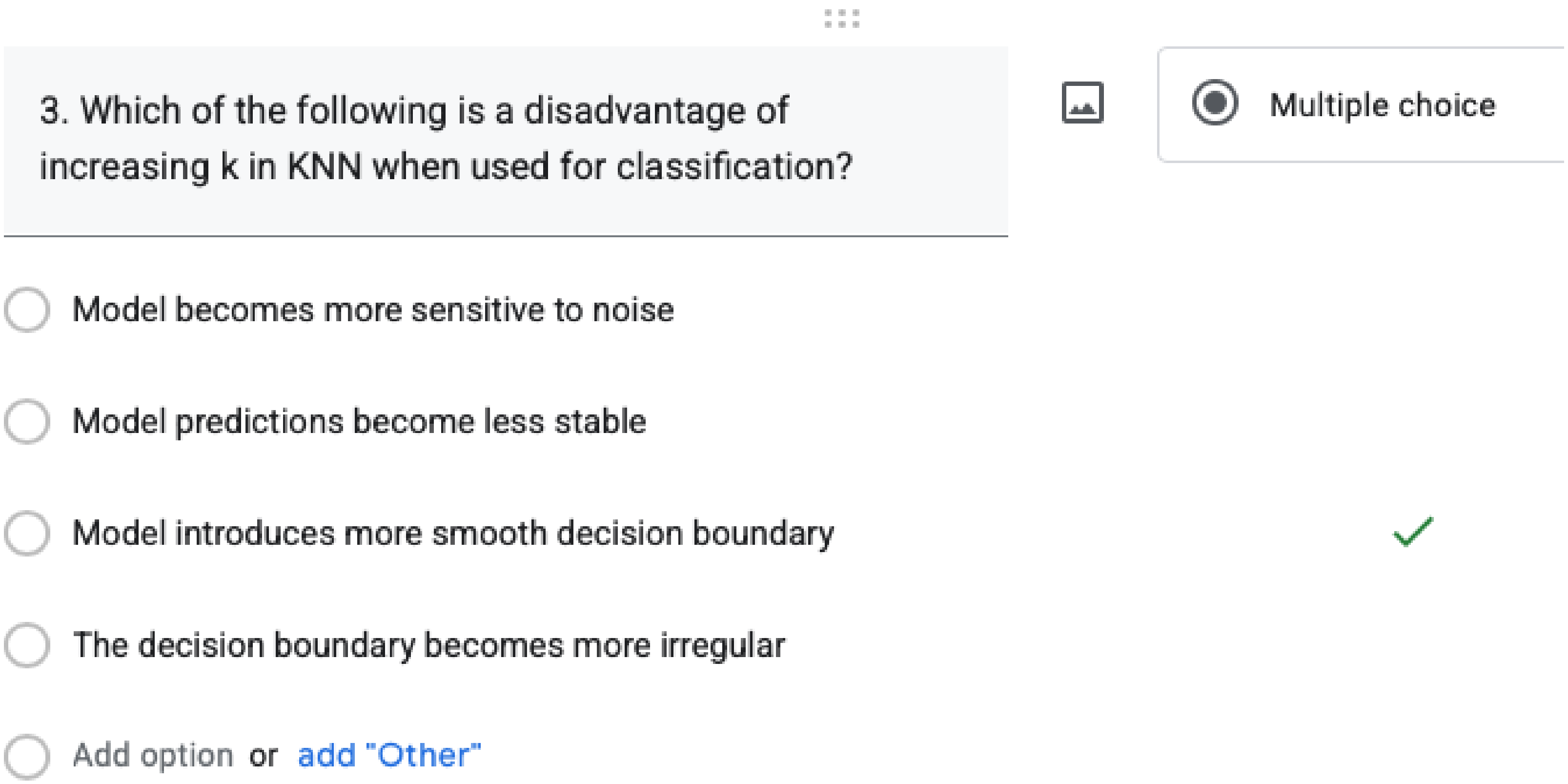
○ No effect, because KNN is scale-invariant

○ Distance measure may get dominated by the weight feature ✓

○ The algorithm will automatically adjust feature ranges

○ Model complexity will increase

Q5



3. Which of the following is a disadvantage of increasing k in KNN when used for classification?

● Multiple choice

○ Model becomes more sensitive to noise

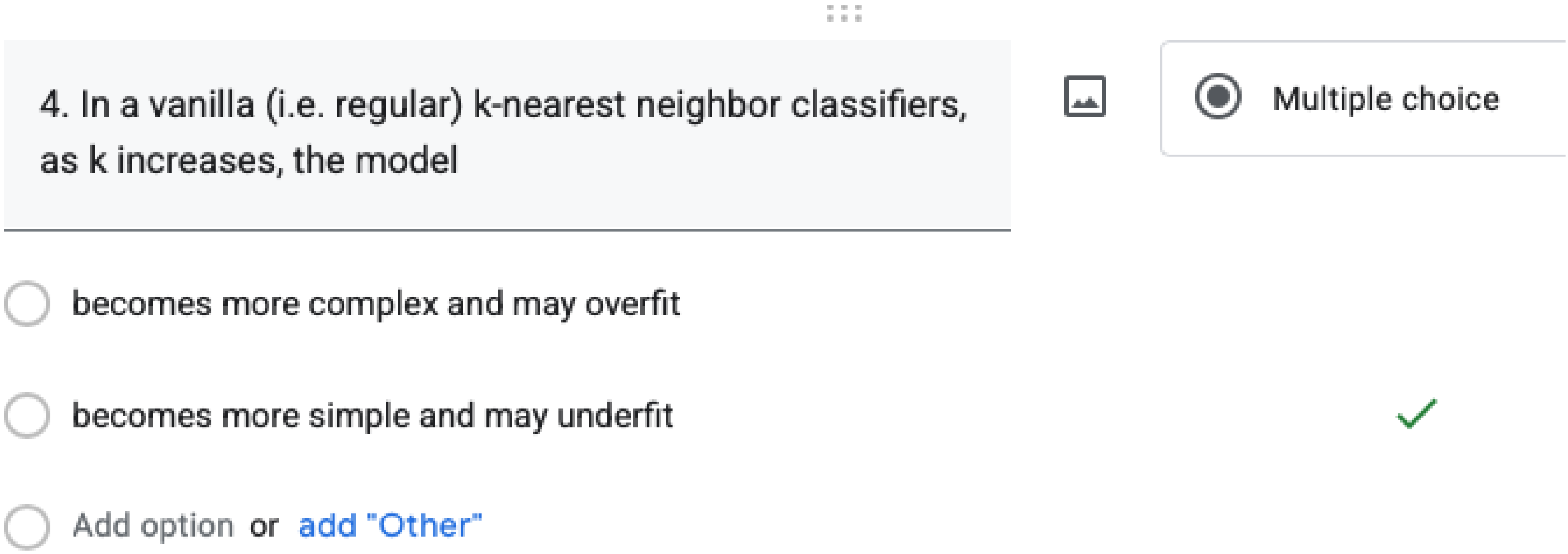
○ Model predictions become less stable

○ Model introduces more smooth decision boundary ✓

○ The decision boundary becomes more irregular

○ Add option or add "Other"

Q5



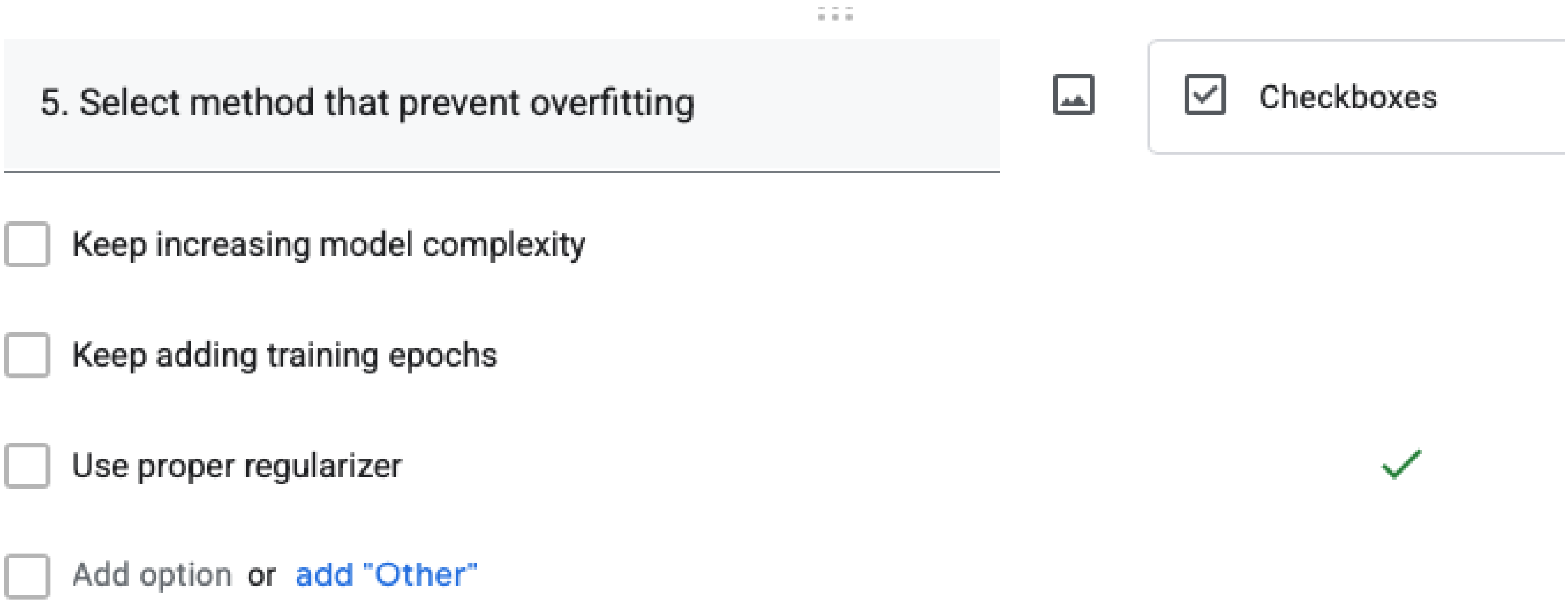
4. In a vanilla (i.e. regular) k-nearest neighbor classifiers, as k increases, the model

● Multiple choice

○ becomes more complex and may overfit

○ becomes more simple and may underfit ✓

Q5



5. Select method that prevent overfitting

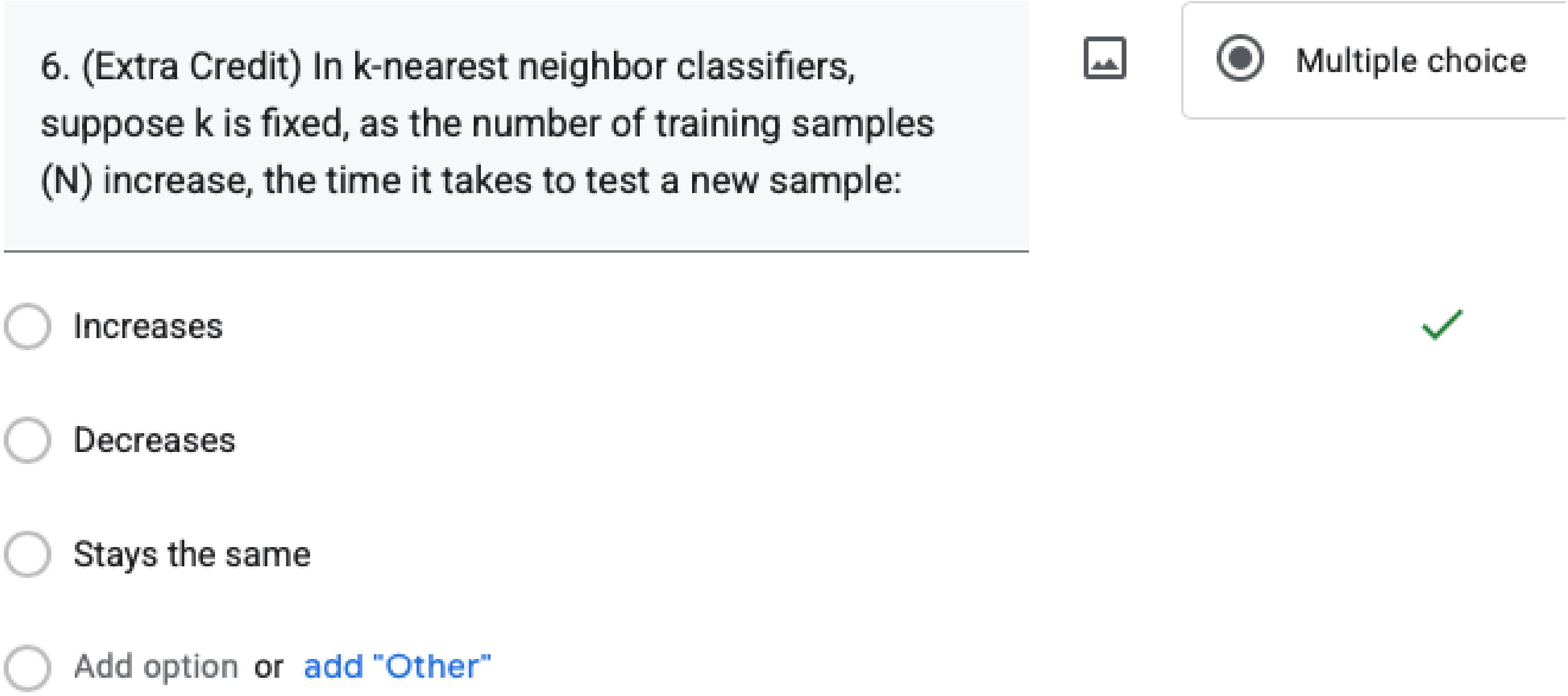
□ Keep increasing model complexity

□ Keep adding training epochs

□ Use proper regularizer ✓

□ Add option or add "Other"

Q5



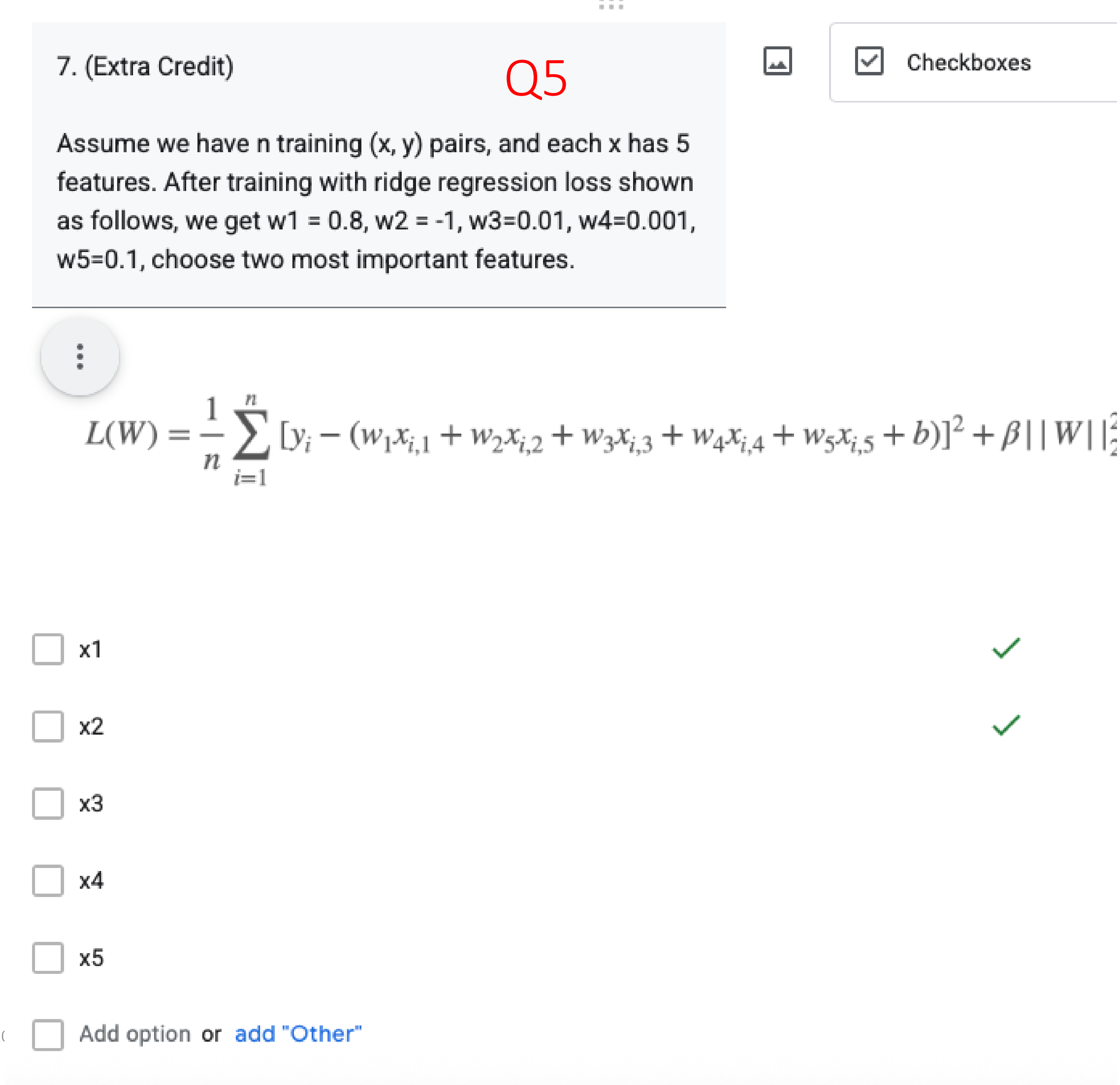
6. (Extra Credit) In k-nearest neighbor classifiers, suppose k is fixed, as the number of training samples (N) increase, the time it takes to test a new sample:

● Multiple choice

○ Increases ✓

○ Decreases

○ Stays the same



7. (Extra Credit)

Q5

☑ Checkboxes

Assume we have n training (x, y) pairs, and each x has 5 features. After training with ridge regression loss shown as follows, we get w1 = 0.8, w2 = -1, w3=0.01, w4=0.001, w5=0.1, choose two most important features.

$$L(W) = \frac{1}{n}\sum_{i=1}^{n}[y_i - (w_1x_{i,1} + w_2x_{i,2} + w_3x_{i,3} + w_4x_{i,4} + w_5x_{i,5} + b)]^2 + \beta||W||_2^2$$

□ x1 ✓

□ x2 ✓

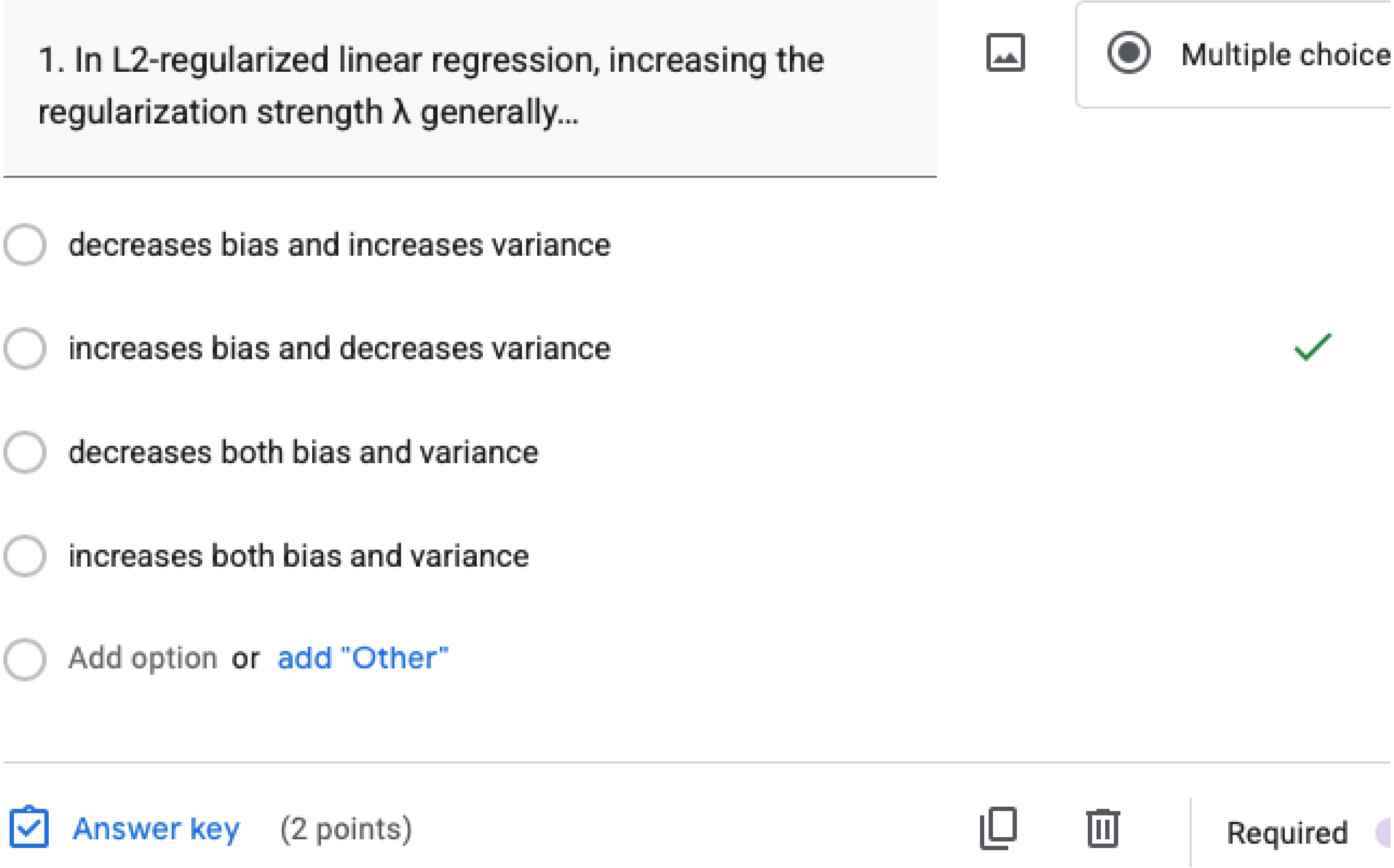
□ x3

□ x4

□ x5

□ Add option or add "Other"

Q6



1. In L2-regularized linear regression, increasing the regularization strength λ generally...

● Multiple choice

○ decreases bias and increases variance

○ increases bias and decreases variance ✓

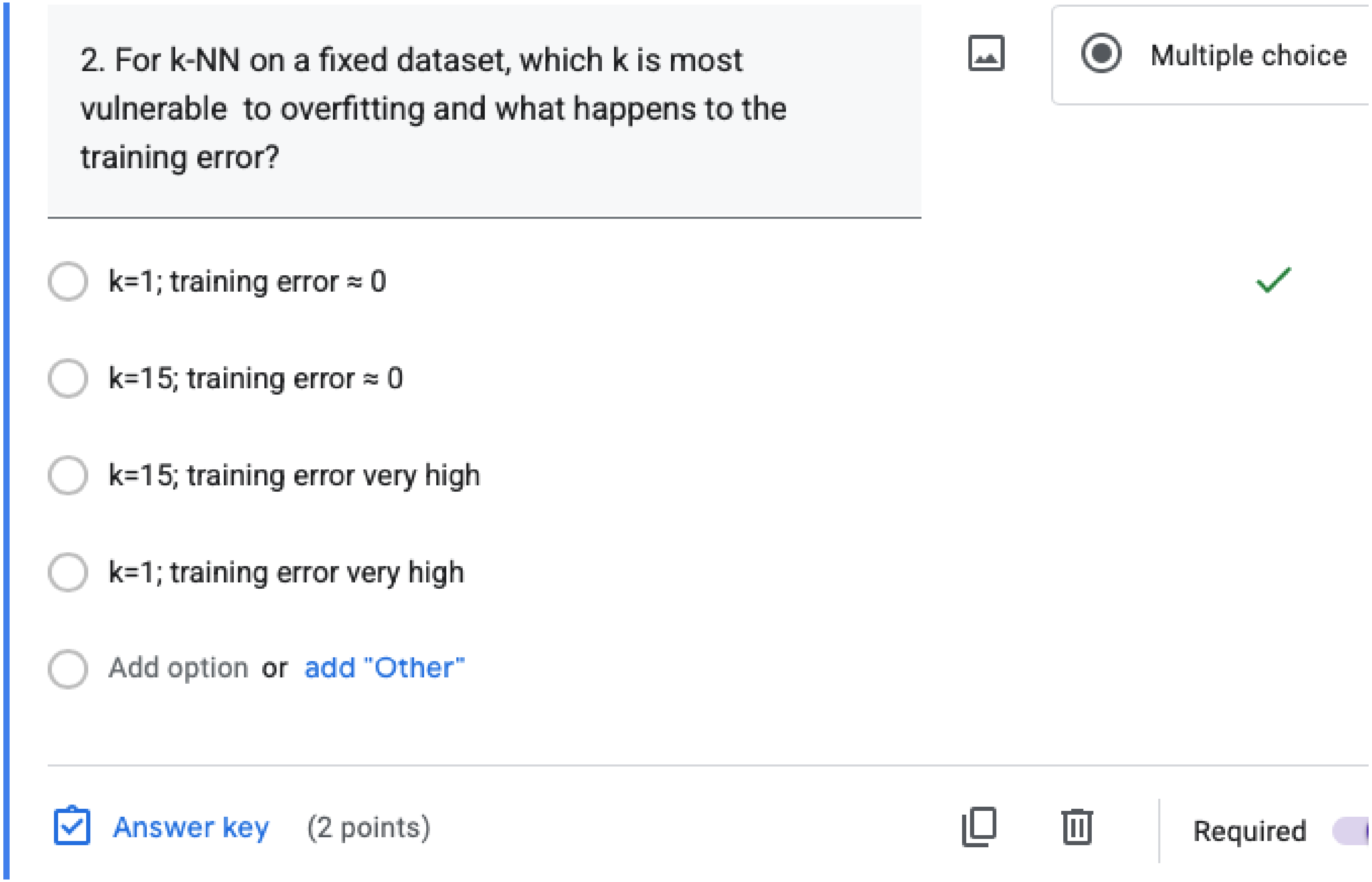
○ decreases both bias and variance

○ increases both bias and variance

○ Add option or add "Other"

☑ Answer key (2 points)

Required



2. For k-NN on a fixed dataset, which k is most vulnerable to overfitting and what happens to the training error?

● Multiple choice

○ k=1; training error ≈ 0 ✓

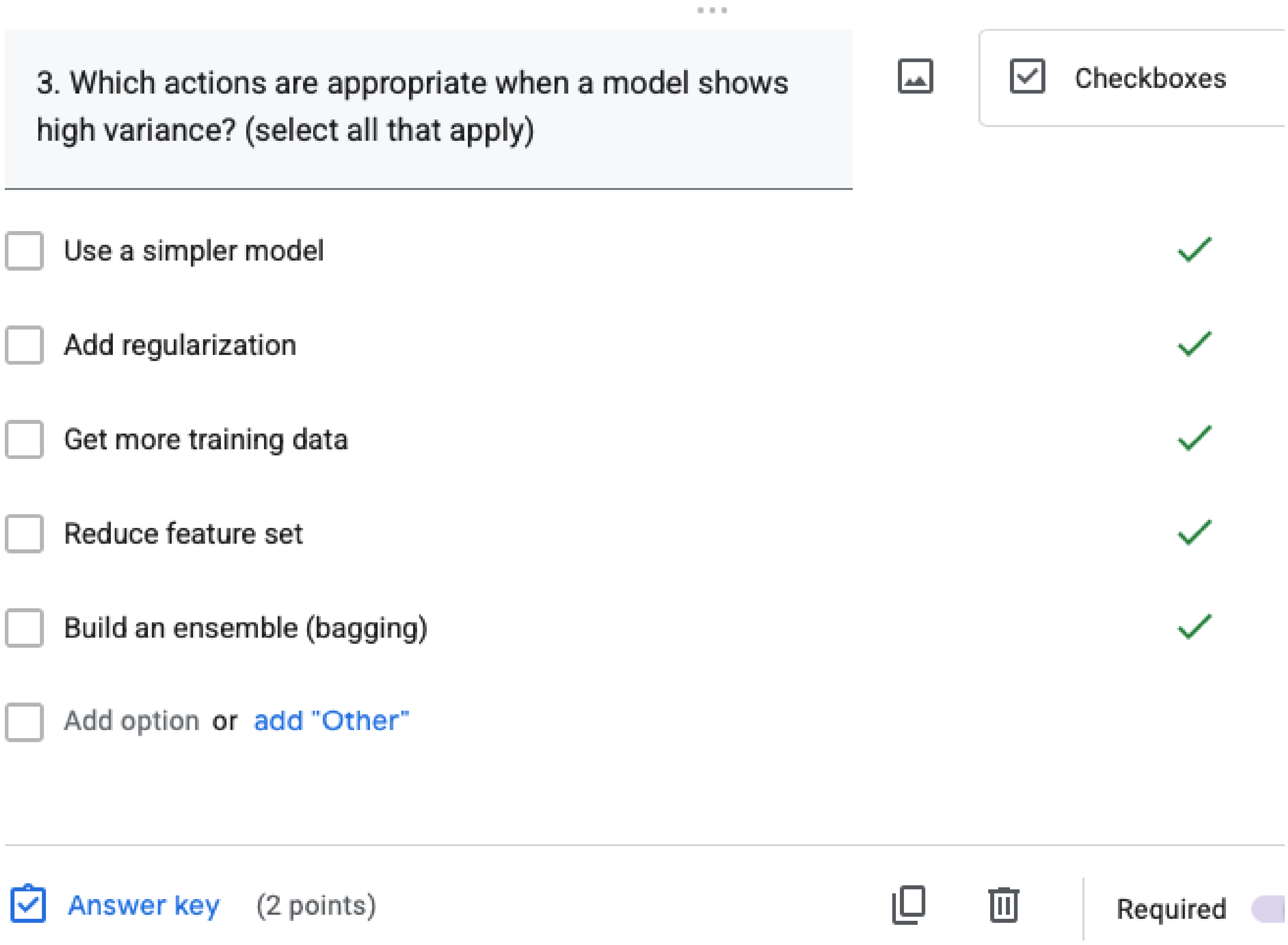
○ k=15; training error ≈ 0

○ k=15; training error very high

○ k=1; training error very high

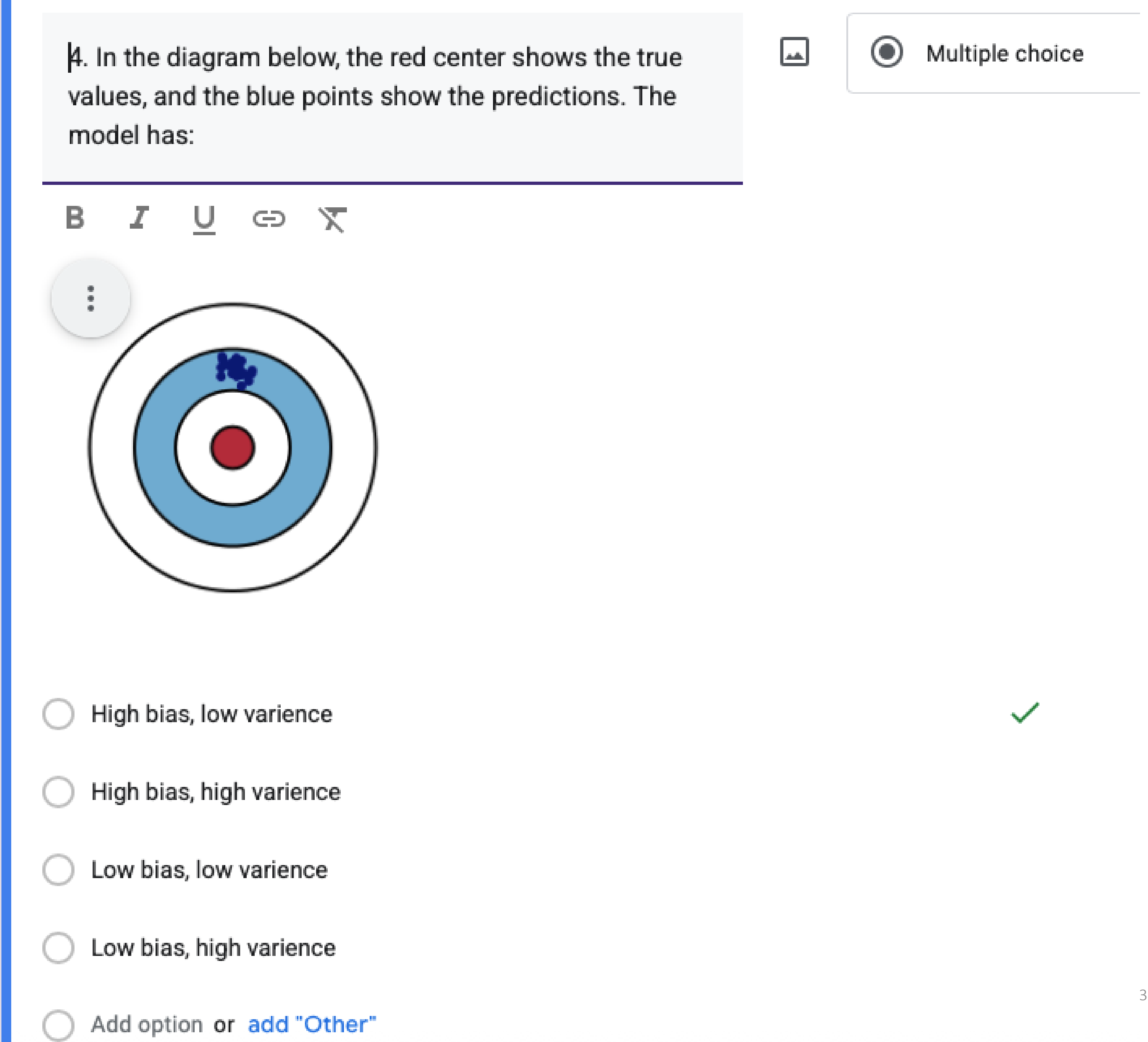
☑ Answer key (2 points)

Required



3. Which actions are appropriate when a model shows high variance? (select all that apply)

- □ Use a simpler model ✓
- □ Add regularization ✓
- □ Get more training data ✓
- □ Reduce feature set ✓
- □ Build an ensemble (bagging) ✓
- □ Add option or add "Other"



4. In the diagram below, the red center shows the true values, and the blue points show the predictions. The model has:

● Multiple choice

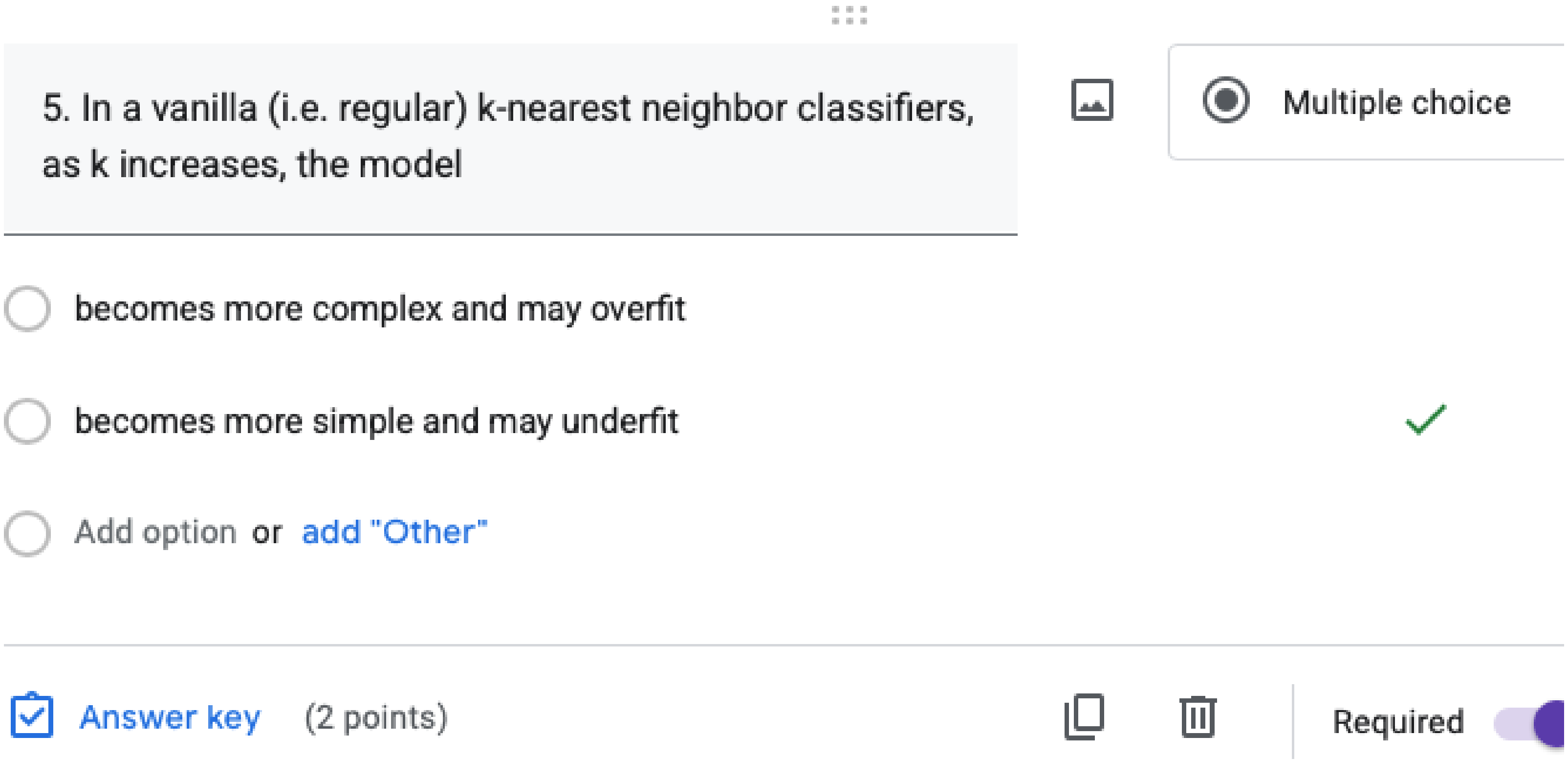
○ High bias, low varience ✓

○ High bias, high varience

○ Low bias, low varience

○ Low bias, high varience

○ Add option or add "Other"

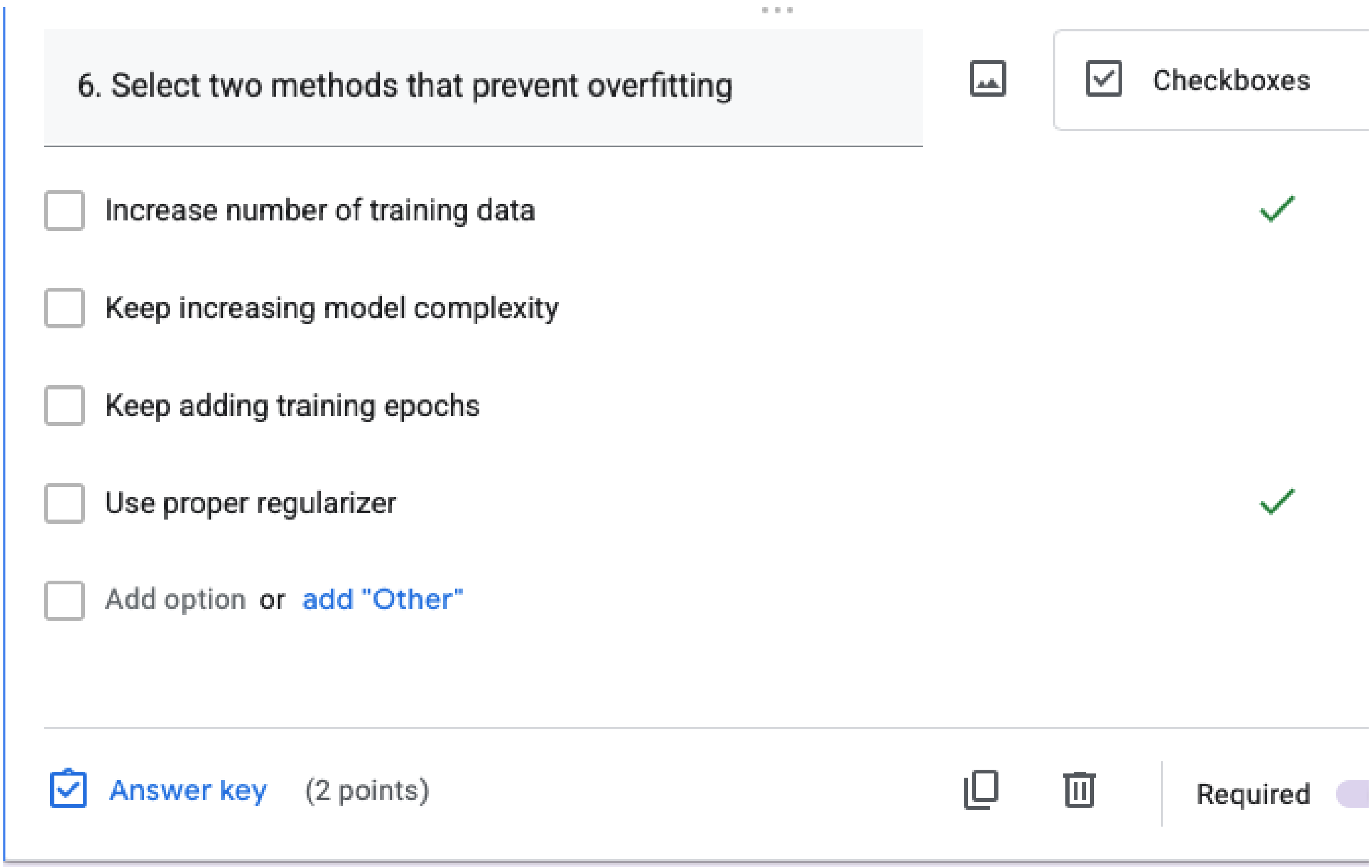


5. In a vanilla (i.e. regular) k-nearest neighbor classifiers, as k increases, the model

○ becomes more complex and may overfit

○ becomes more simple and may underfit ✓

○ Add option or add "Other"



6. Select two methods that prevent overfitting

☑ Checkboxes

□ Increase number of training data ✓

□ Keep increasing model complexity

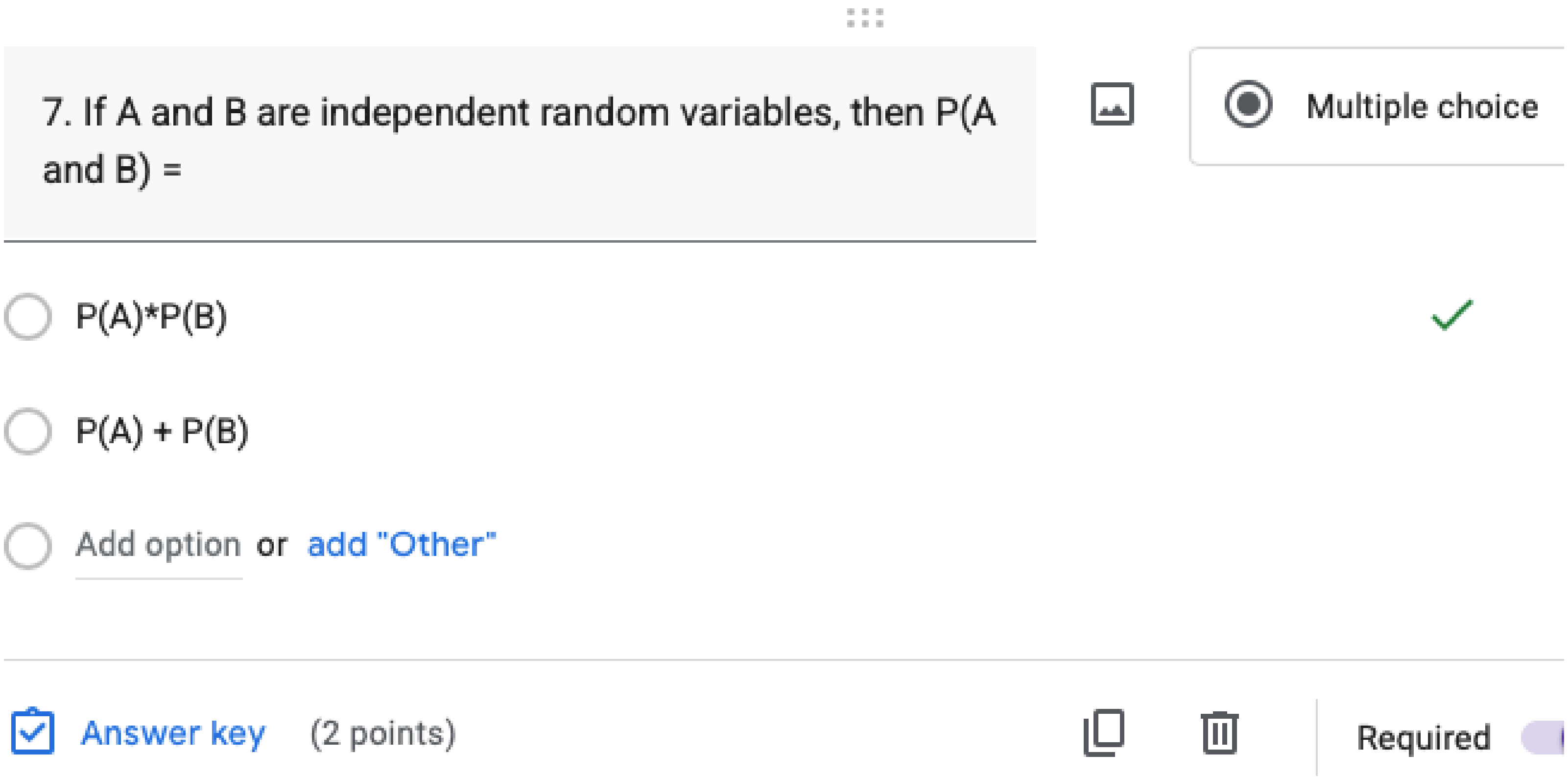
□ Keep adding training epochs

□ Use proper regularizer ✓

□ Add option or add "Other"

☑ Answer key (2 points)

Required

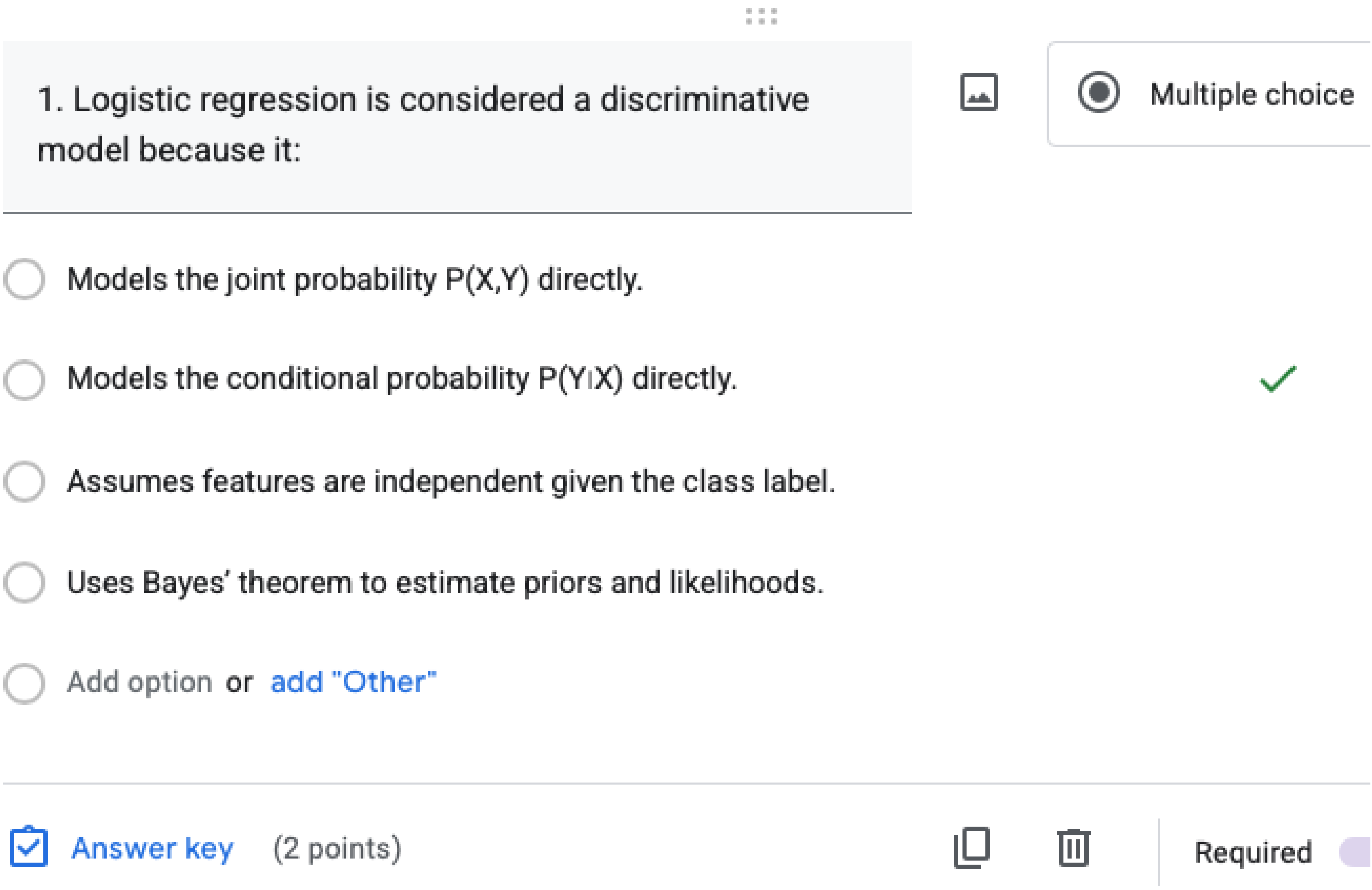


7. If A and B are independent random variables, then P(A and B) =

○ P(A)*P(B) ✓

○ P(A) + P(B)

Q7



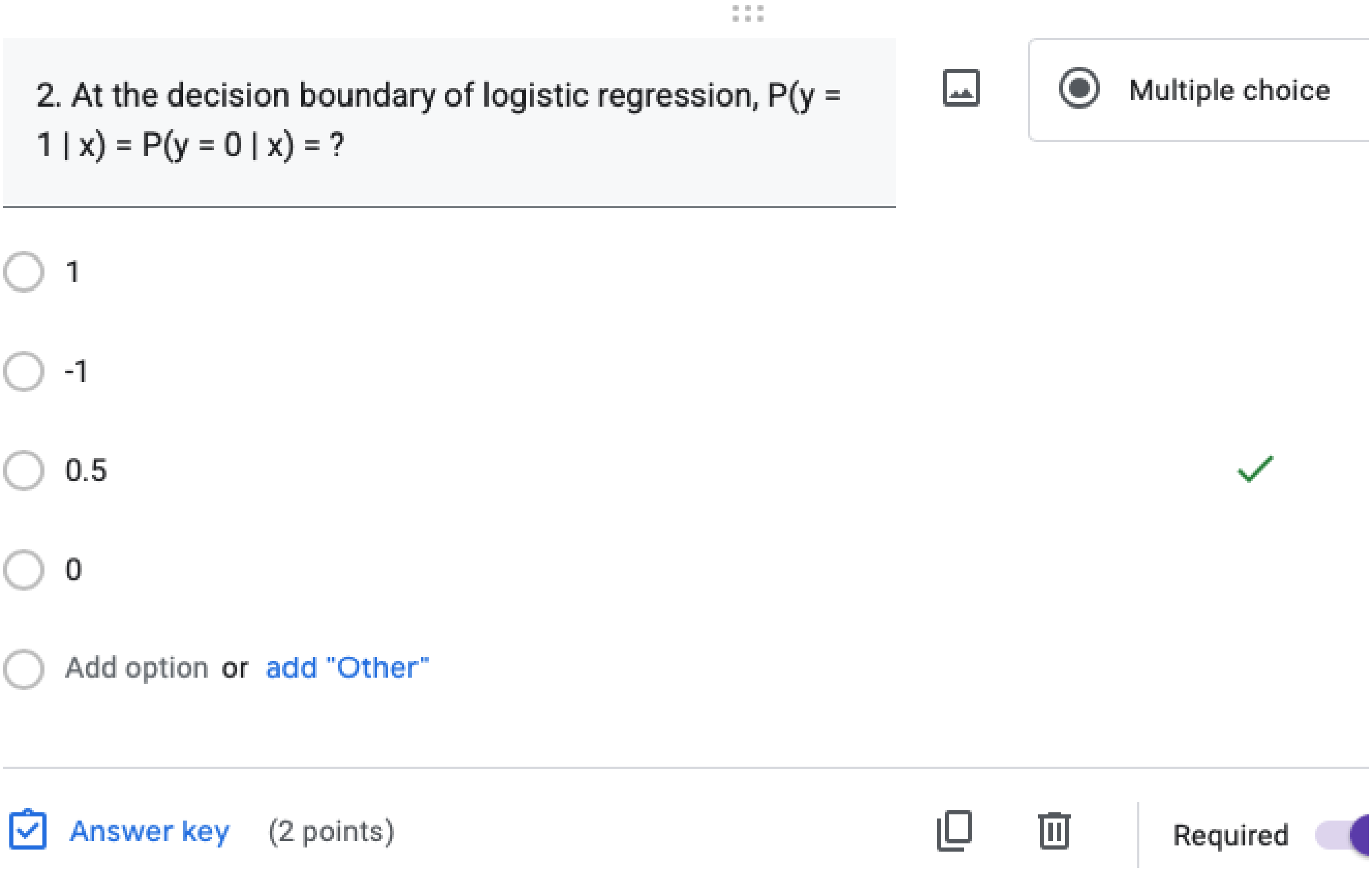
1. Logistic regression is considered a discriminative model because it:

○ Models the joint probability P(X,Y) directly.

○ Models the conditional probability P(Y|X) directly. ✓

○ Assumes features are independent given the class label.

○ Uses Bayes' theorem to estimate priors and likelihoods.



2. At the decision boundary of logistic regression, P(y = 1 | x) = P(y = 0 | x) = ?

○ 1

○ -1

○ 0.5 ✓

○ 0

(2 points)

3. True or False? Logistic regression models have a linear decision boundary.

● Multiple choice

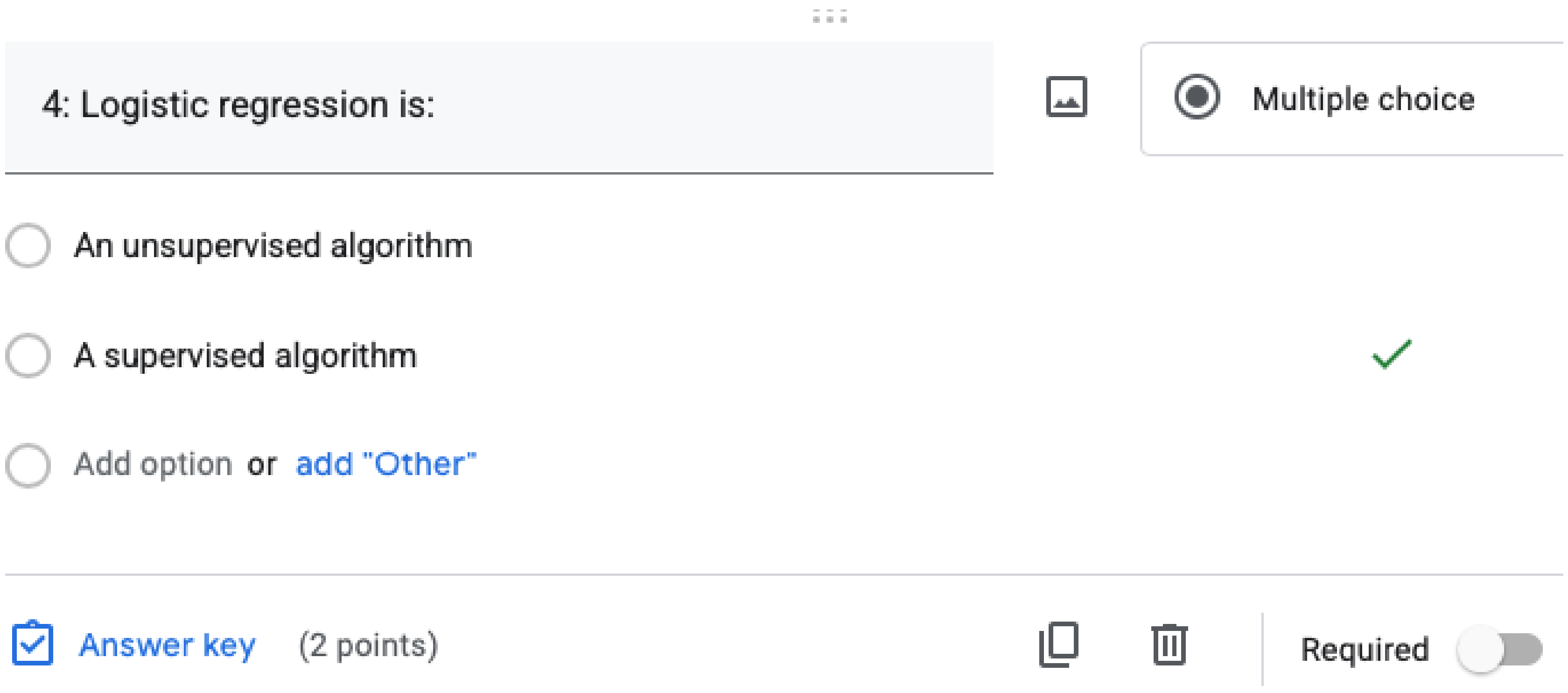
○ True ✓

○ False

○ Add option or add "Other"

☑ Answer key (2 points)

Required



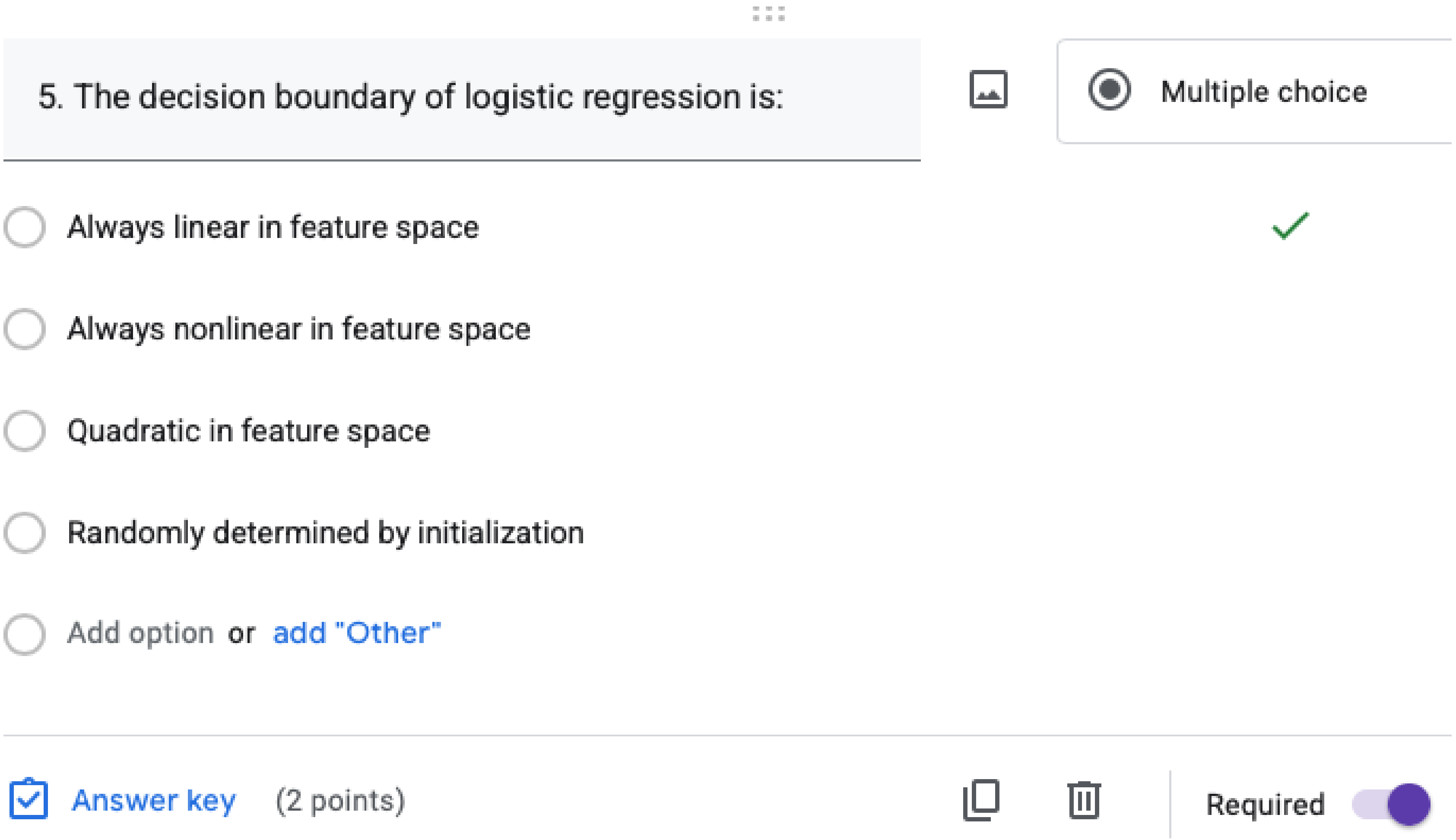
4: Logistic regression is:

● Multiple choice

○ An unsupervised algorithm

○ A supervised algorithm ✓

○ Add option or add "Other"



5. The decision boundary of logistic regression is:

● Multiple choice

○ Always linear in feature space ✓

○ Always nonlinear in feature space

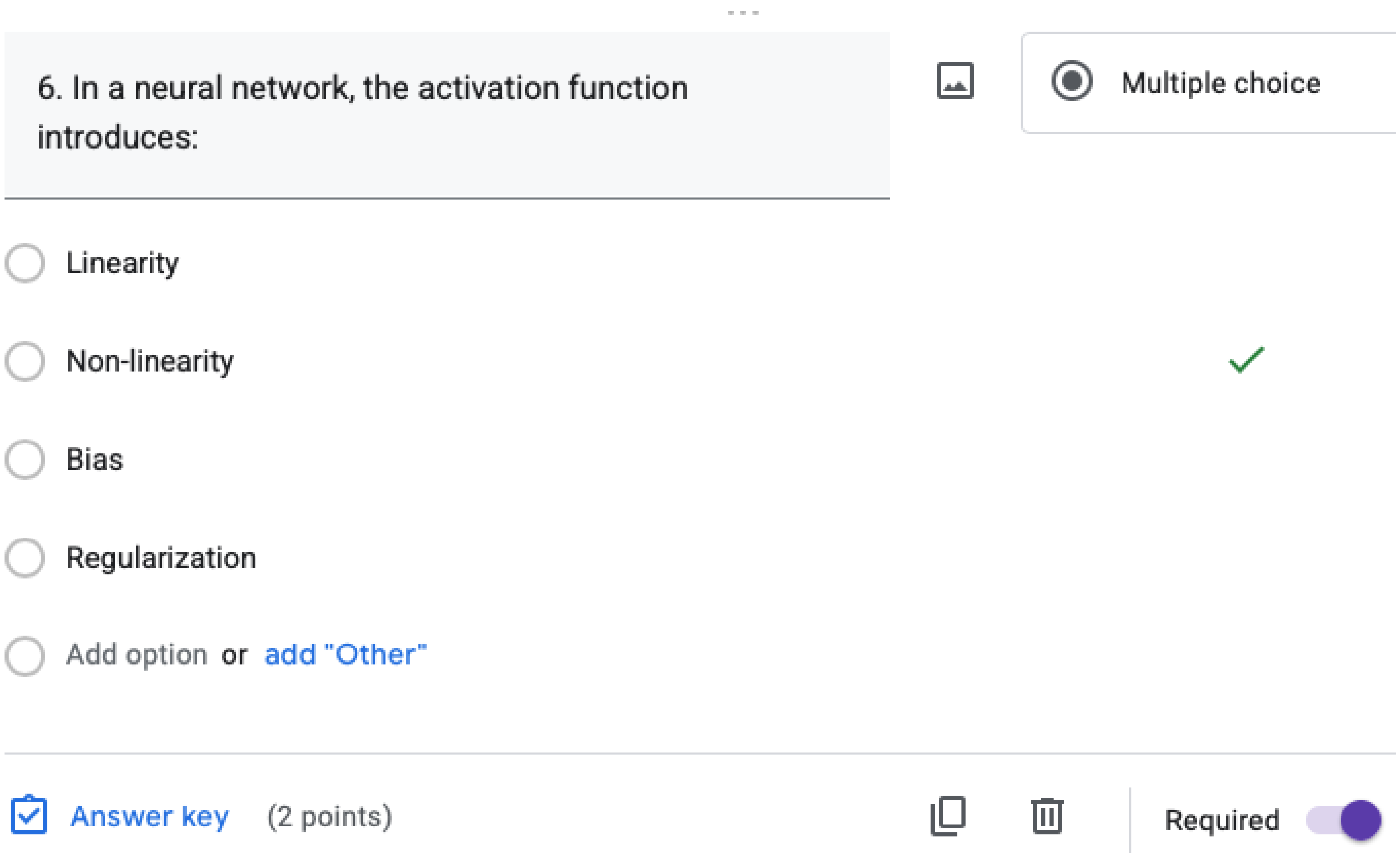
○ Quadratic in feature space

○ Randomly determined by initialization

○ Add option or add "Other"

☑ Answer key (2 points)

Required



● Multiple choice

6. In a neural network, the activation function introduces:

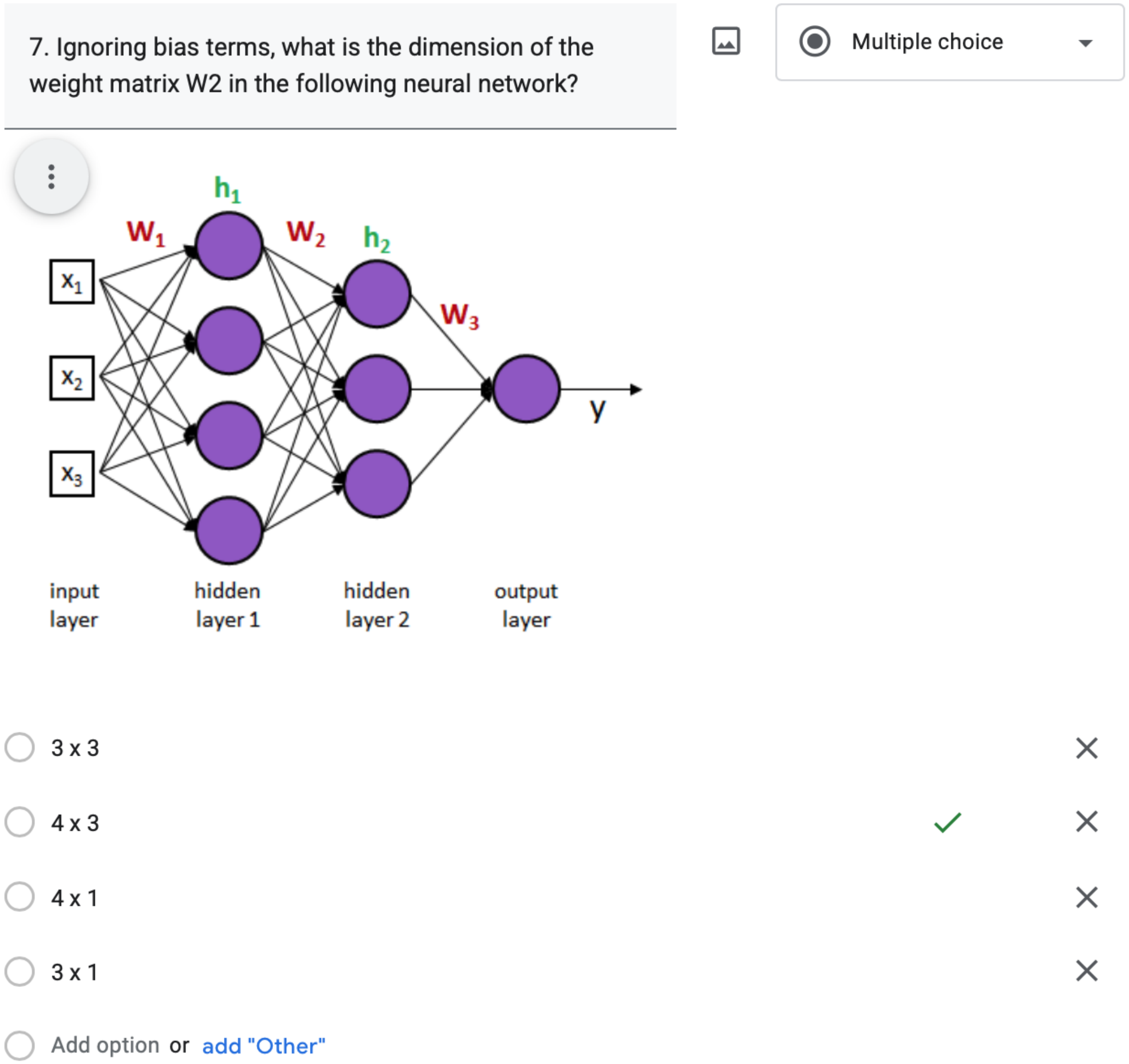
○ Linearity

○ Non-linearity ✓

○ Bias

○ Regularization

○ Add option or add "Other"



7. Ignoring bias terms, what is the dimension of the weight matrix W2 in the following neural network?

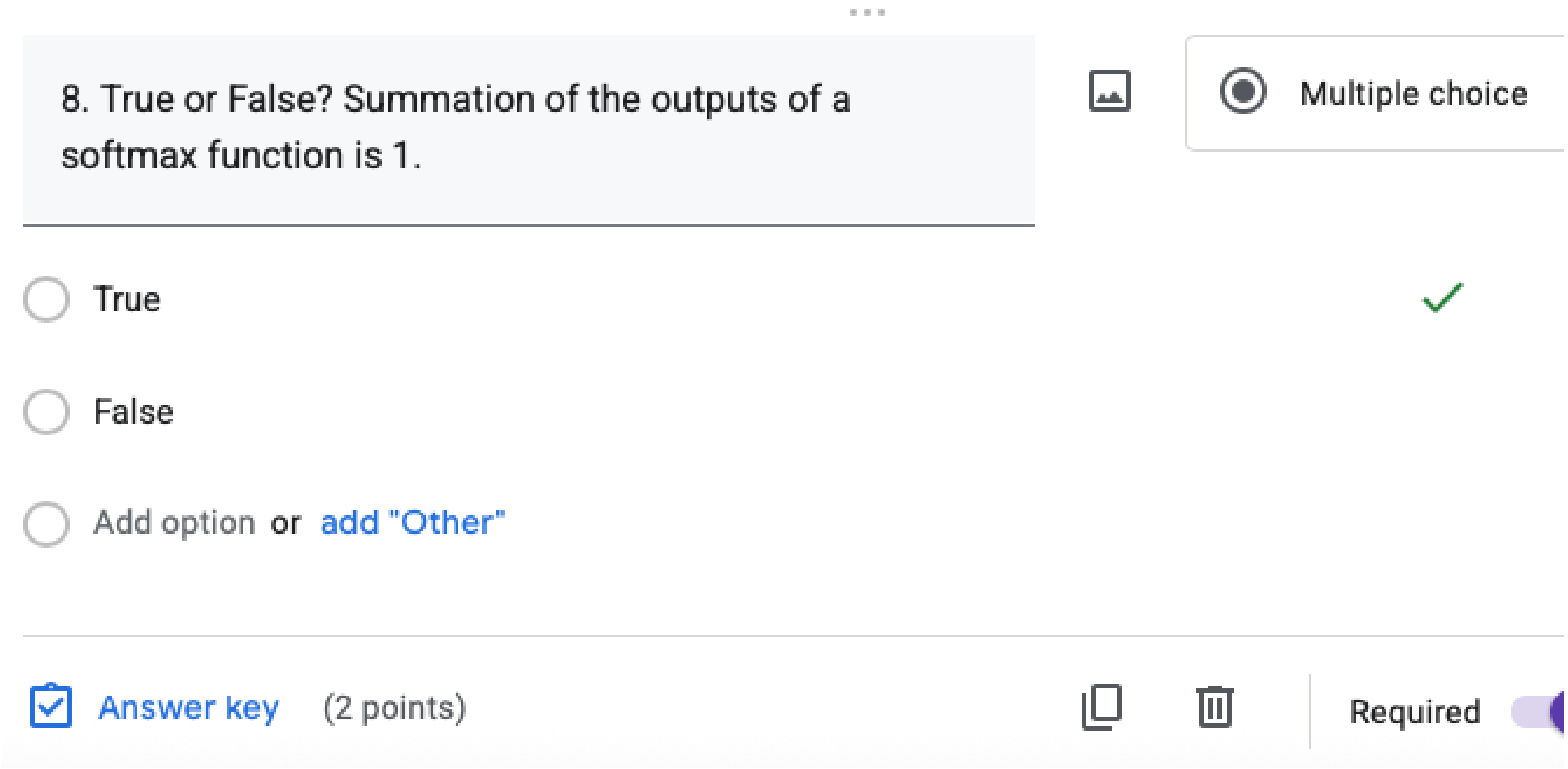
○ 3 x 3 ✕

○ 4 x 3 ✓ ✕

○ 4 x 1 ✕

○ 3 x 1 ✕

○ Add option or add "Other"



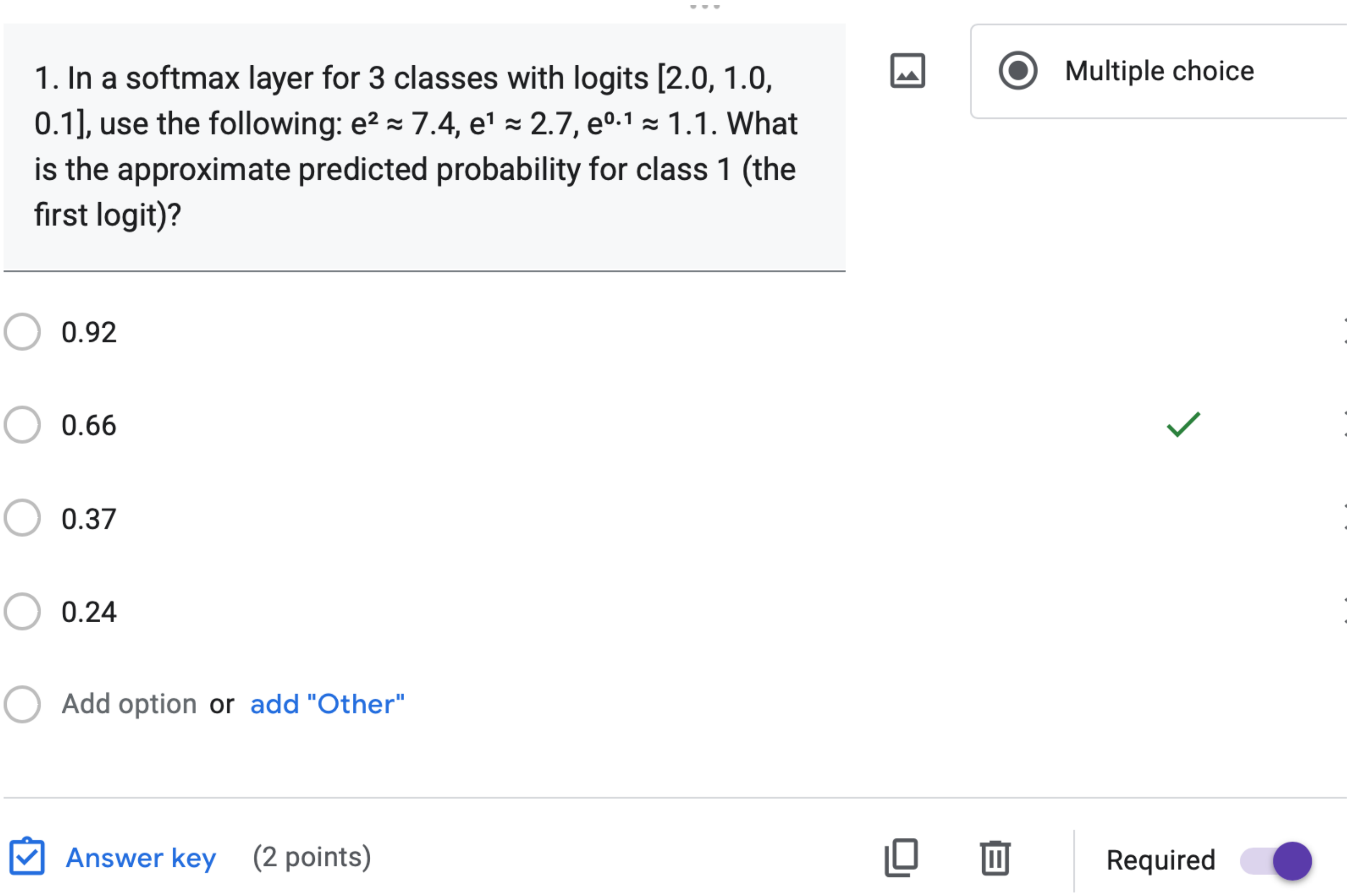
8. True or False? Summation of the outputs of a softmax function is 1.

● Multiple choice

○ True ✓

○ False

Q8



1. In a softmax layer for 3 classes with logits [2.0, 1.0, 0.1], use the following: $e^2 \approx 7.4$, $e^1 \approx 2.7$, $e^{0.1} \approx 1.1$. What is the approximate predicted probability for class 1 (the first logit)?

● Multiple choice

○ 0.92

○ 0.66 ✓

○ 0.37

○ 0.24

○ Add option or add "Other"

2. Which property of images allows CNNs to reuse the same weights across different regions?

● Multiple choice

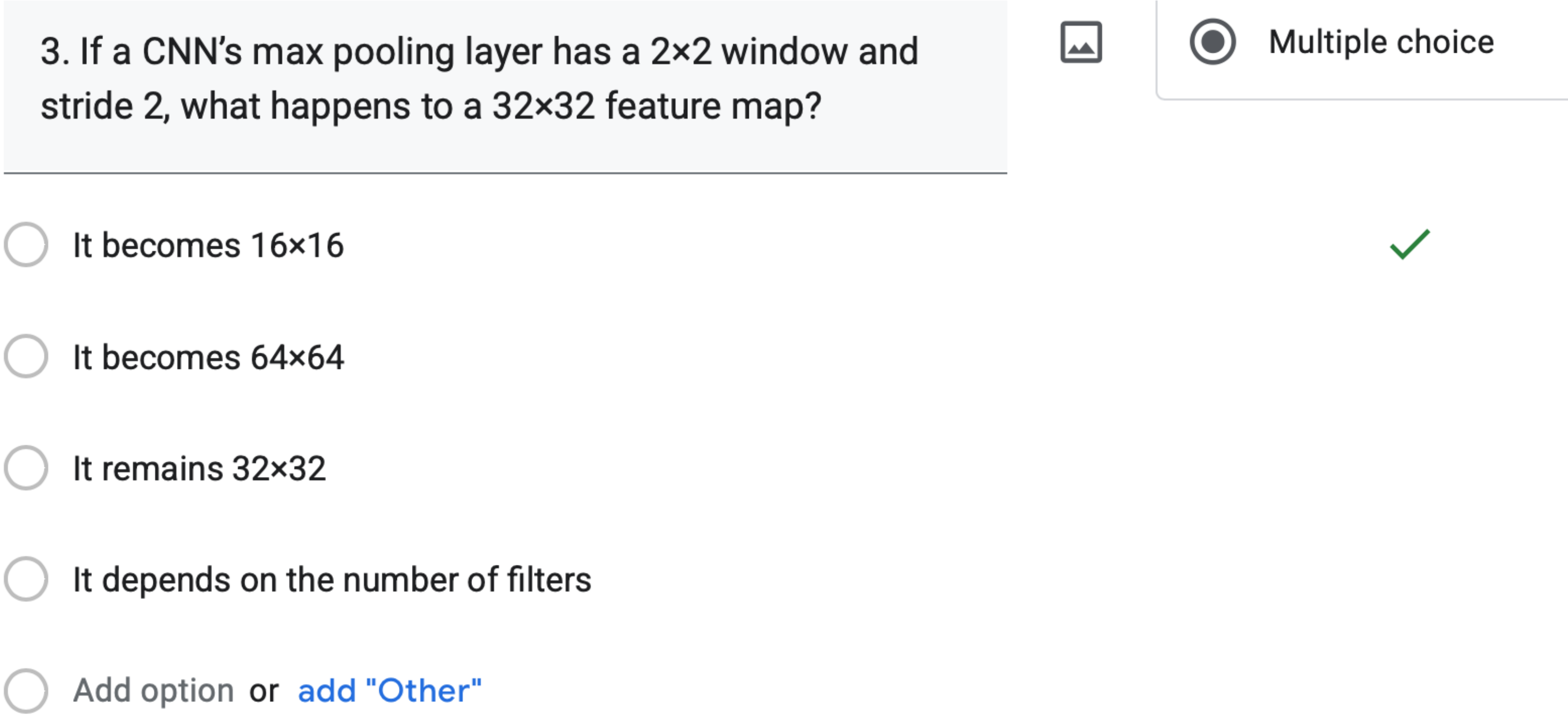
○ Manifold structure

○ Translation invariance ✓

○ Subsampling

○ Dimensionality reduction

○ Add option or add "Other"



3. If a CNN's max pooling layer has a 2×2 window and stride 2, what happens to a 32×32 feature map?

● Multiple choice

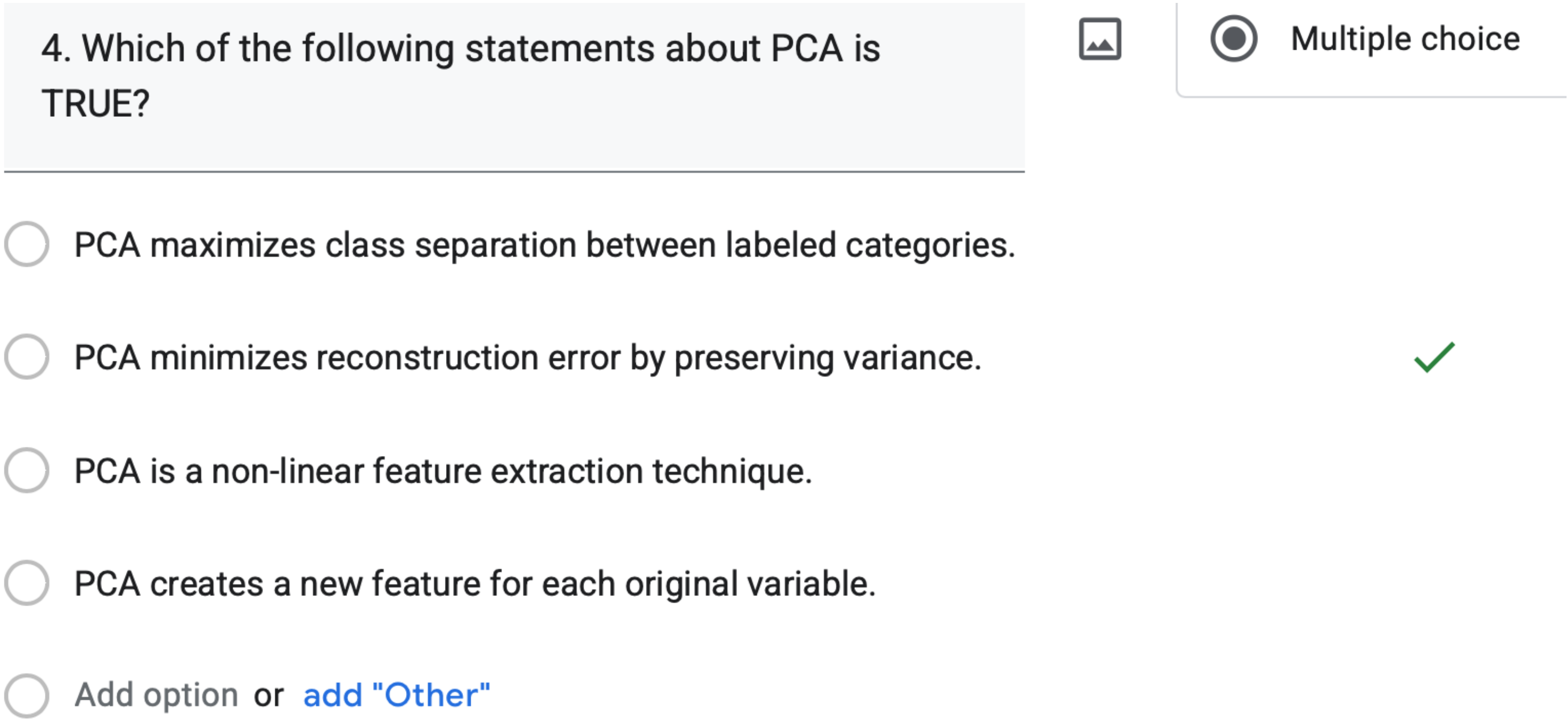
○ It becomes 16×16 ✓

○ It becomes 64×64

○ It remains 32×32

○ It depends on the number of filters

○ Add option or add "Other"



4. Which of the following statements about PCA is TRUE?

● Multiple choice

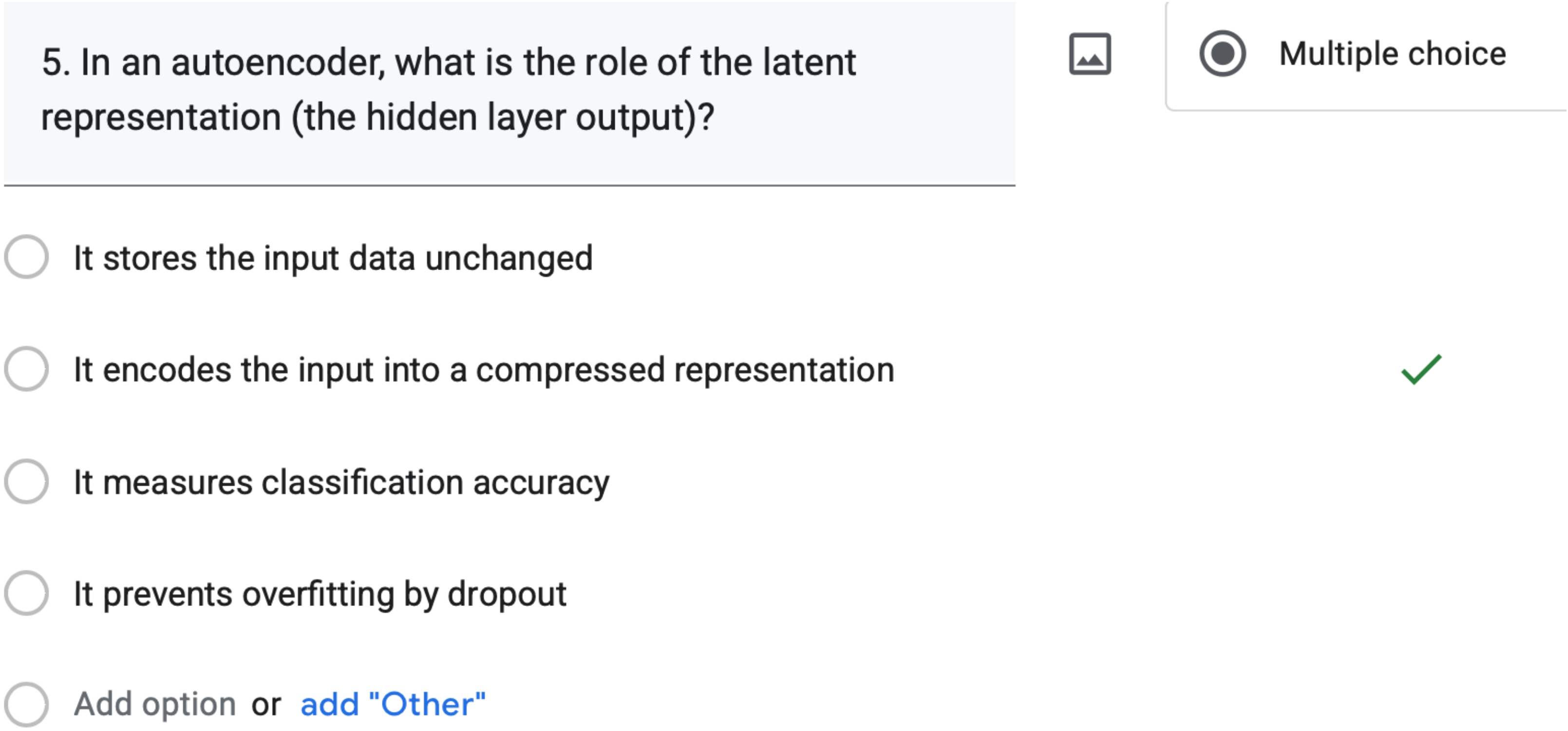
○ PCA maximizes class separation between labeled categories.

○ PCA minimizes reconstruction error by preserving variance. ✓

○ PCA is a non-linear feature extraction technique.

○ PCA creates a new feature for each original variable.

○ Add option or add "Other"



5. In an autoencoder, what is the role of the latent representation (the hidden layer output)?

● Multiple choice

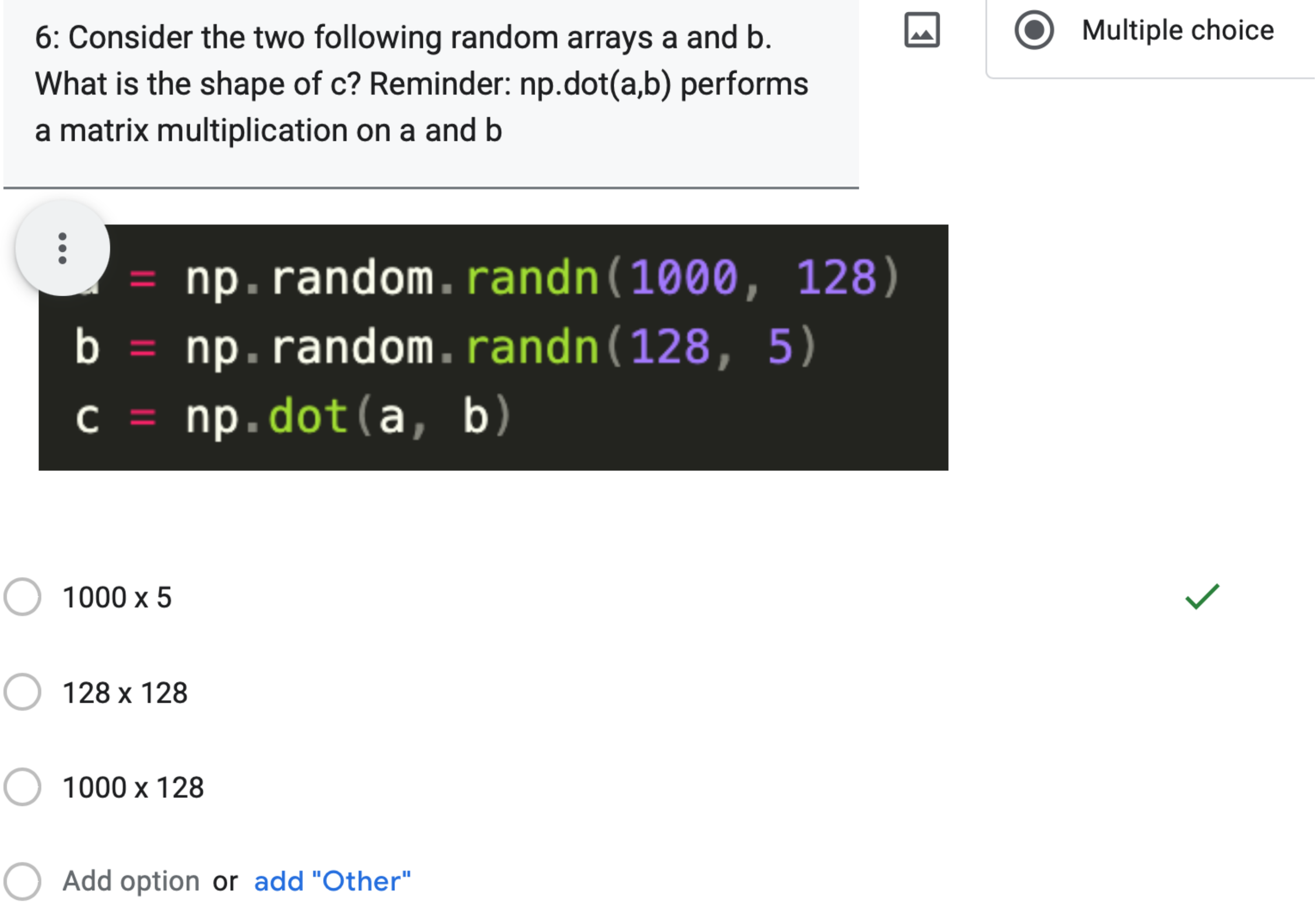
○ It stores the input data unchanged

○ It encodes the input into a compressed representation ✓

○ It measures classification accuracy

○ It prevents overfitting by dropout

○ Add option or add "Other"



6: Consider the two following random arrays a and b. What is the shape of c? Reminder: np.dot(a,b) performs a matrix multiplication on a and b

● Multiple choice

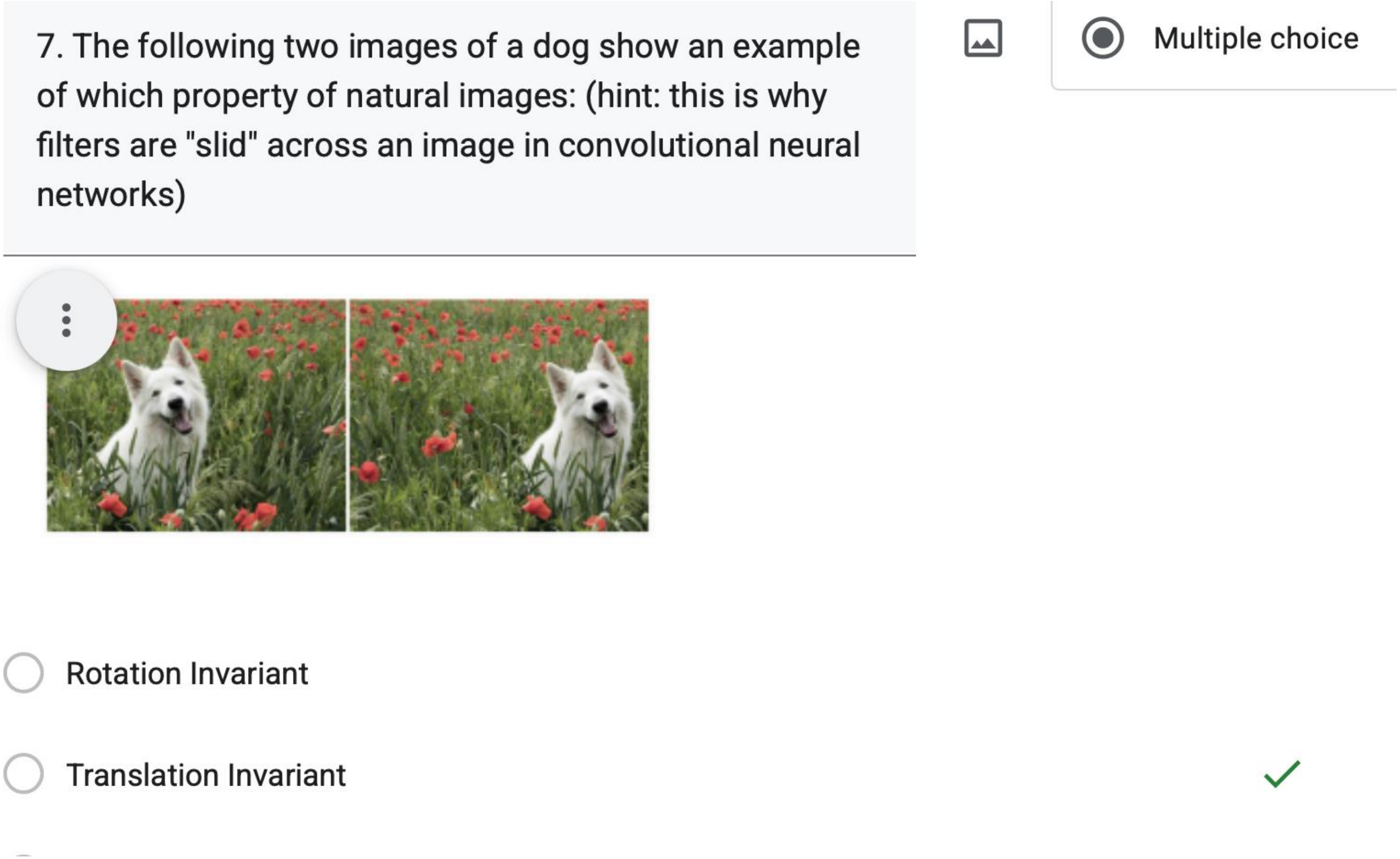
```
a = np.random.randn(1000, 128)
b = np.random.randn(128, 5)
c = np.dot(a, b)
```

○ 1000 x 5 ✓

○ 128 x 128

○ 1000 x 128

○ Add option or add "Other"

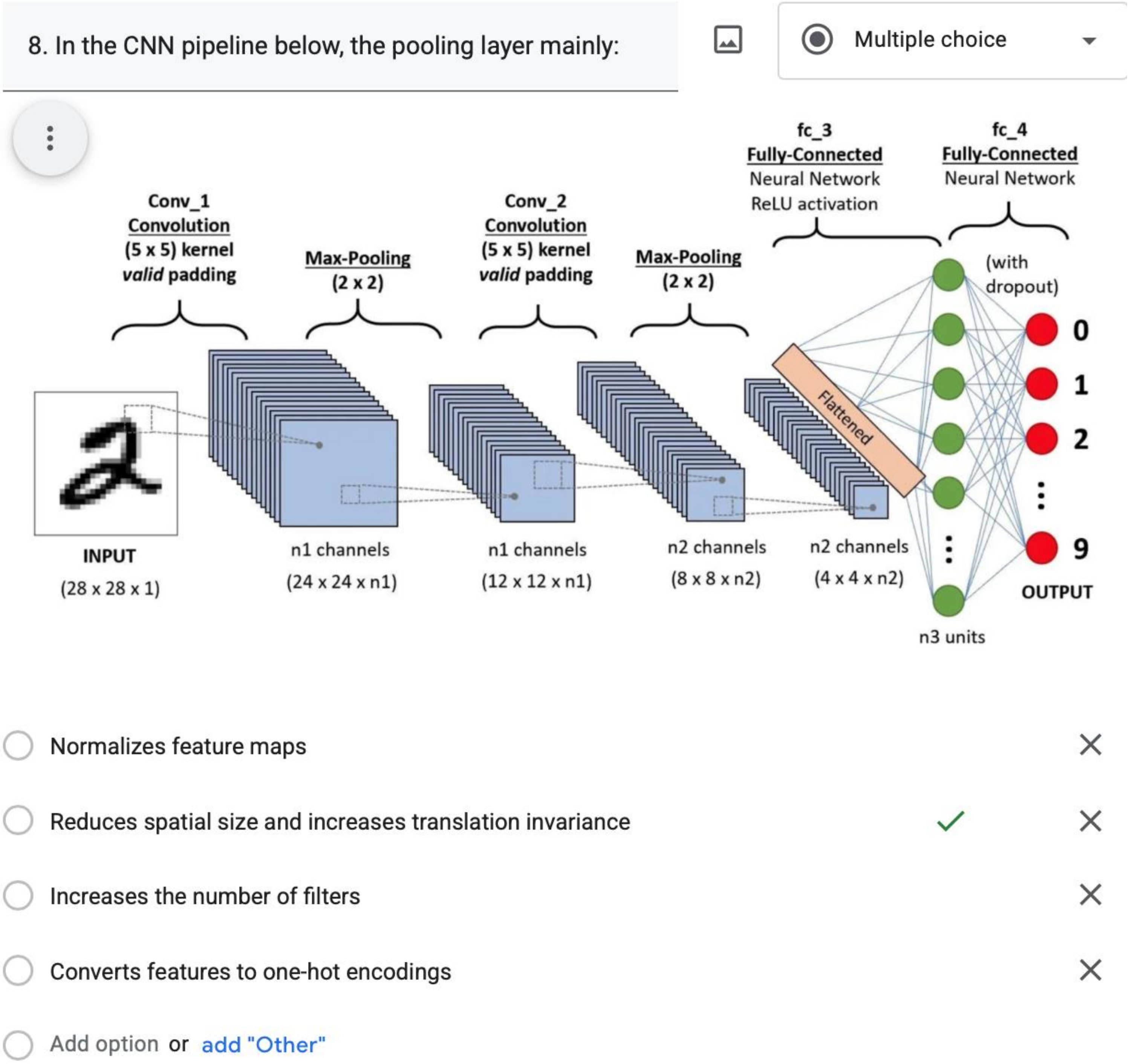


7. The following two images of a dog show an example of which property of natural images: (hint: this is why filters are "slid" across an image in convolutional neural networks)

● Multiple choice

○ Rotation Invariant

○ Translation Invariant ✓



8. In the CNN pipeline below, the pooling layer mainly:

○ Normalizes feature maps ×

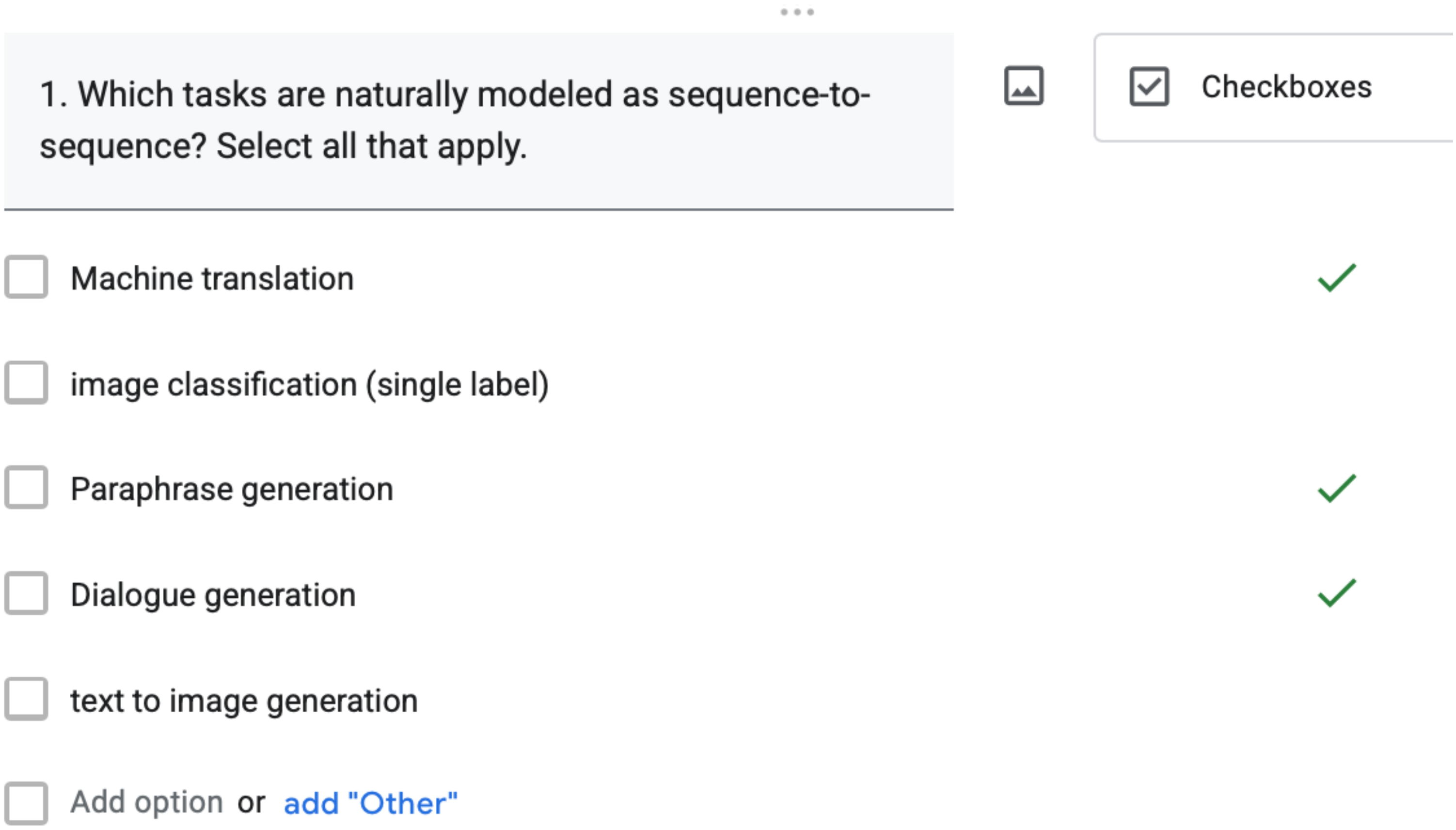
○ Reduces spatial size and increases translation invariance ✓ ×

○ Increases the number of filters ×

○ Converts features to one-hot encodings ×

○ Add option or add "Other"

Q9



1. Which tasks are naturally modeled as sequence-to-sequence? Select all that apply.

☑ Checkboxes

□ Machine translation ✓

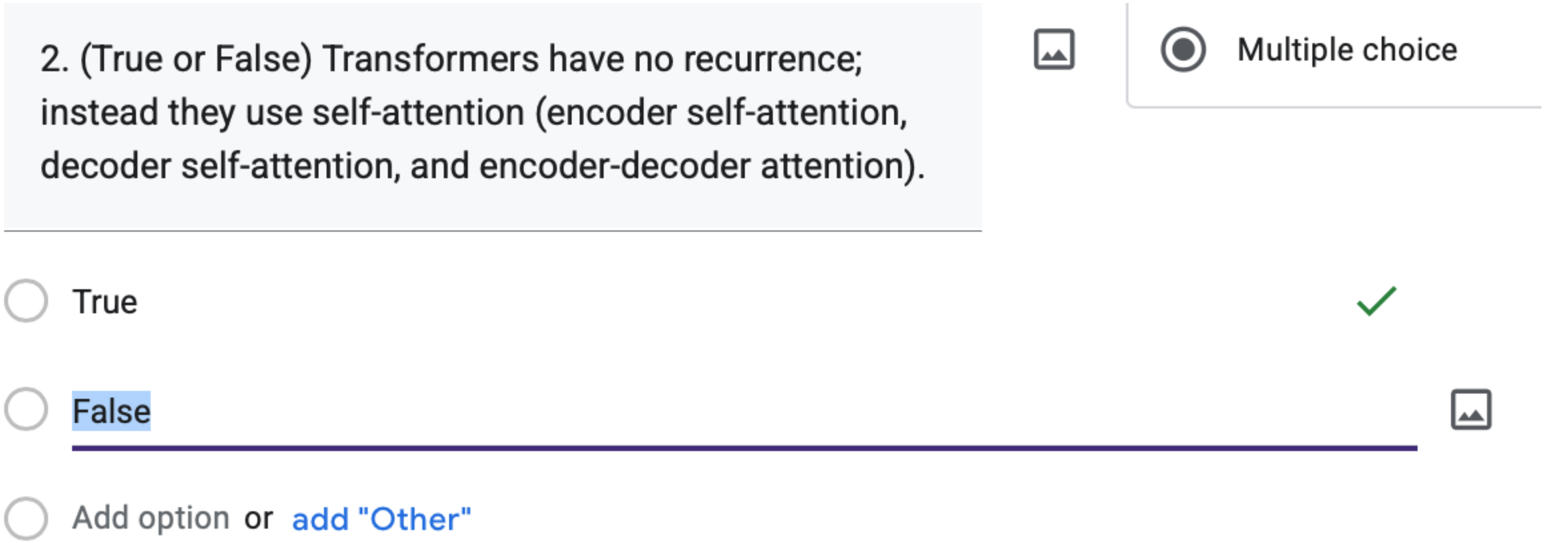
□ image classification (single label)

□ Paraphrase generation ✓

□ Dialogue generation ✓

□ text to image generation

□ Add option or add "Other"



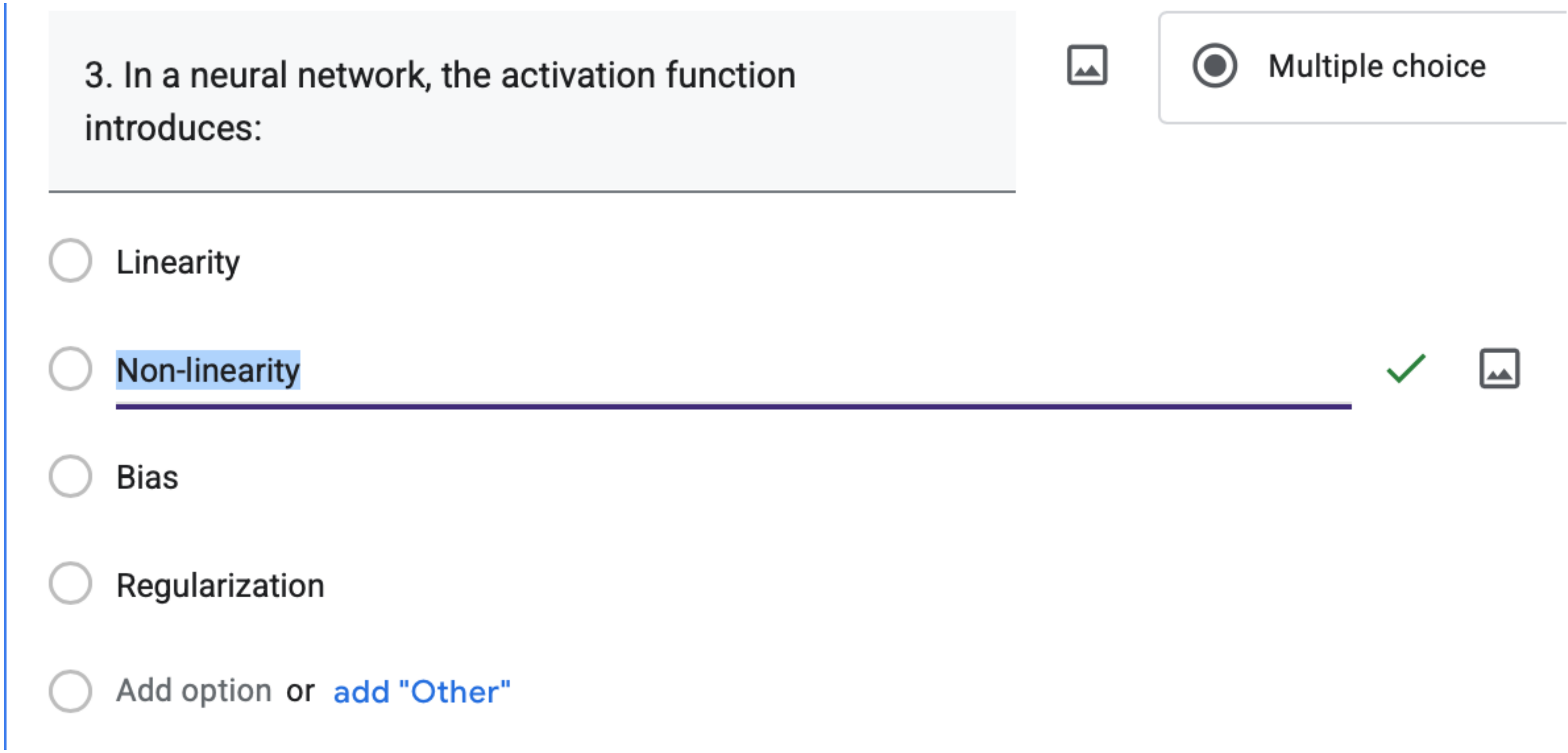
2. (True or False) Transformers have no recurrence; instead they use self-attention (encoder self-attention, decoder self-attention, and encoder-decoder attention).

● Multiple choice

○ True ✓

○ False

○ Add option or add "Other"



3. In a neural network, the activation function introduces:

● Multiple choice

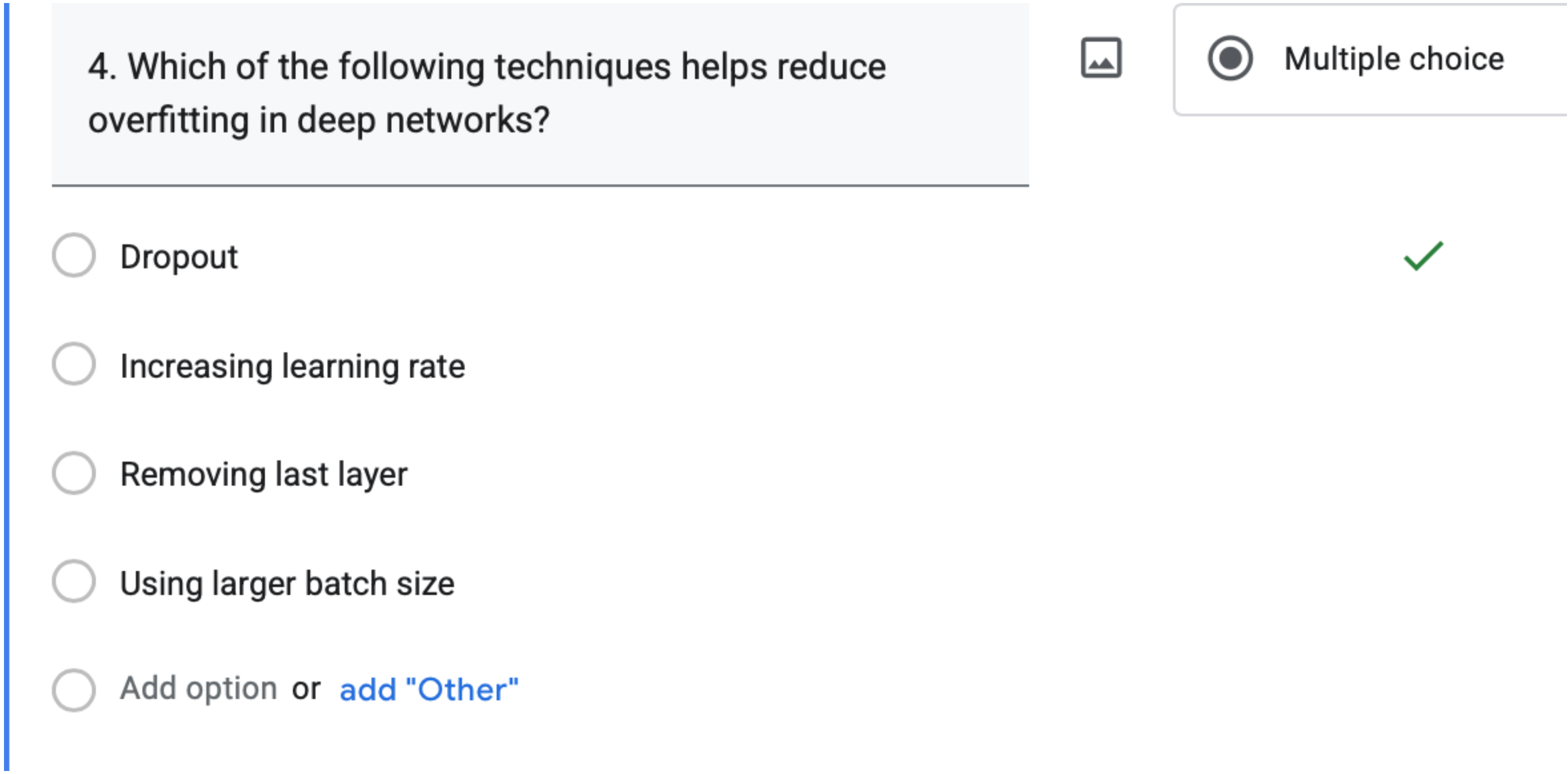
○ Linearity

○ Non-linearity ✓

○ Bias

○ Regularization

○ Add option or add "Other"



4. Which of the following techniques helps reduce overfitting in deep networks?

● Multiple choice

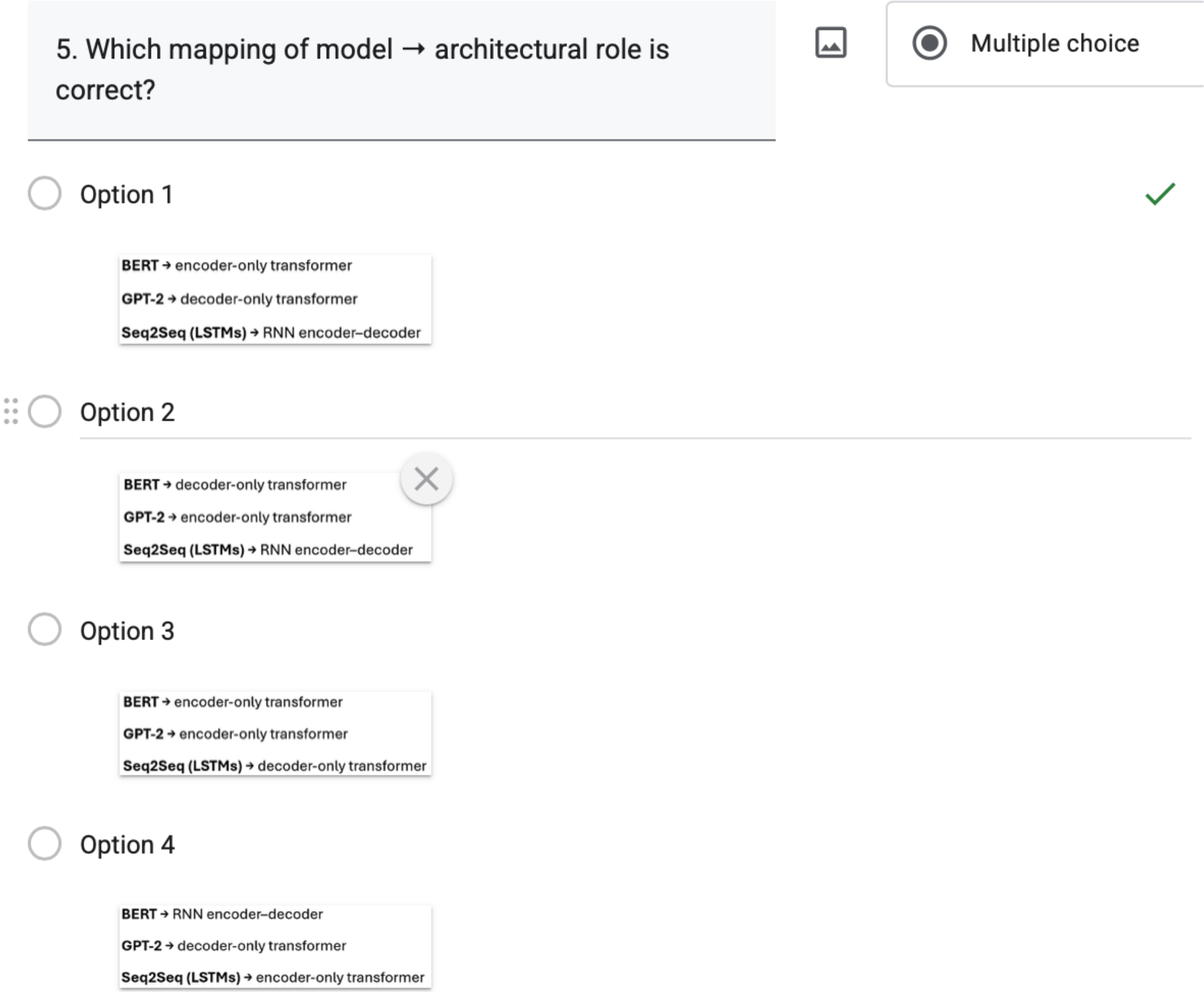
○ Dropout ✓

○ Increasing learning rate

○ Removing last layer

○ Using larger batch size

○ Add option or add "Other"



5. Which mapping of model → architectural role is correct?

● Multiple choice

○ Option 1 ✓

BERT → encoder-only transformer

GPT-2 → decoder-only transformer

Seq2Seq (LSTMs) → RNN encoder–decoder

○ Option 2

BERT → decoder-only transformer

GPT-2 → encoder-only transformer

Seq2Seq (LSTMs) → RNN encoder–decoder

○ Option 3

BERT → encoder-only transformer

GPT-2 → encoder-only transformer

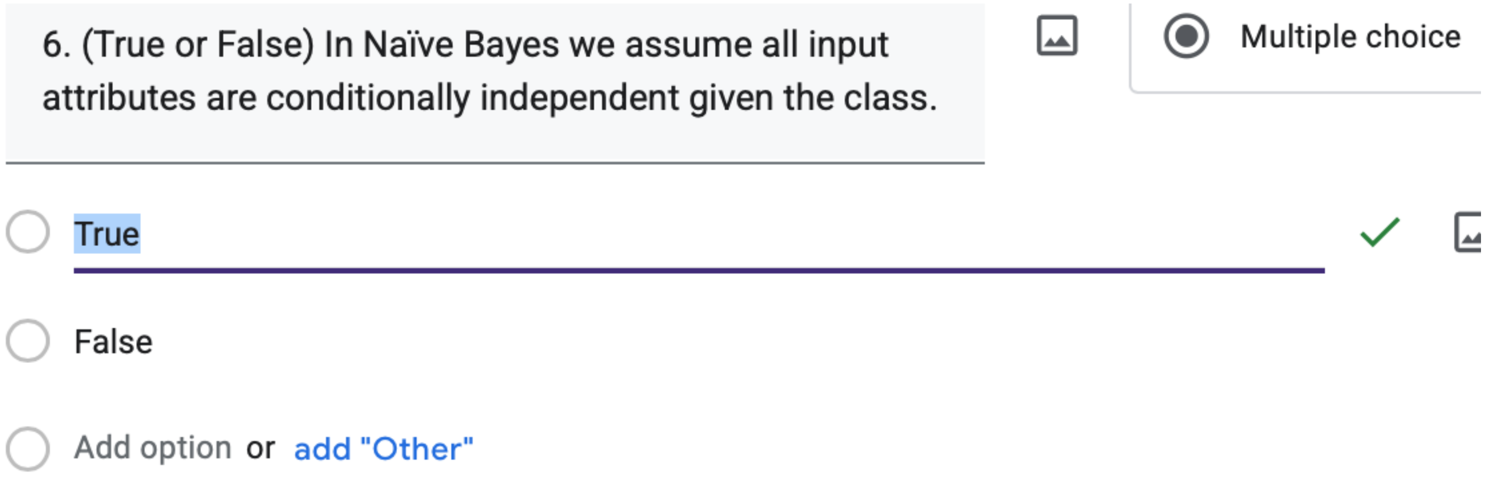
Seq2Seq (LSTMs) → decoder-only transformer

○ Option 4

BERT → RNN encoder–decoder

GPT-2 → decoder-only transformer

Seq2Seq (LSTMs) → encoder-only transformer



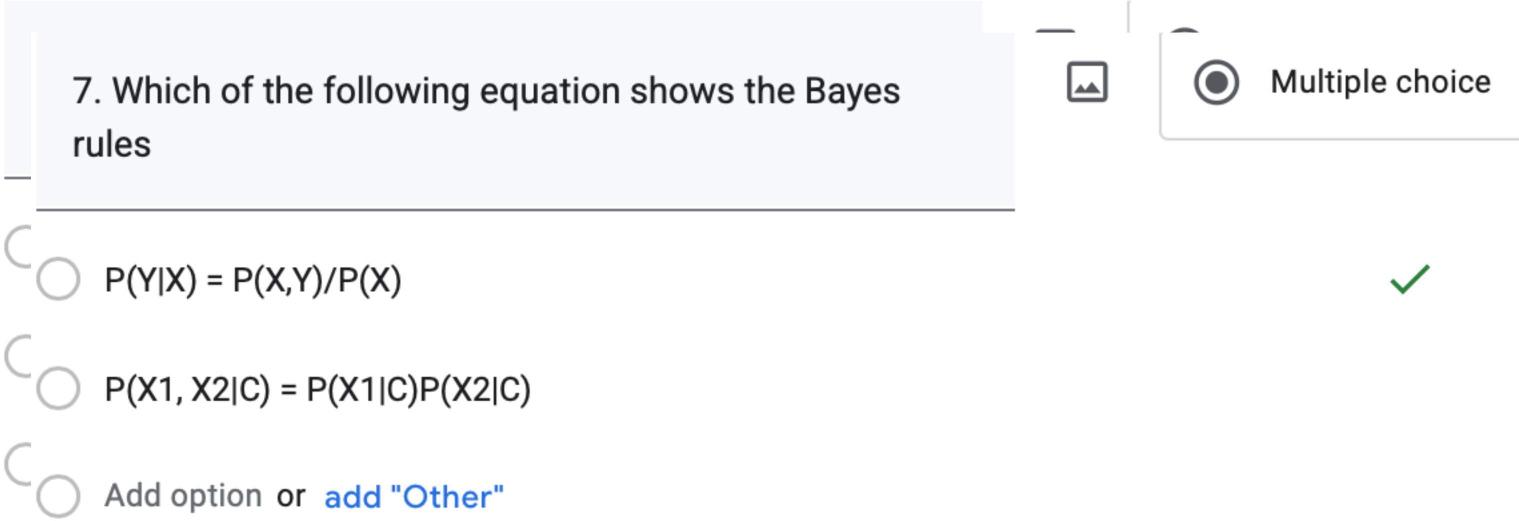
6. (True or False) In Naïve Bayes we assume all input attributes are conditionally independent given the class.

● Multiple choice

○ True ✓

○ False

○ Add option or add "Other"



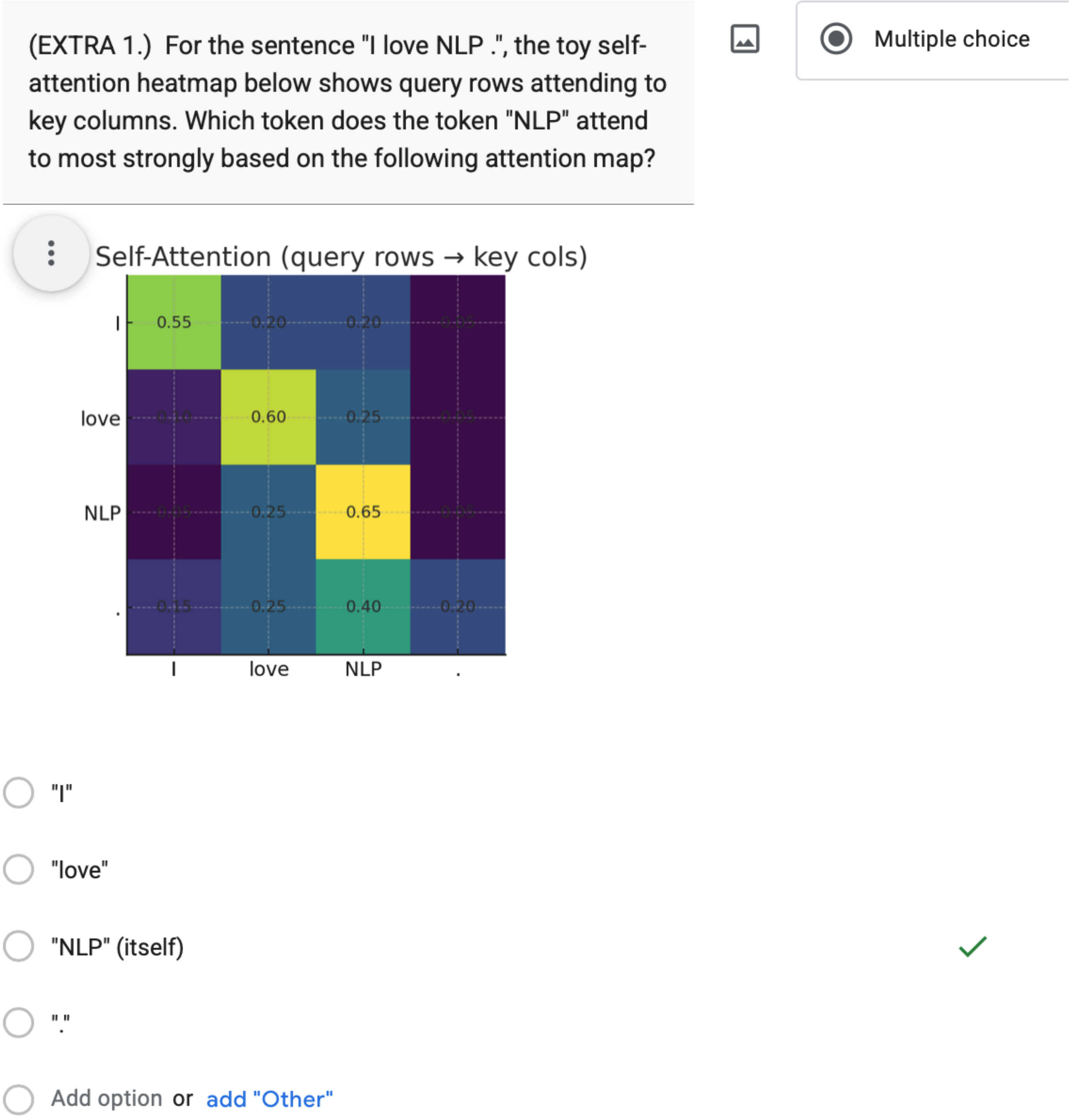
7. Which of the following equation shows the Bayes rules

● Multiple choice

○ P(Y|X) = P(X,Y)/P(X) ✓

○ P(X1, X2|C) = P(X1|C)P(X2|C)

○ Add option or add "Other"



(EXTRA 1.) For the sentence "I love NLP .", the toy self-attention heatmap below shows query rows attending to key columns. Which token does the token "NLP" attend to most strongly based on the following attention map?

Self-Attention (query rows → key cols)

	I	love	NLP	.
I	0.55	0.20	0.20	0.05
love	0.10	0.60	0.25	0.05
NLP	0.05	0.25	0.65	0.05
.	0.15	0.25	0.40	0.20

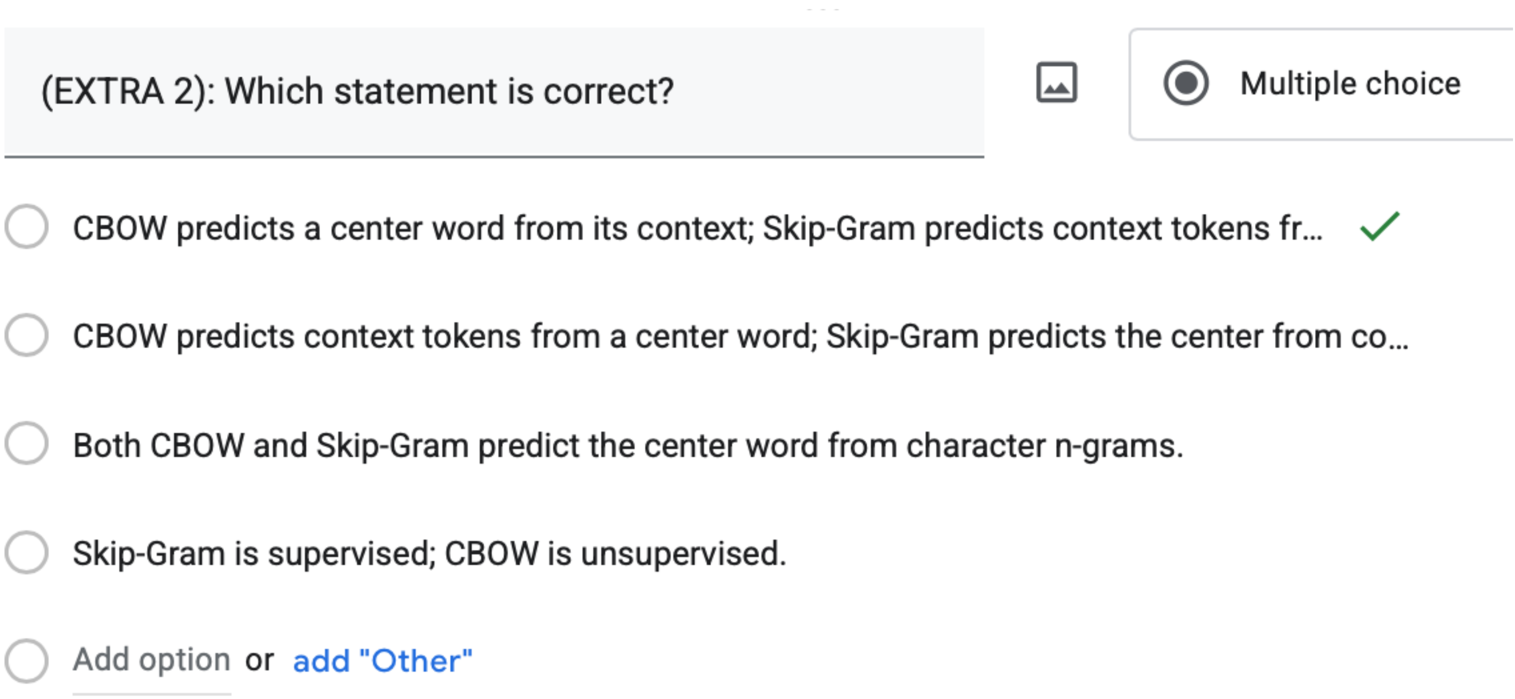
○ "I"

○ "love"

○ "NLP" (itself) ✓

○ "."

○ Add option or add "Other"



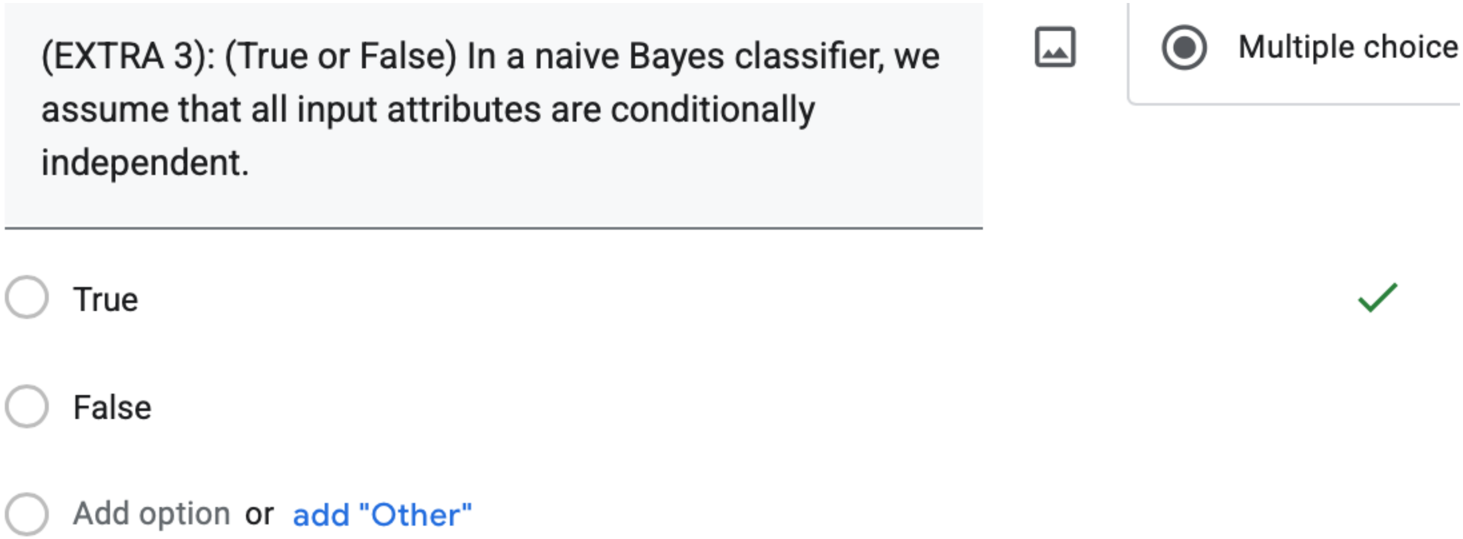
(EXTRA 2): Which statement is correct?

○ CBOW predicts a center word from its context; Skip-Gram predicts context tokens fr… ✓

○ CBOW predicts context tokens from a center word; Skip-Gram predicts the center from co…

○ Both CBOW and Skip-Gram predict the center word from character n-grams.

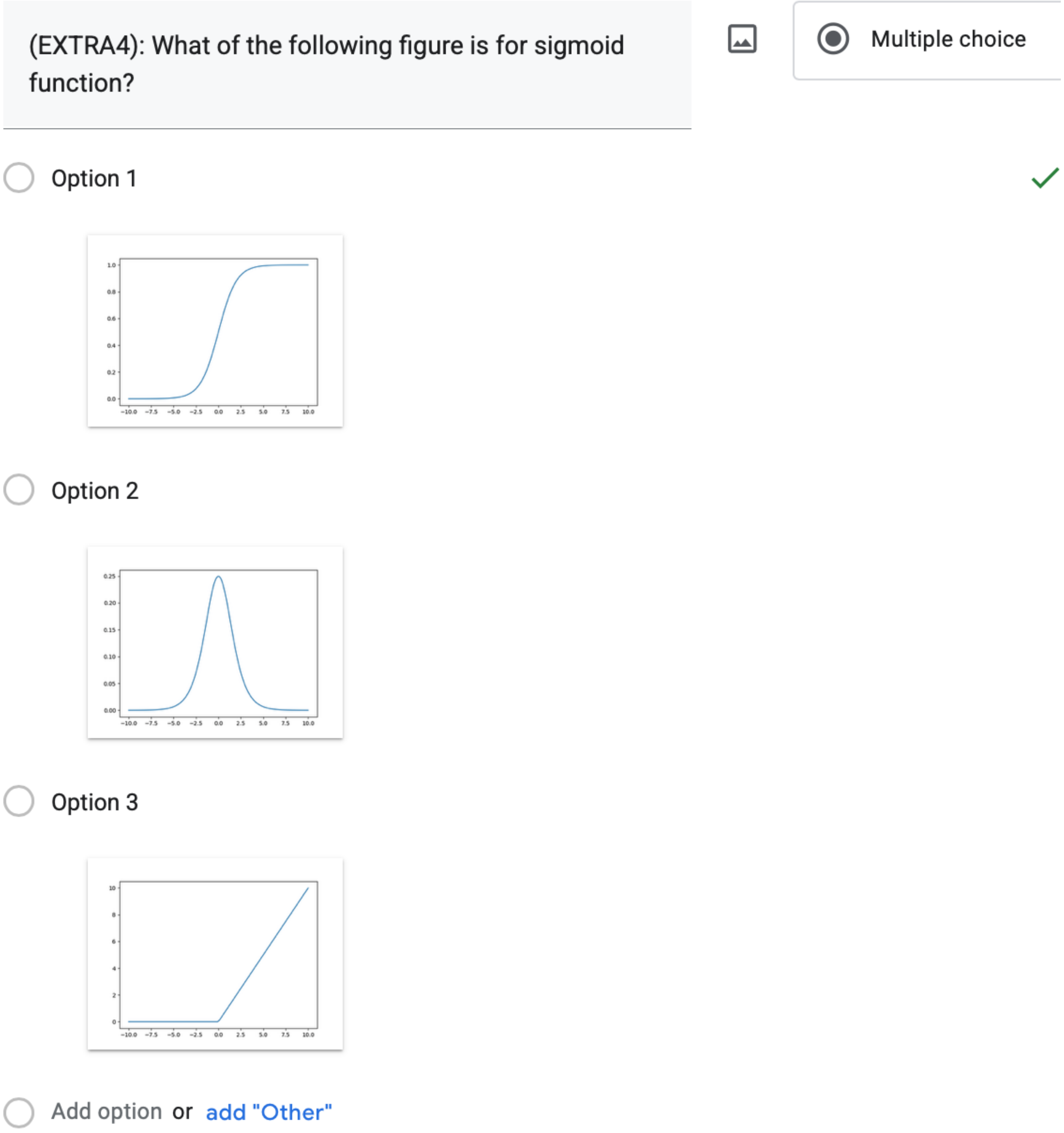
○ Skip-Gram is supervised; CBOW is unsupervised.



(EXTRA 3): (True or False) In a naive Bayes classifier, we assume that all input attributes are conditionally independent.

○ True ✓

○ False



(EXTRA4): What of the following figure is for sigmoid function?

● Multiple choice

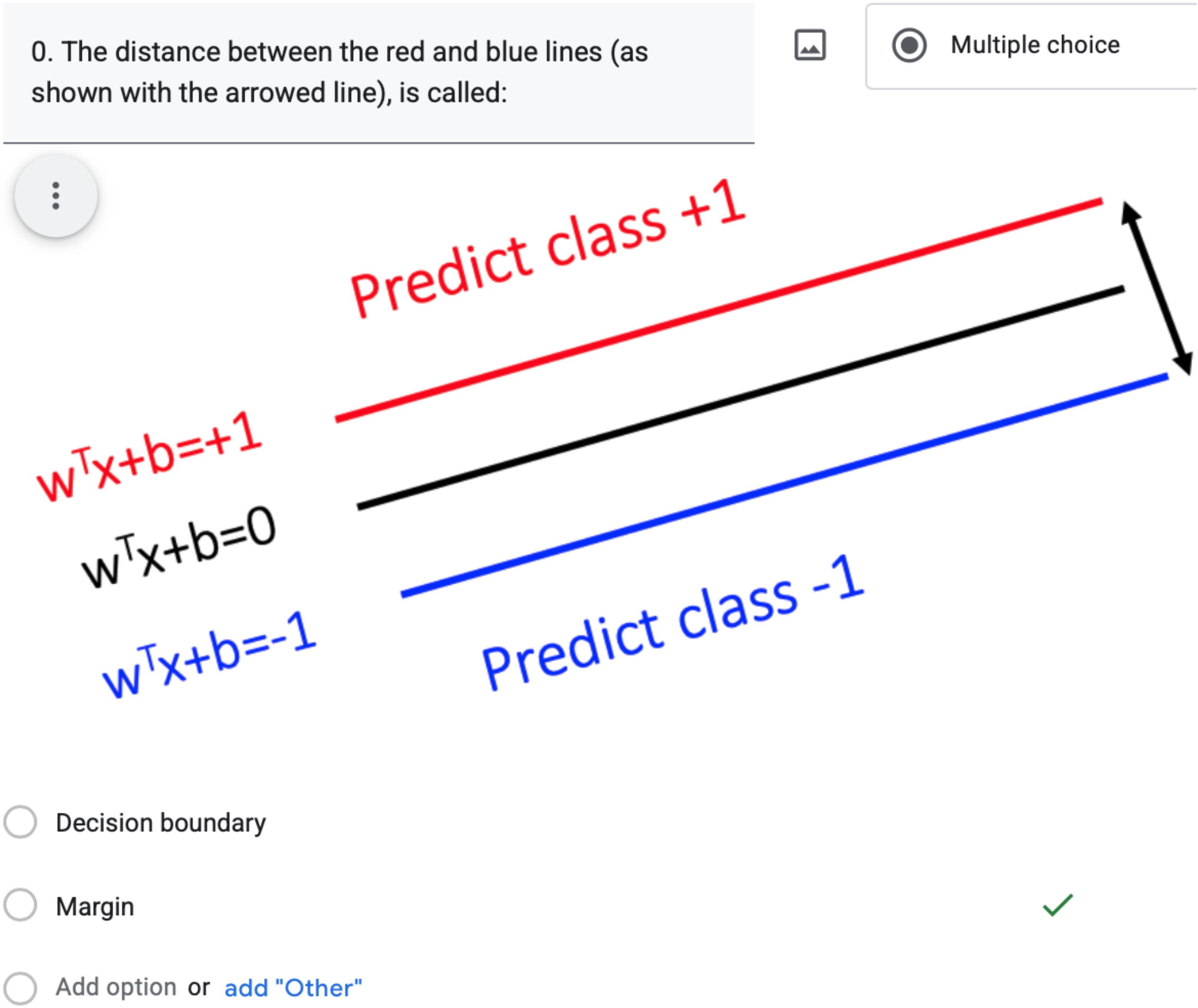
○ Option 1 ✓

○ Option 2

○ Option 3

○ Add option or add "Other"

Q10



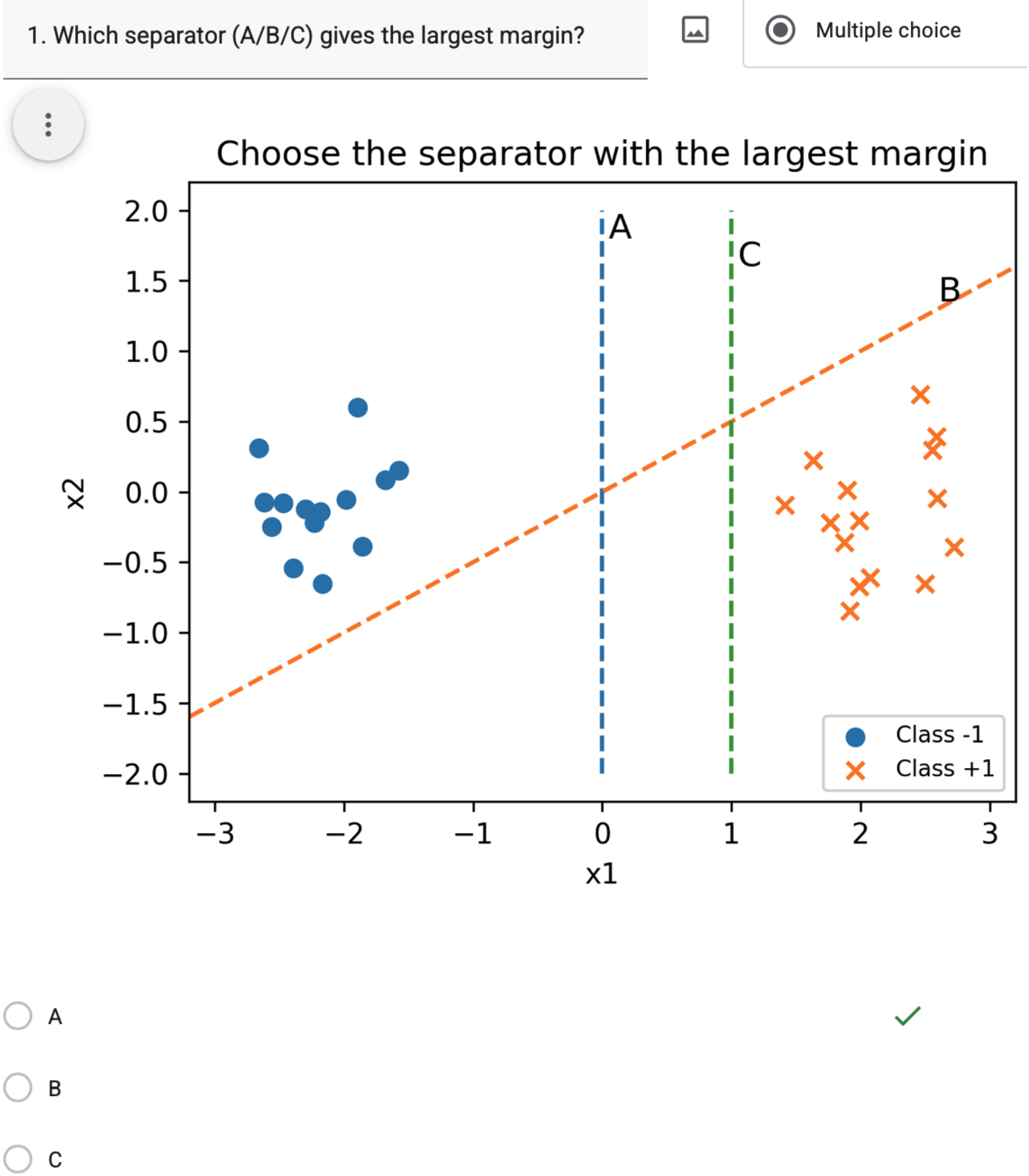
0. The distance between the red and blue lines (as shown with the arrowed line), is called:

● Multiple choice

○ Decision boundary

○ Margin ✓

○ Add option or add "Other"



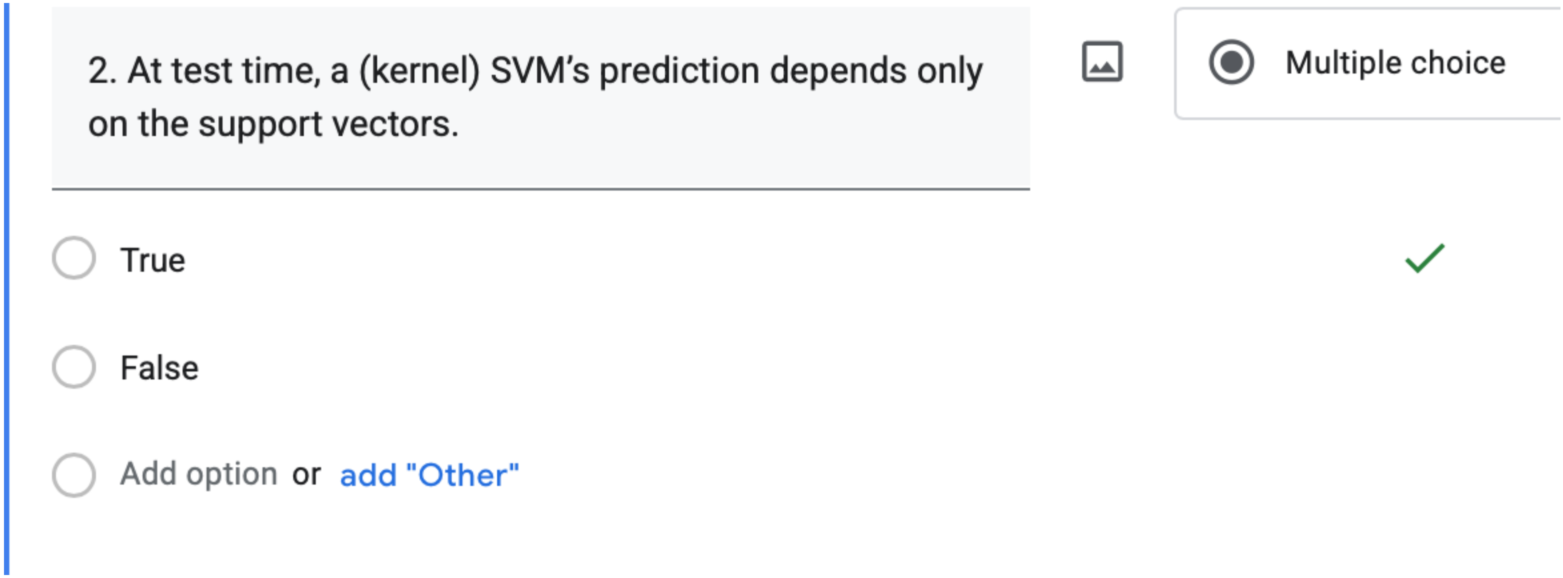
1. Which separator (A/B/C) gives the largest margin?

● Multiple choice

○ A ✓

○ B

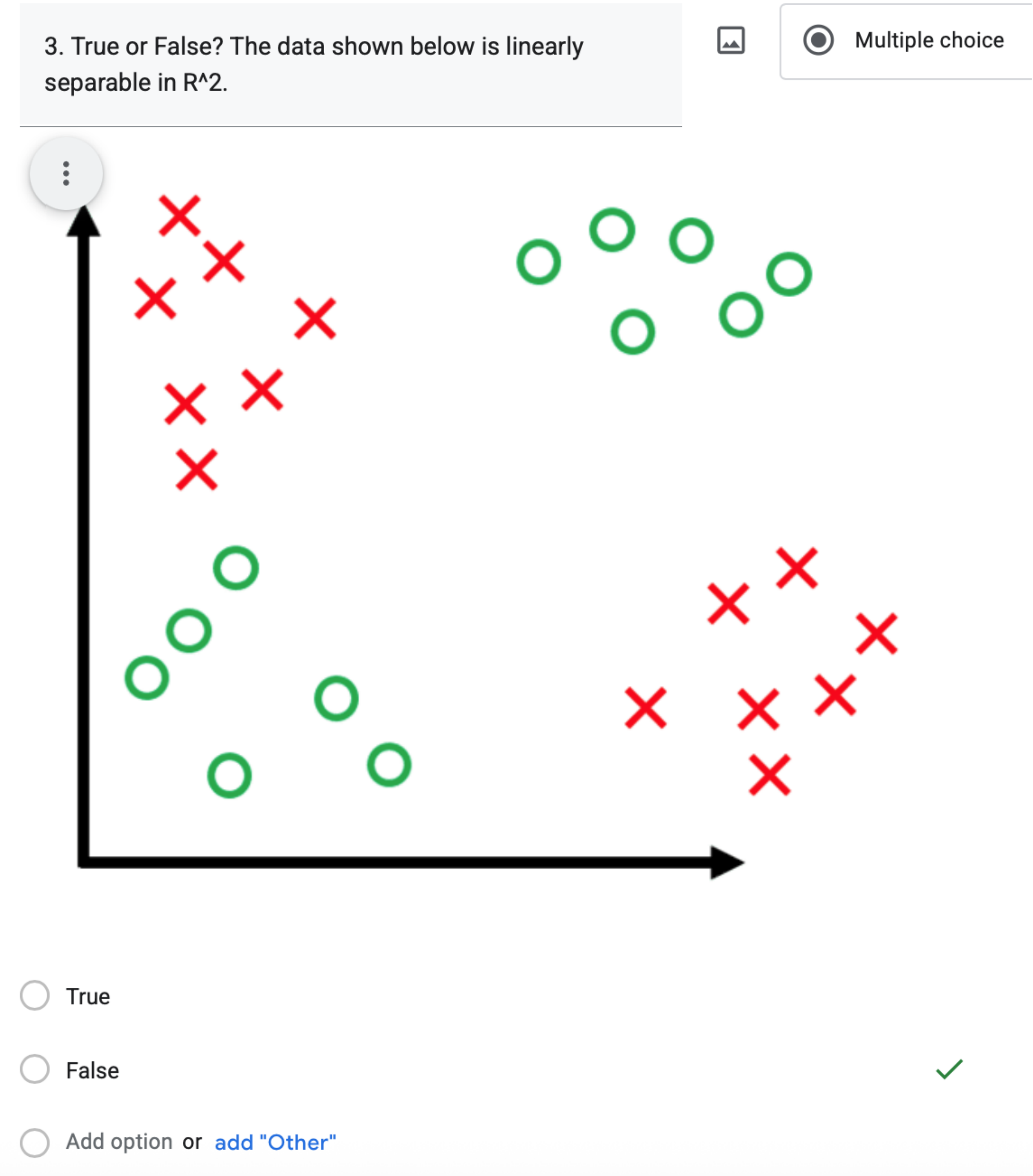
○ C



2. At test time, a (kernel) SVM’s prediction depends only on the support vectors.

○ True ✓

○ False

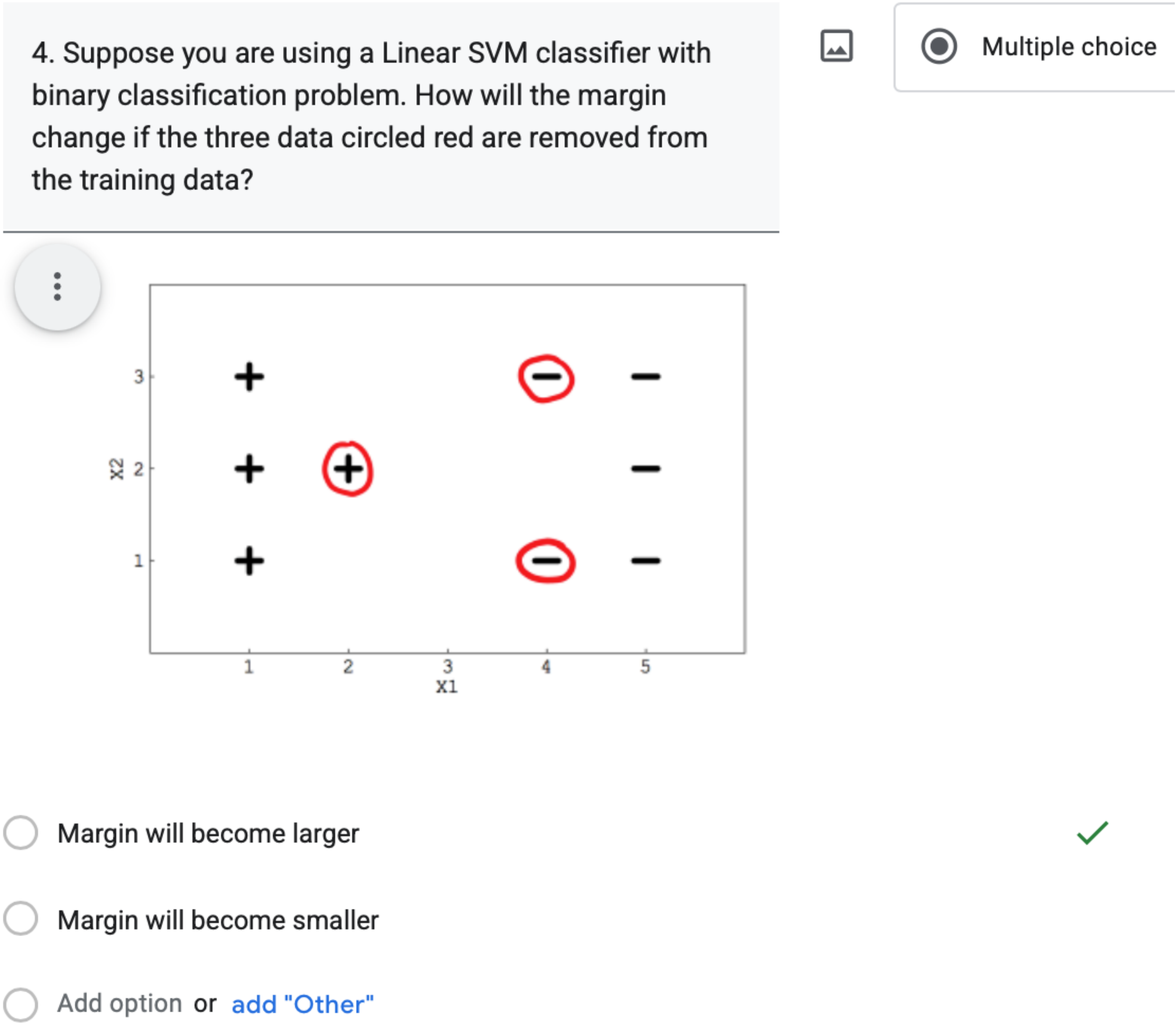


3. True or False? The data shown below is linearly separable in R^2.

○ True

○ False ✓

○ Add option or add "Other"



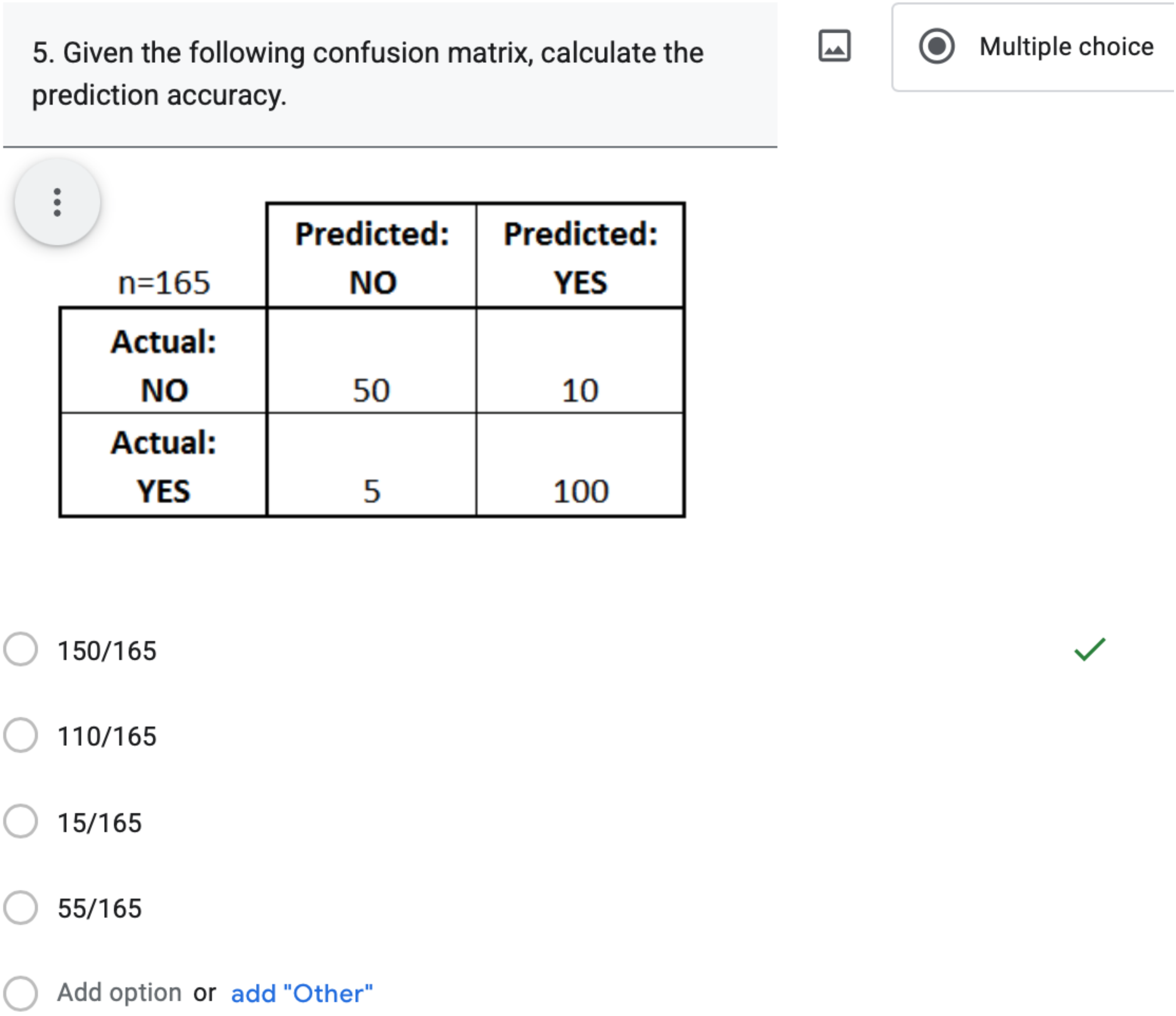
4. Suppose you are using a Linear SVM classifier with binary classification problem. How will the margin change if the three data circled red are removed from the training data?

● Multiple choice

○ Margin will become larger ✓

○ Margin will become smaller

○ Add option or add "Other"



5. Given the following confusion matrix, calculate the prediction accuracy.

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

○ 150/165 ✓

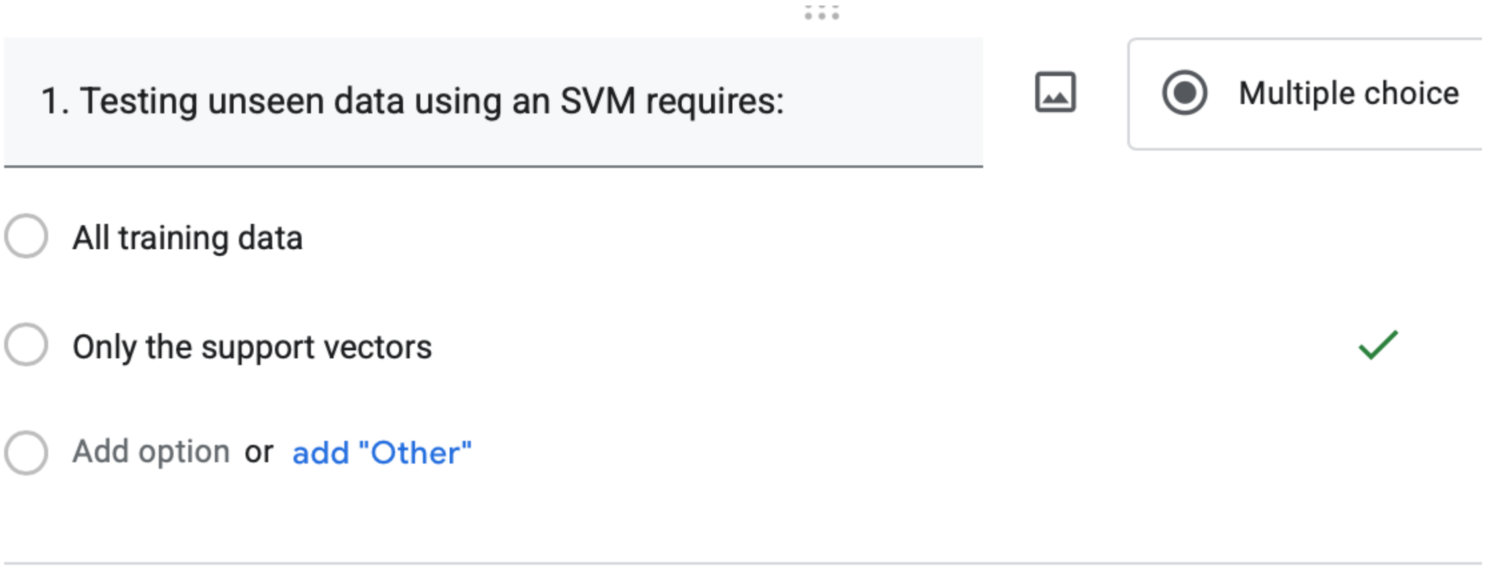
○ 110/165

○ 15/165

○ 55/165

○ Add option or add "Other"

Q11

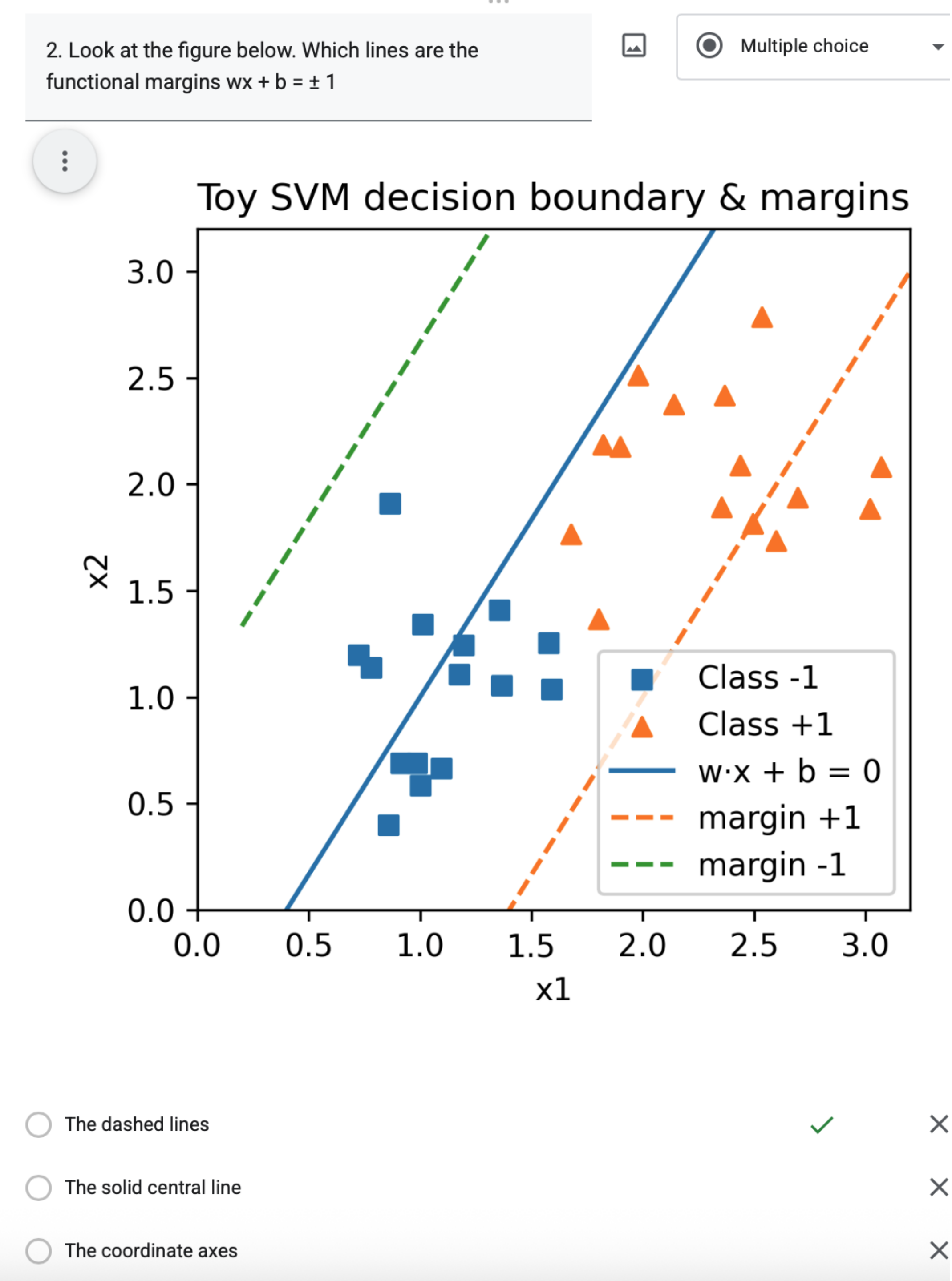


1. Testing unseen data using an SVM requires:

● Multiple choice

○ All training data

○ Only the support vectors ✓



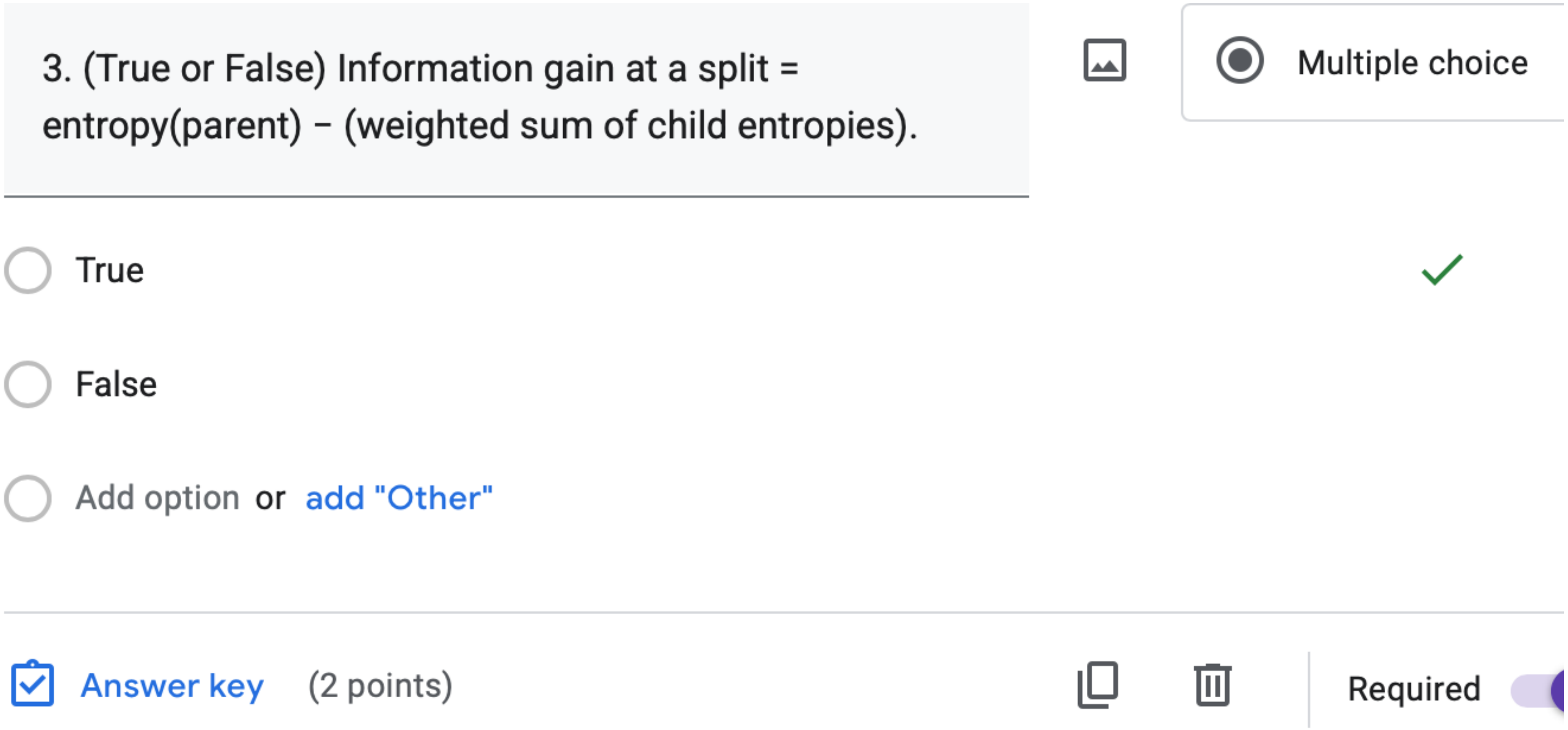
2. Look at the figure below. Which lines are the functional margins wx + b = ± 1

● Multiple choice

○ The dashed lines ✓ ✕

○ The solid central line ✕

○ The coordinate axes ✕



3. (True or False) Information gain at a split = entropy(parent) – (weighted sum of child entropies).

● Multiple choice

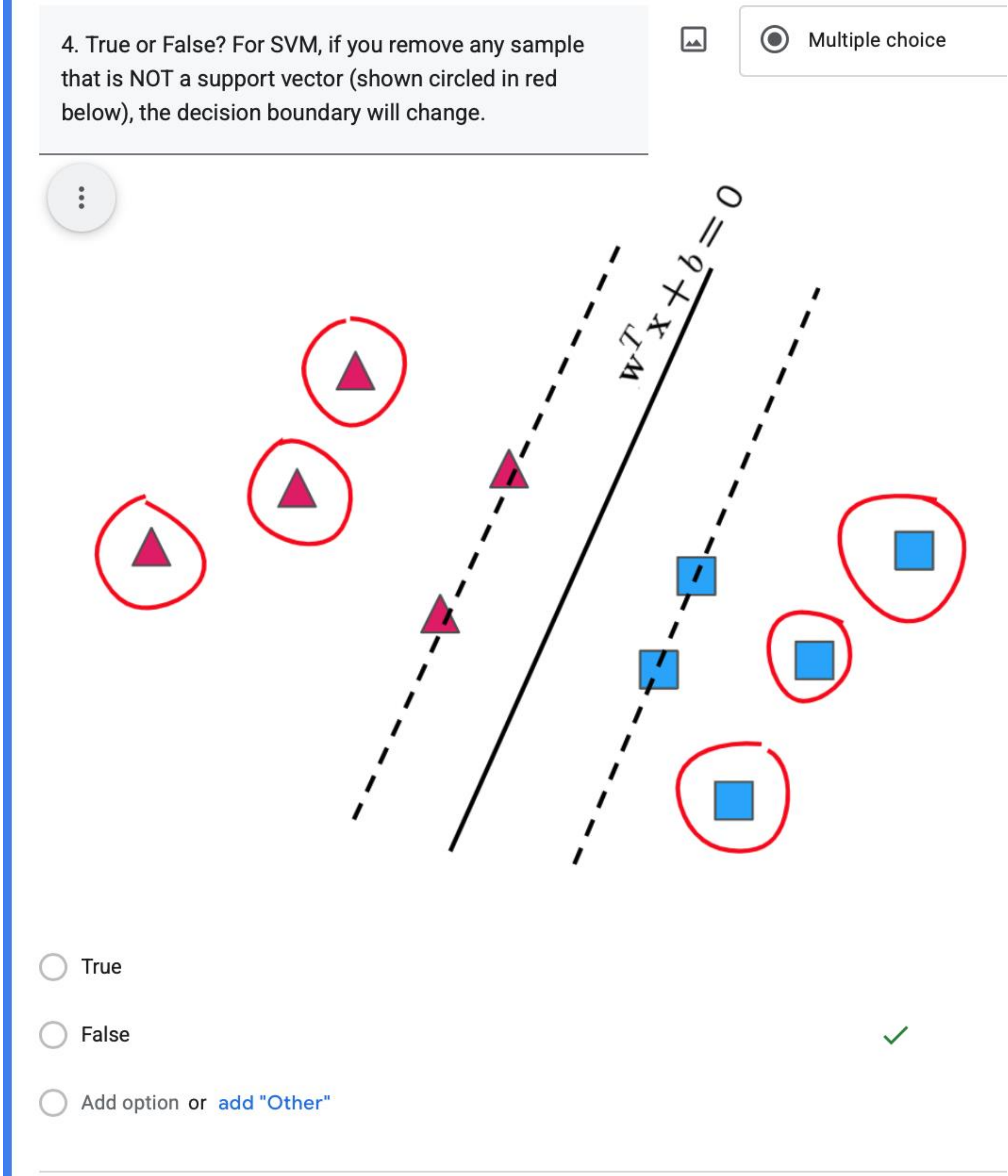
○ True ✓

○ False

○ Add option or add "Other"

☑ Answer key (2 points)

Required



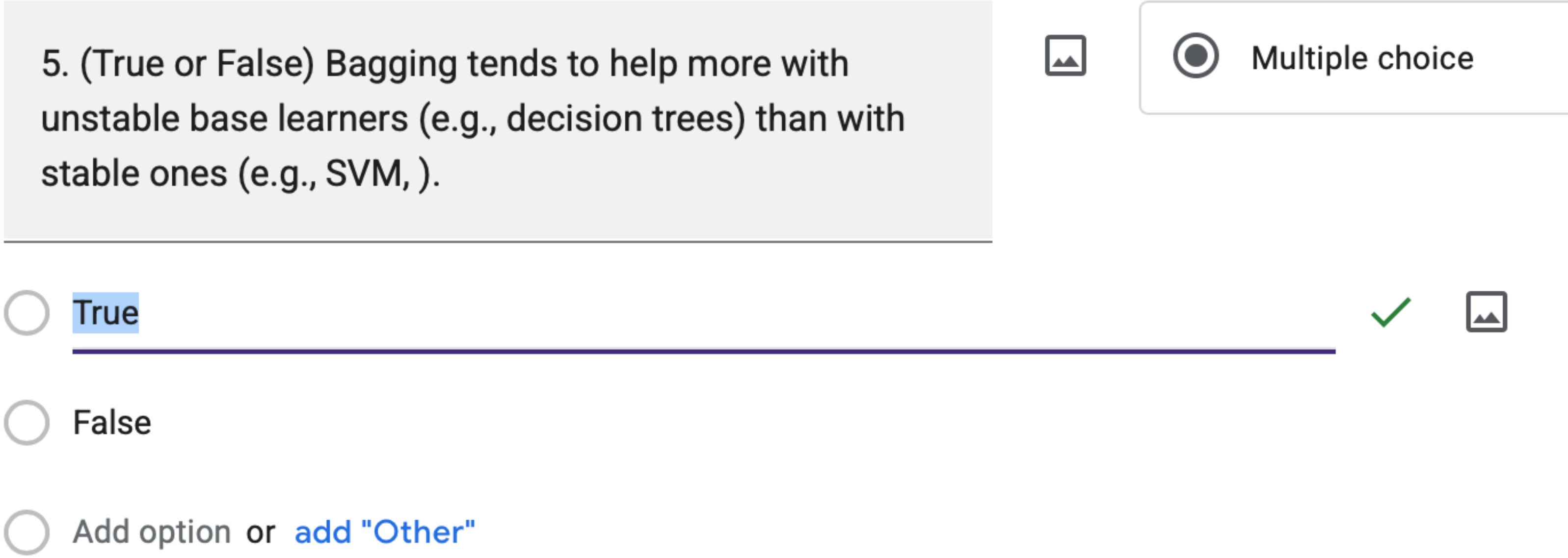
4. True or False? For SVM, if you remove any sample that is NOT a support vector (shown circled in red below), the decision boundary will change.

● Multiple choice

○ True

○ False ✓

○ Add option or add "Other"

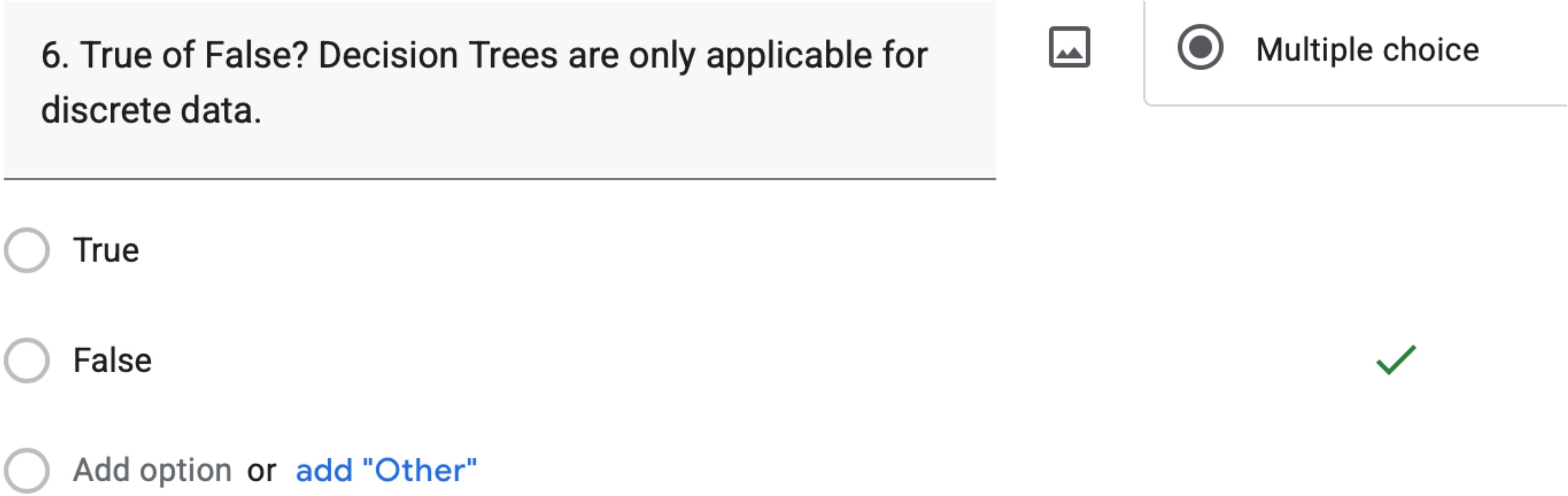


5. (True or False) Bagging tends to help more with unstable base learners (e.g., decision trees) than with stable ones (e.g., SVM,).

● Multiple choice

○ True ✓

○ False



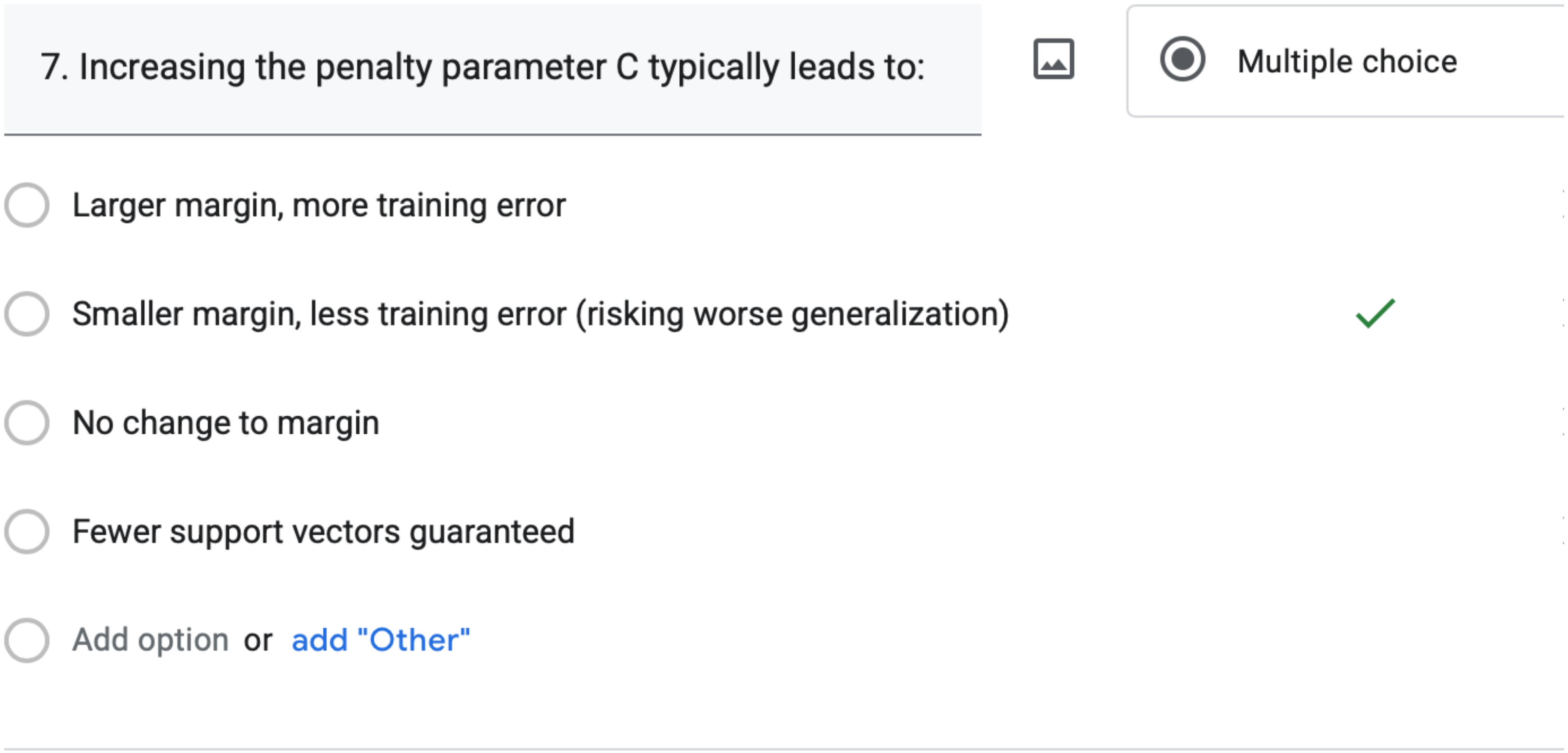
6. True of False? Decision Trees are only applicable for discrete data.

● Multiple choice

○ True

○ False ✓

○ Add option or add "Other"



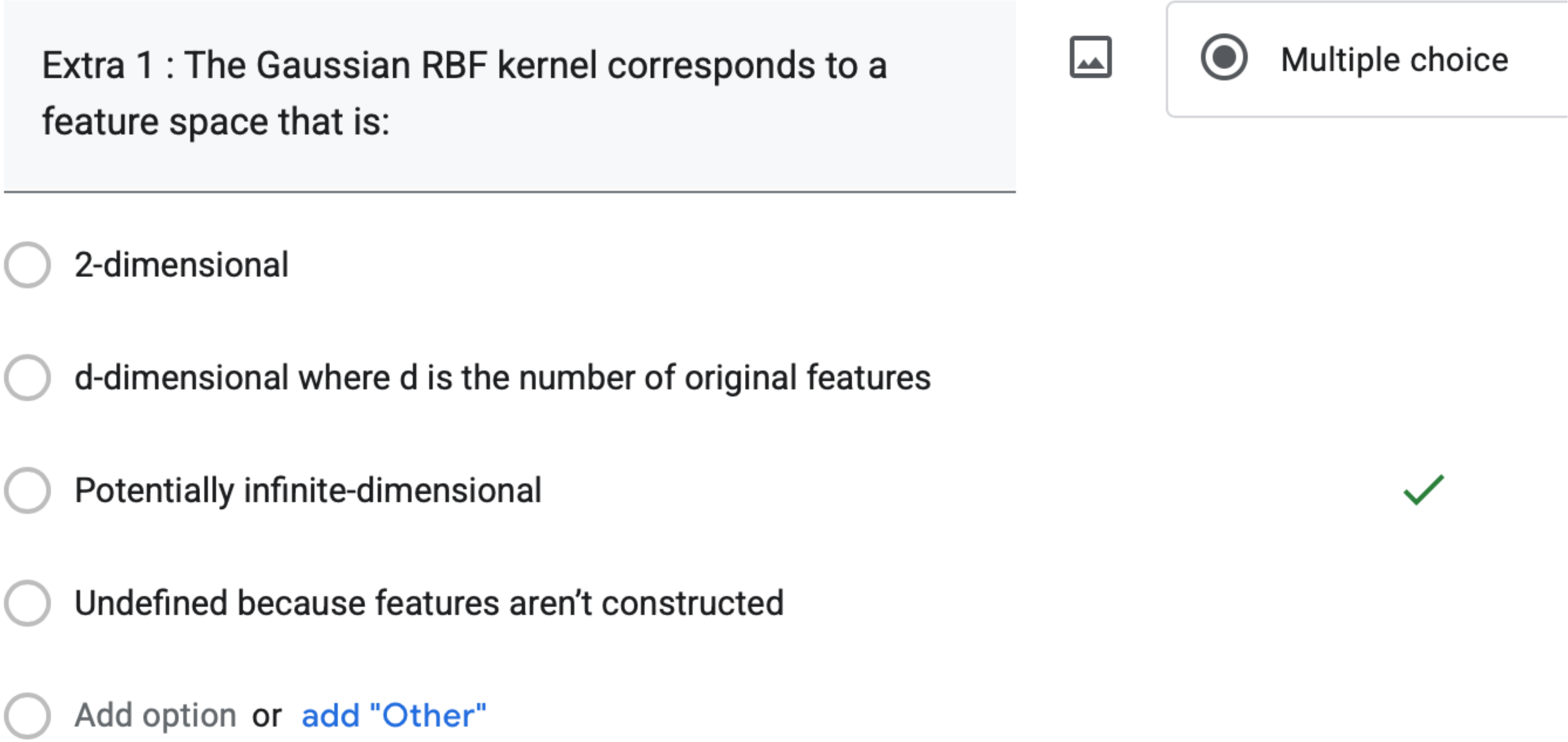
7. Increasing the penalty parameter C typically leads to:

○ Larger margin, more training error

○ Smaller margin, less training error (risking worse generalization) ✓

○ No change to margin

○ Fewer support vectors guaranteed



Extra 1 : The Gaussian RBF kernel corresponds to a feature space that is:

● Multiple choice

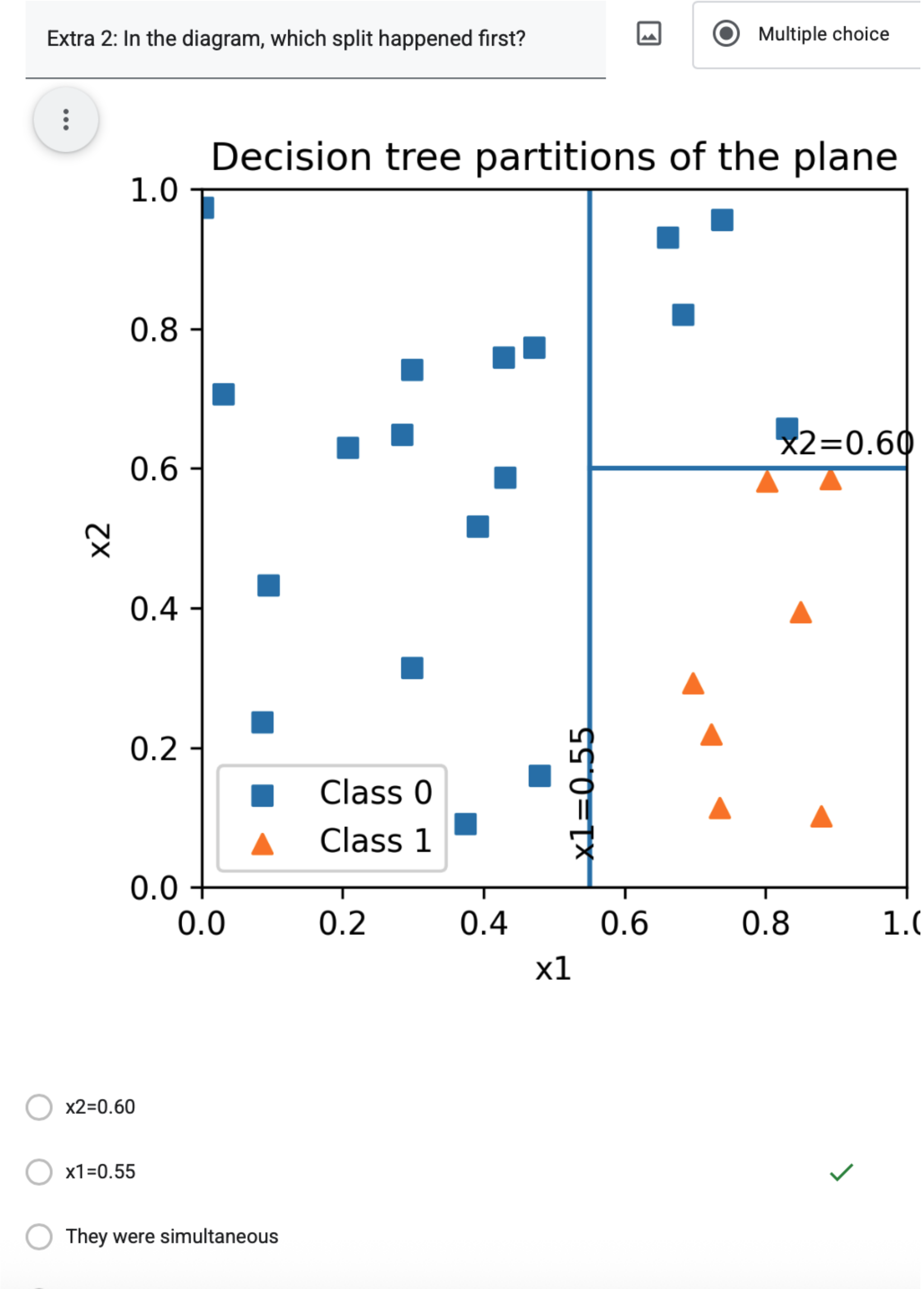
○ 2-dimensional

○ d-dimensional where d is the number of original features

○ Potentially infinite-dimensional ✓

○ Undefined because features aren’t constructed

○ Add option or add "Other"

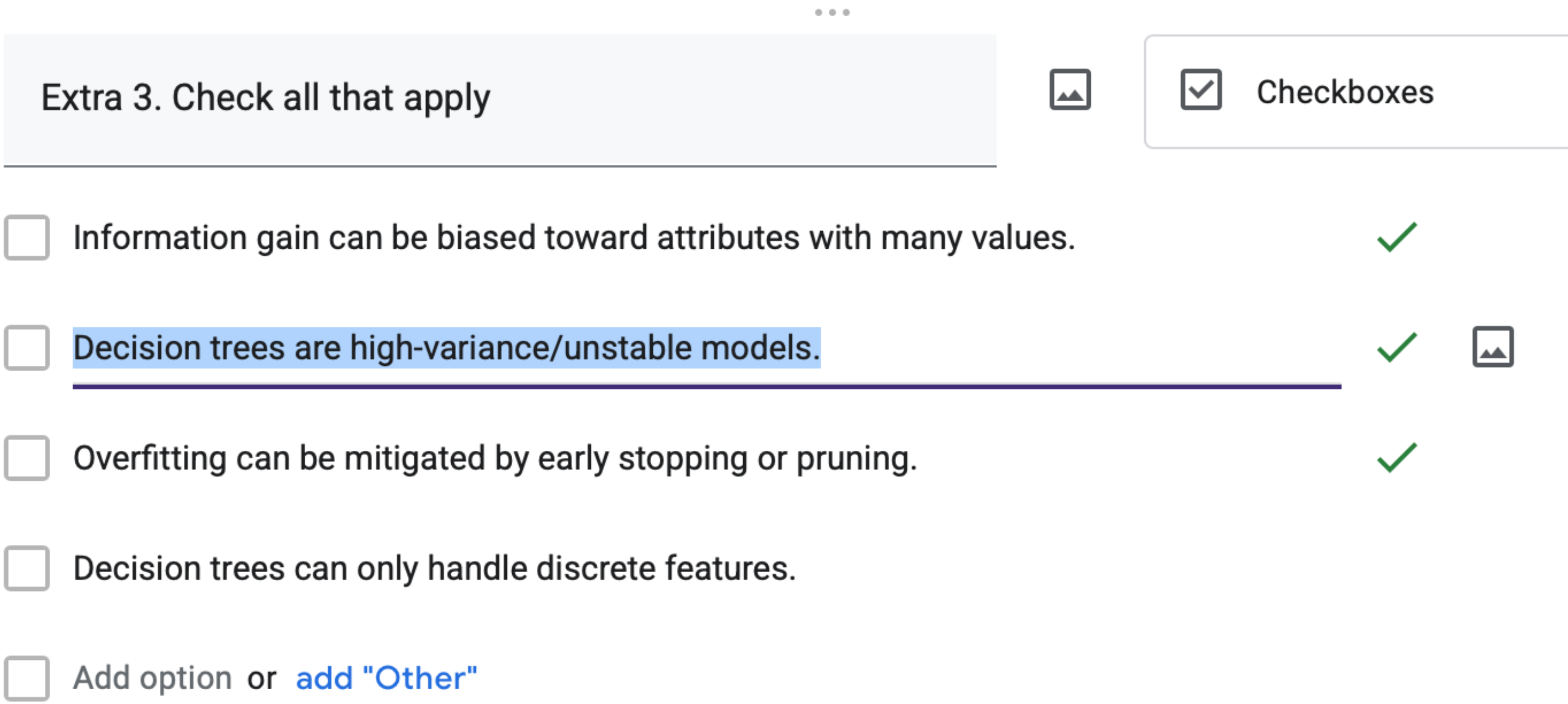


Extra 2: In the diagram, which split happened first?

○ x2=0.60

○ x1=0.55 ✓

○ They were simultaneous



Extra 3. Check all that apply

☑ Checkboxes

□ Information gain can be biased toward attributes with many values. ✓

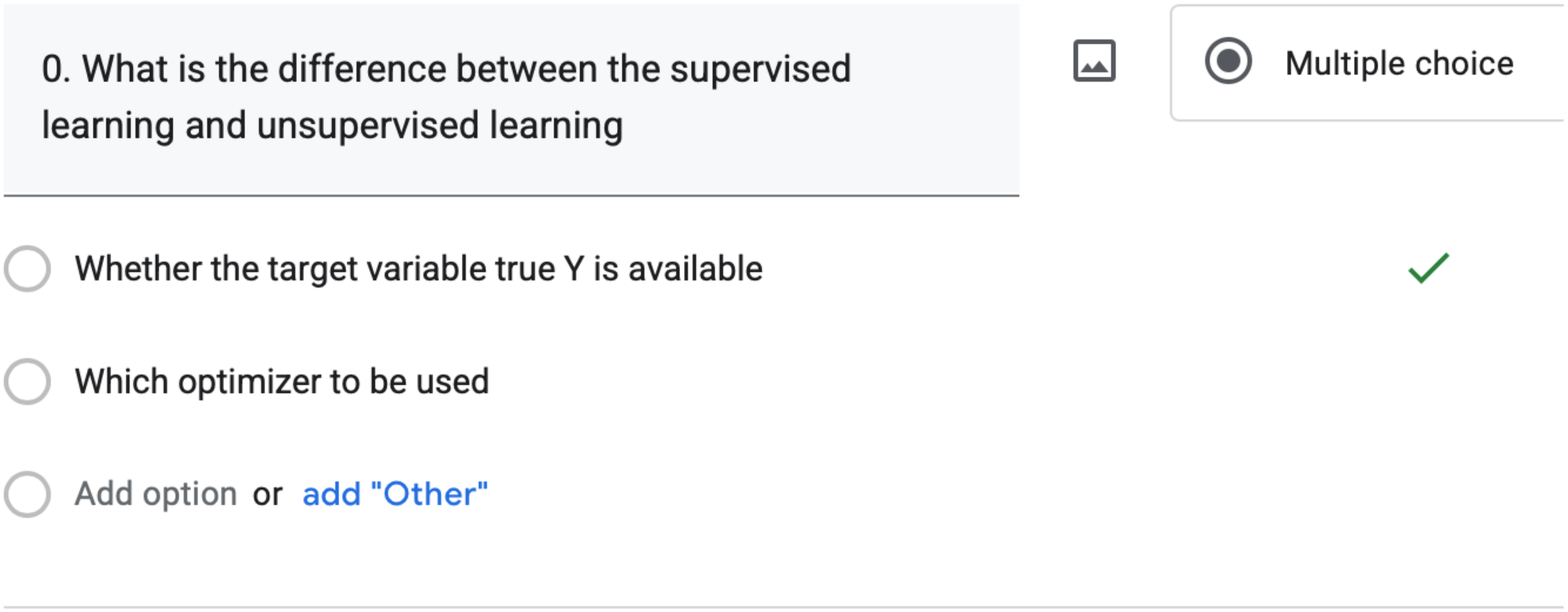
□ Decision trees are high-variance/unstable models. ✓

□ Overfitting can be mitigated by early stopping or pruning. ✓

□ Decision trees can only handle discrete features.

□ Add option or add "Other"

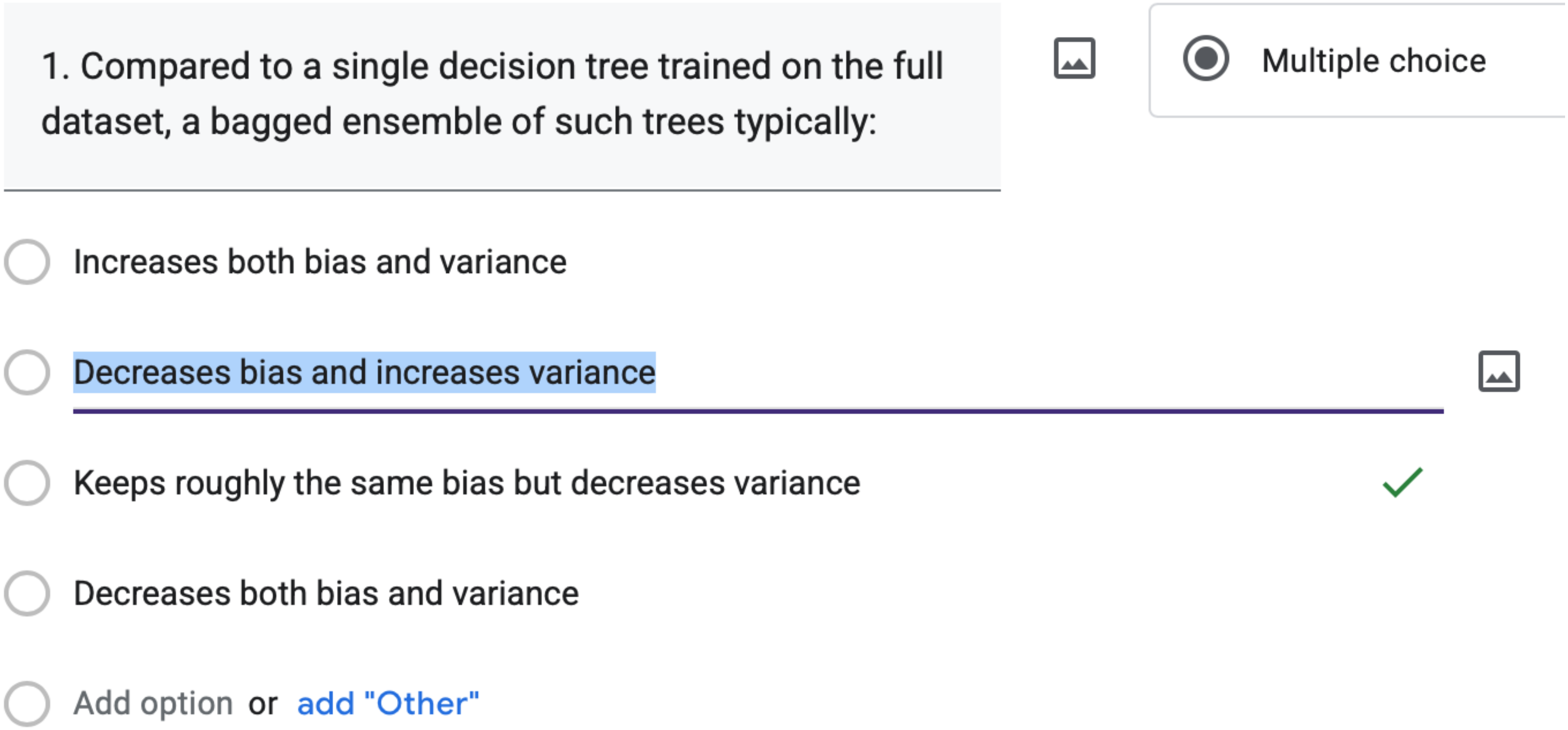
Q12



0. What is the difference between the supervised learning and unsupervised learning

○ Whether the target variable true Y is available ✓

○ Which optimizer to be used



1. Compared to a single decision tree trained on the full dataset, a bagged ensemble of such trees typically:

◉ Multiple choice

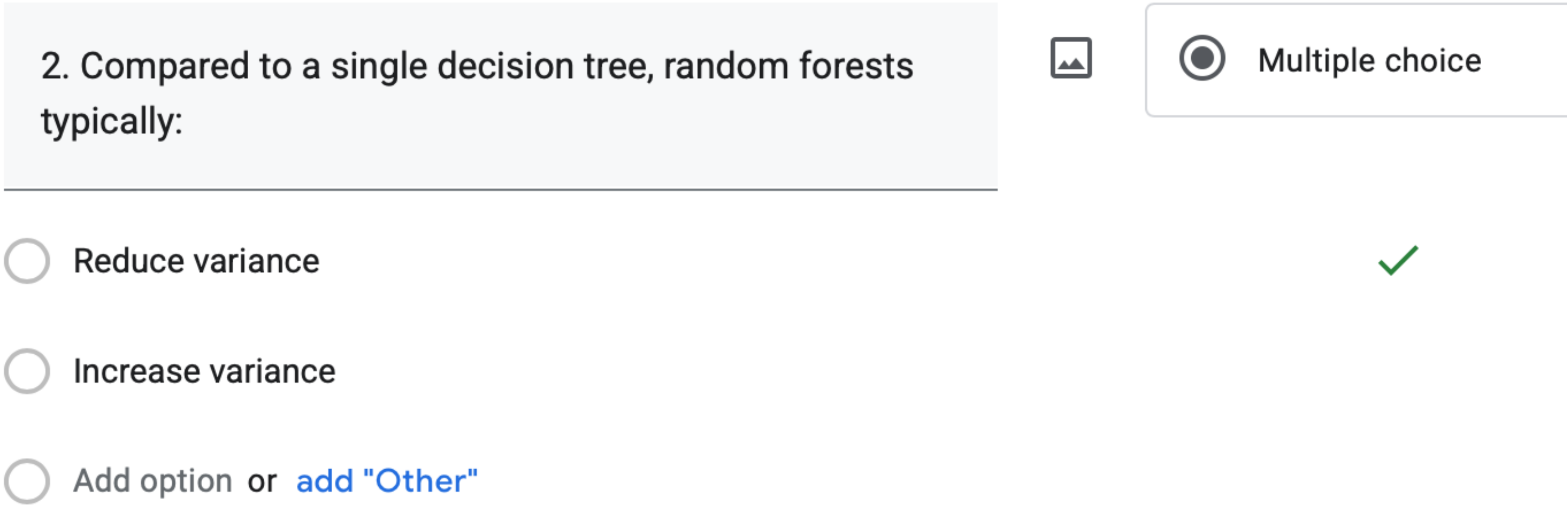
○ Increases both bias and variance

○ Decreases bias and increases variance

○ Keeps roughly the same bias but decreases variance ✓

○ Decreases both bias and variance

○ Add option or add "Other"



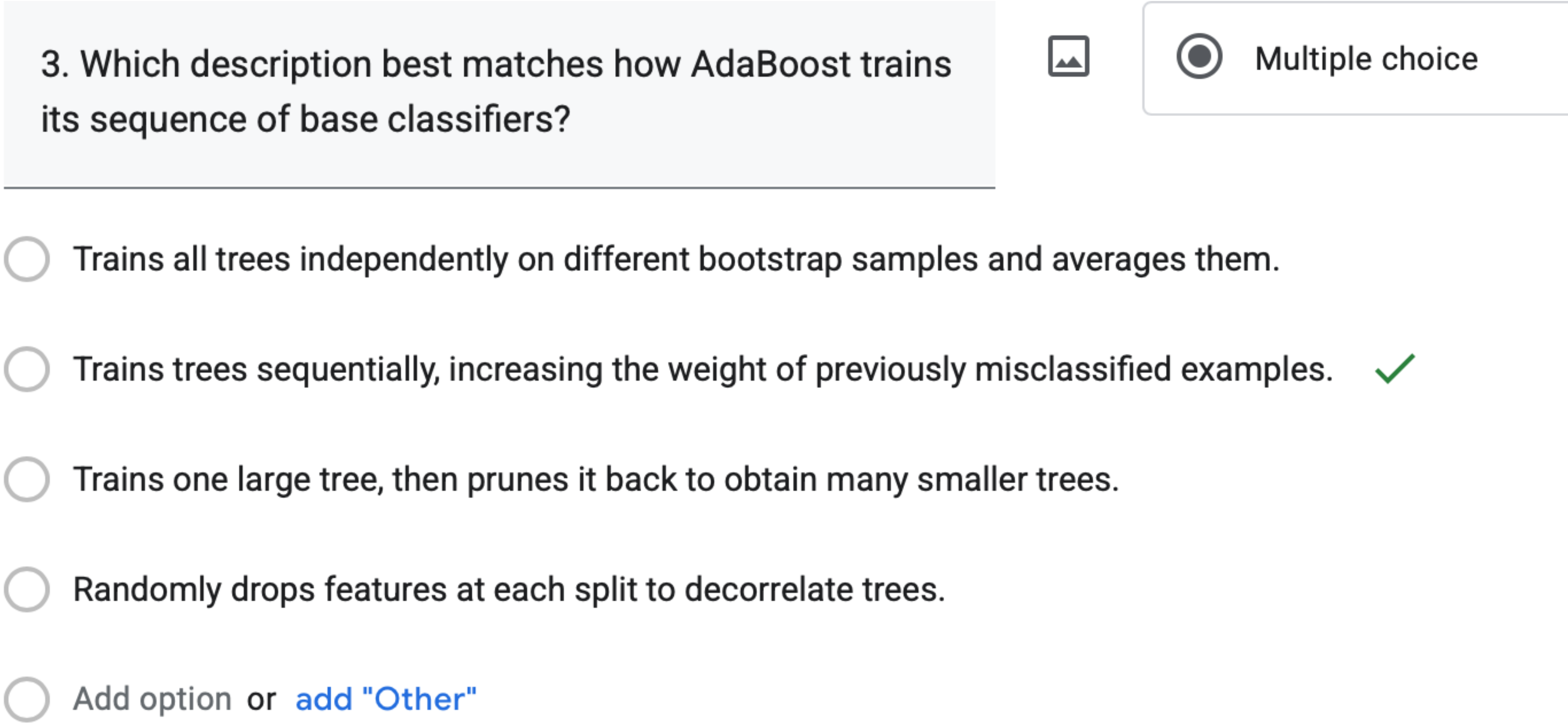
2. Compared to a single decision tree, random forests typically:

● Multiple choice

○ Reduce variance ✓

○ Increase variance

○ Add option or add "Other"



3. Which description best matches how AdaBoost trains its sequence of base classifiers?

● Multiple choice

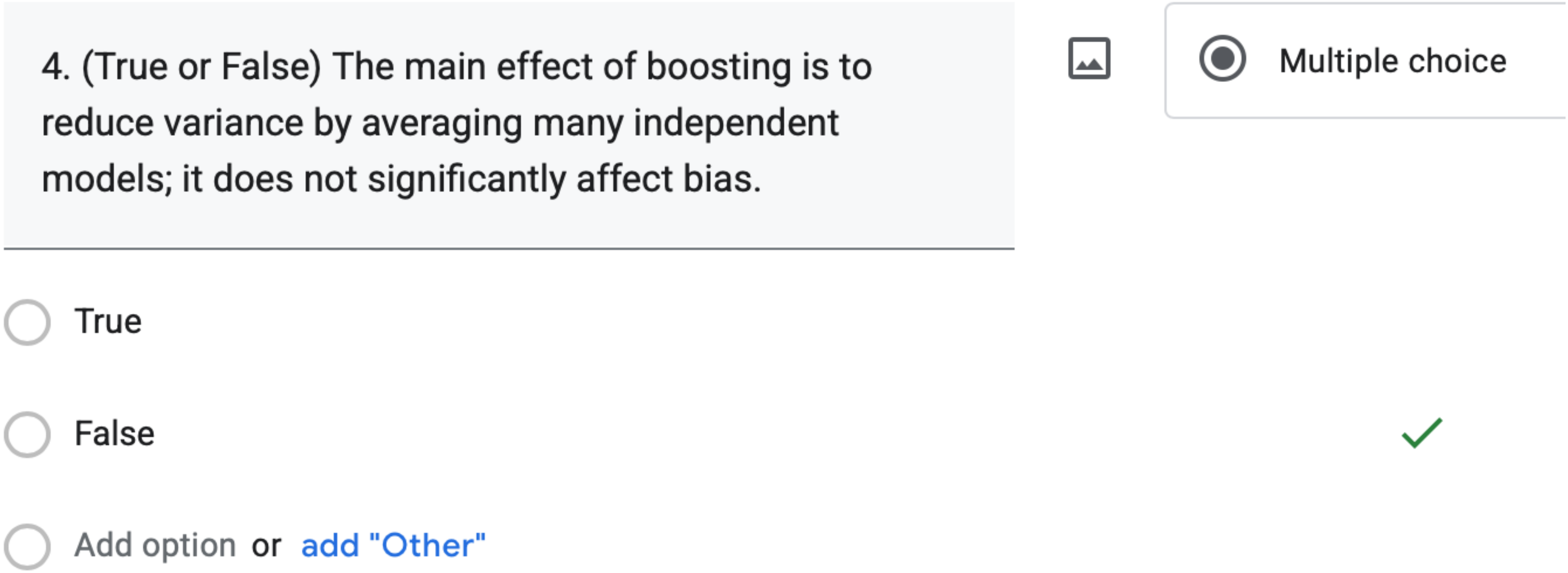
○ Trains all trees independently on different bootstrap samples and averages them.

○ Trains trees sequentially, increasing the weight of previously misclassified examples. ✓

○ Trains one large tree, then prunes it back to obtain many smaller trees.

○ Randomly drops features at each split to decorrelate trees.

○ Add option or add "Other"



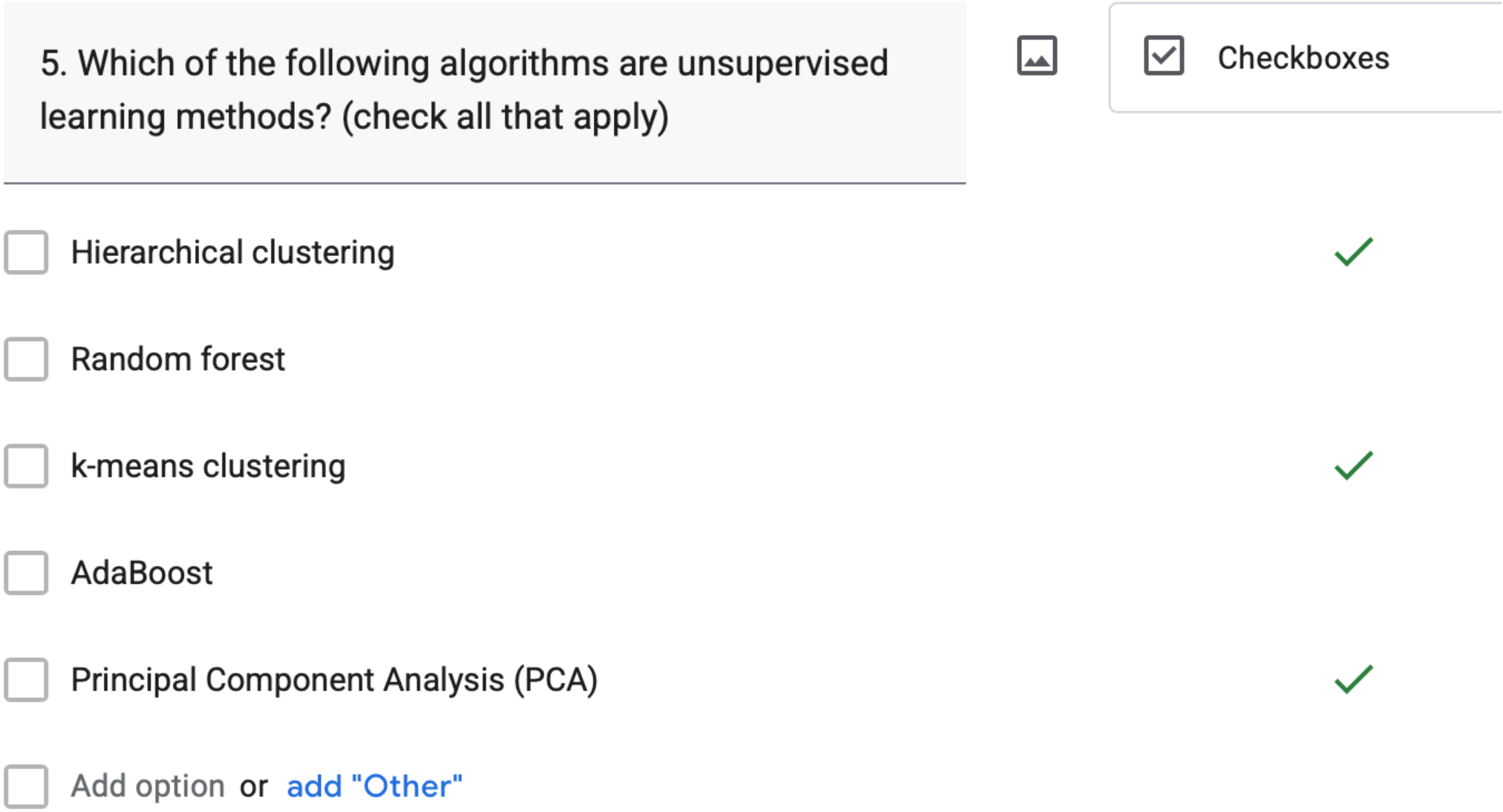
4. (True or False) The main effect of boosting is to reduce variance by averaging many independent models; it does not significantly affect bias.

● Multiple choice

○ True

○ False ✓

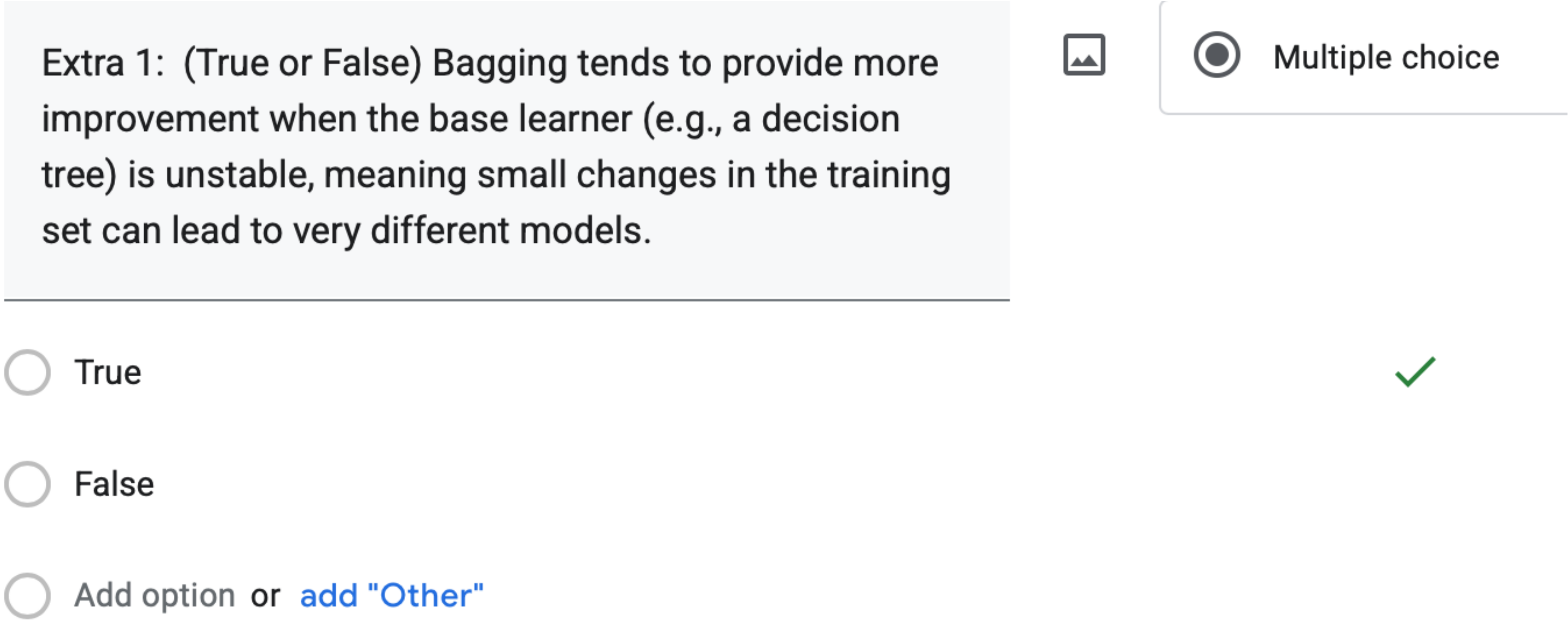
○ Add option or add "Other"



5. Which of the following algorithms are unsupervised learning methods? (check all that apply)

☑ Checkboxes

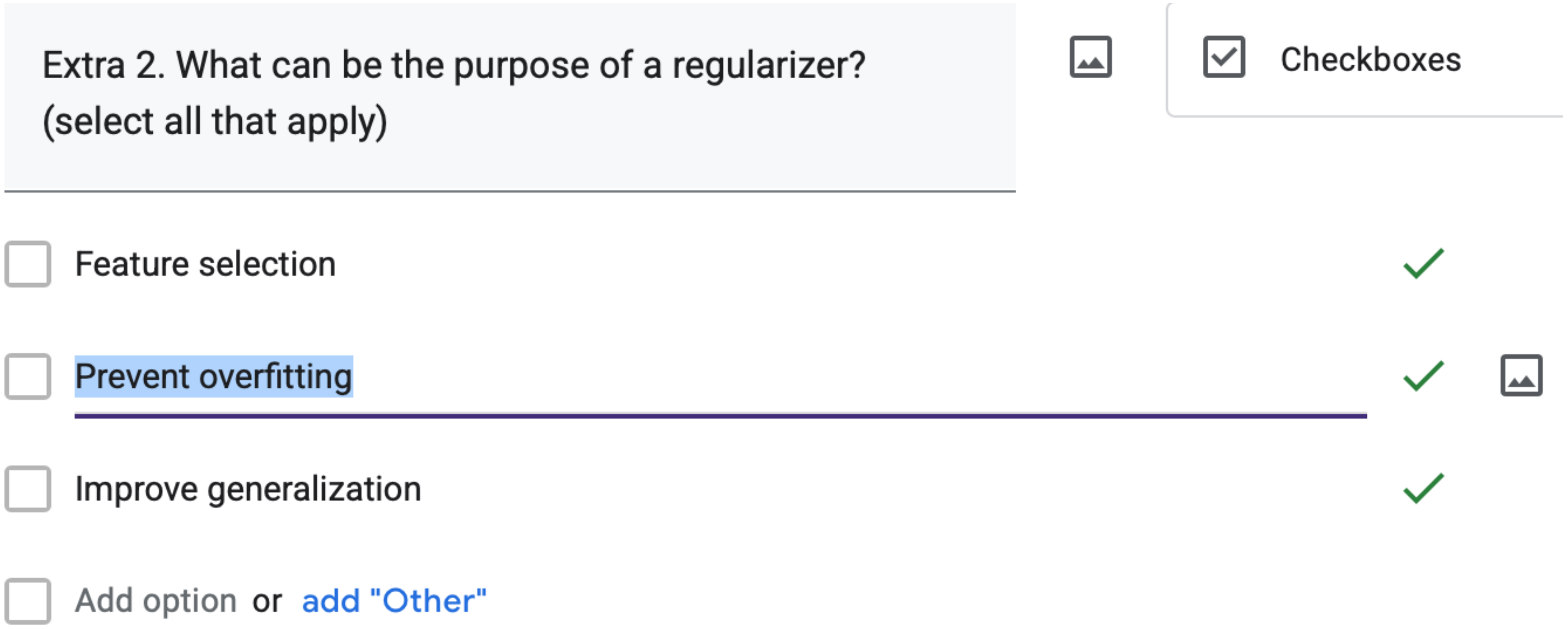
- □ Hierarchical clustering ✓
- □ Random forest
- □ k-means clustering ✓
- □ AdaBoost
- □ Principal Component Analysis (PCA) ✓
- □ Add option or add "Other"



Extra 1: (True or False) Bagging tends to provide more improvement when the base learner (e.g., a decision tree) is unstable, meaning small changes in the training set can lead to very different models.

○ True ✓

○ False



Extra 2. What can be the purpose of a regularizer? (select all that apply)

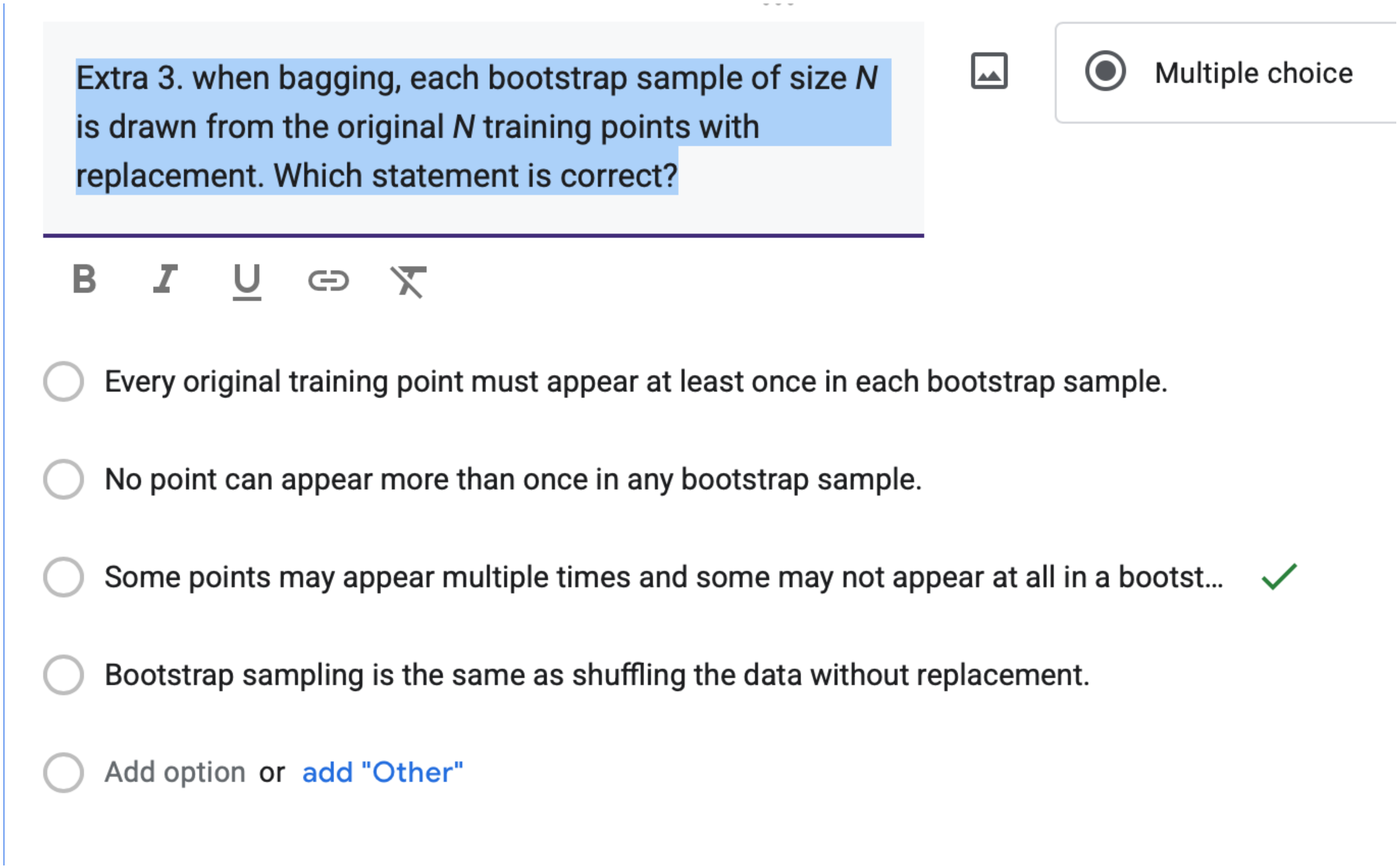
☑ Checkboxes

□ Feature selection ✓

□ Prevent overfitting ✓

□ Improve generalization ✓

□ Add option or add "Other"



Extra 3. when bagging, each bootstrap sample of size *N* is drawn from the original *N* training points with replacement. Which statement is correct?

● Multiple choice

○ Every original training point must appear at least once in each bootstrap sample.

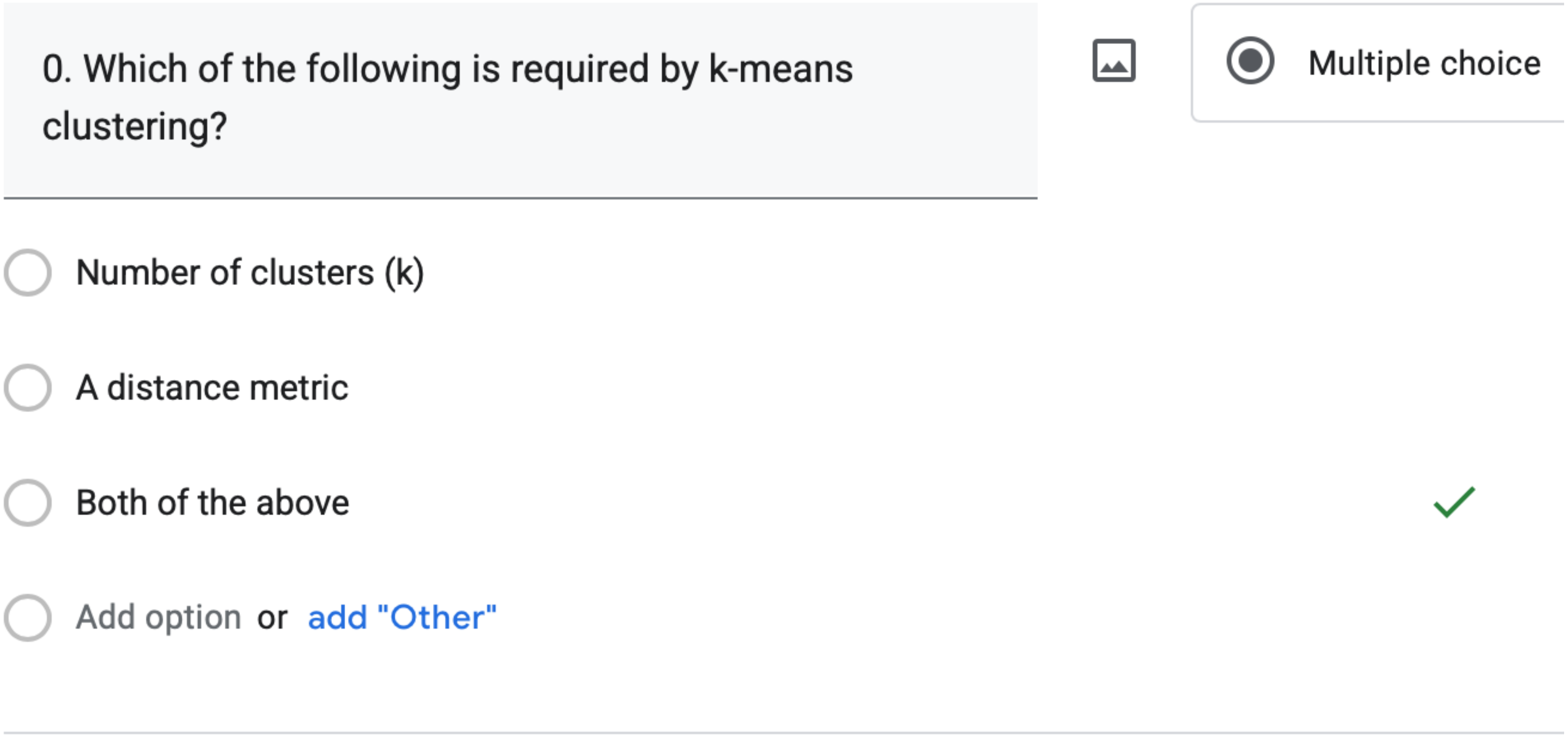
○ No point can appear more than once in any bootstrap sample.

○ Some points may appear multiple times and some may not appear at all in a bootst... ✓

○ Bootstrap sampling is the same as shuffling the data without replacement.

○ Add option or add "Other"

Q13



0. Which of the following is required by k-means clustering?

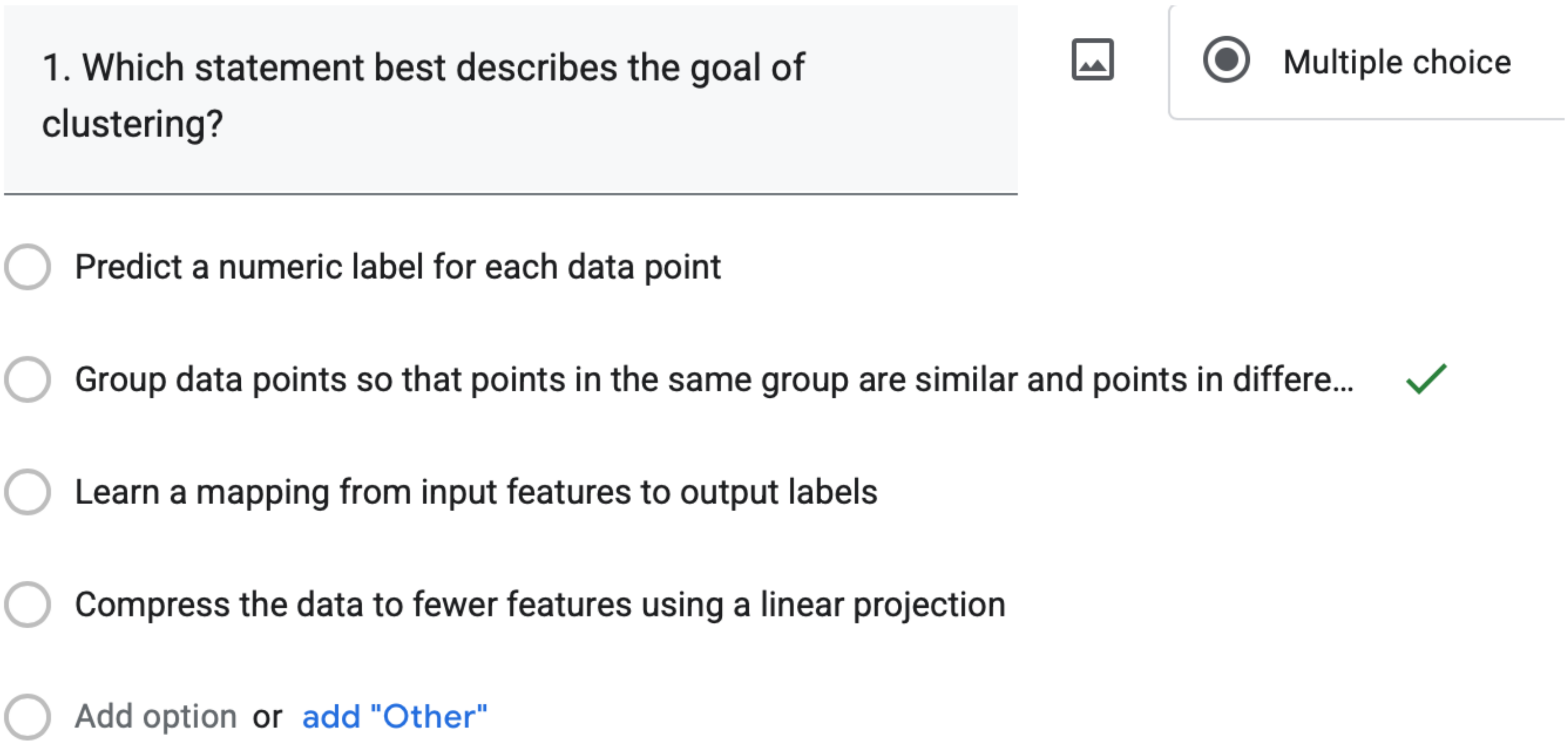
● Multiple choice

○ Number of clusters (k)

○ A distance metric

○ Both of the above ✓

○ Add option or add "Other"



1. Which statement best describes the goal of clustering?

● Multiple choice

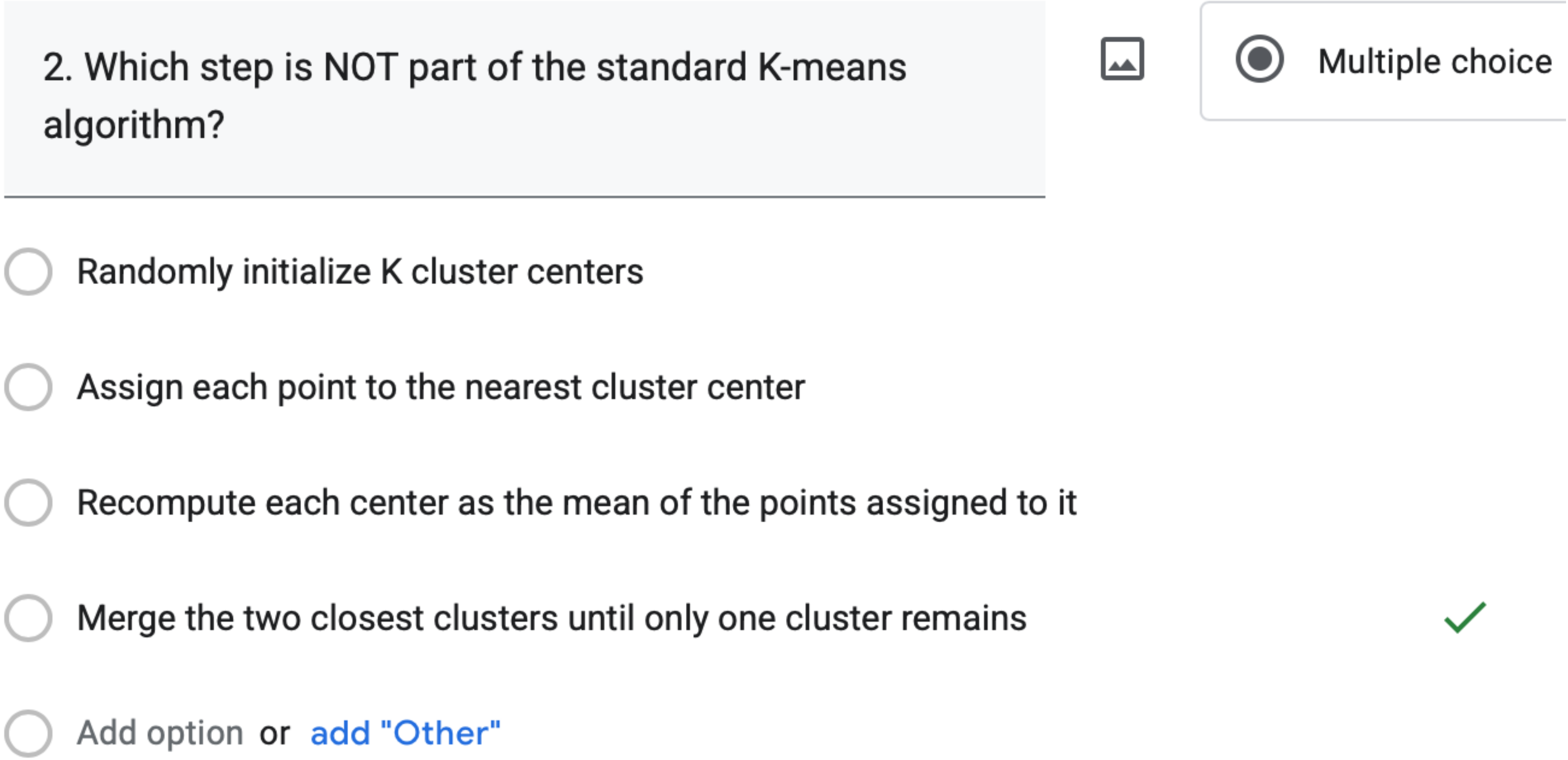
○ Predict a numeric label for each data point

○ Group data points so that points in the same group are similar and points in differe... ✓

○ Learn a mapping from input features to output labels

○ Compress the data to fewer features using a linear projection

○ Add option or add "Other"



2. Which step is NOT part of the standard K-means algorithm?

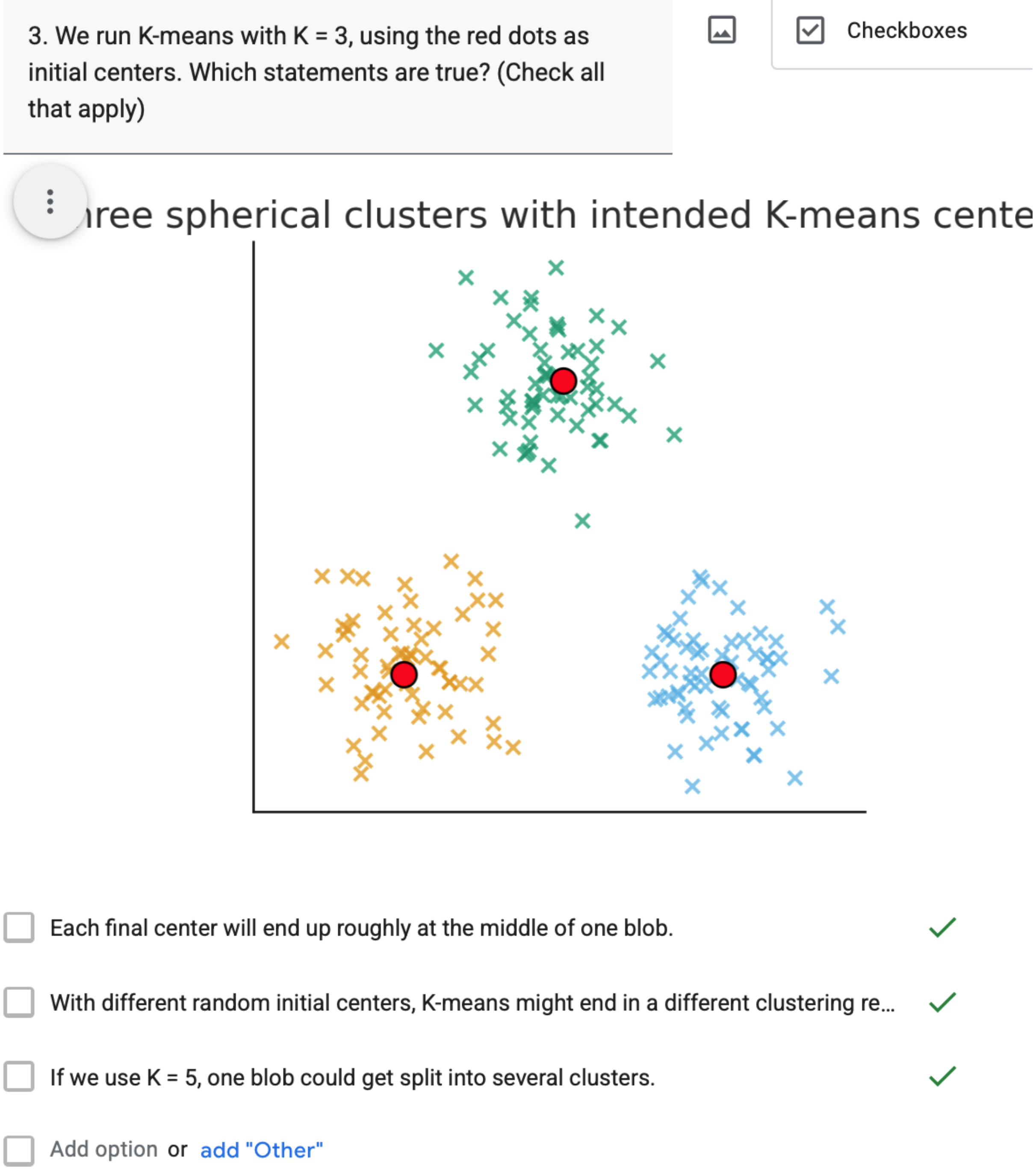
● Multiple choice

○ Randomly initialize K cluster centers

○ Assign each point to the nearest cluster center

○ Recompute each center as the mean of the points assigned to it

○ Merge the two closest clusters until only one cluster remains ✓



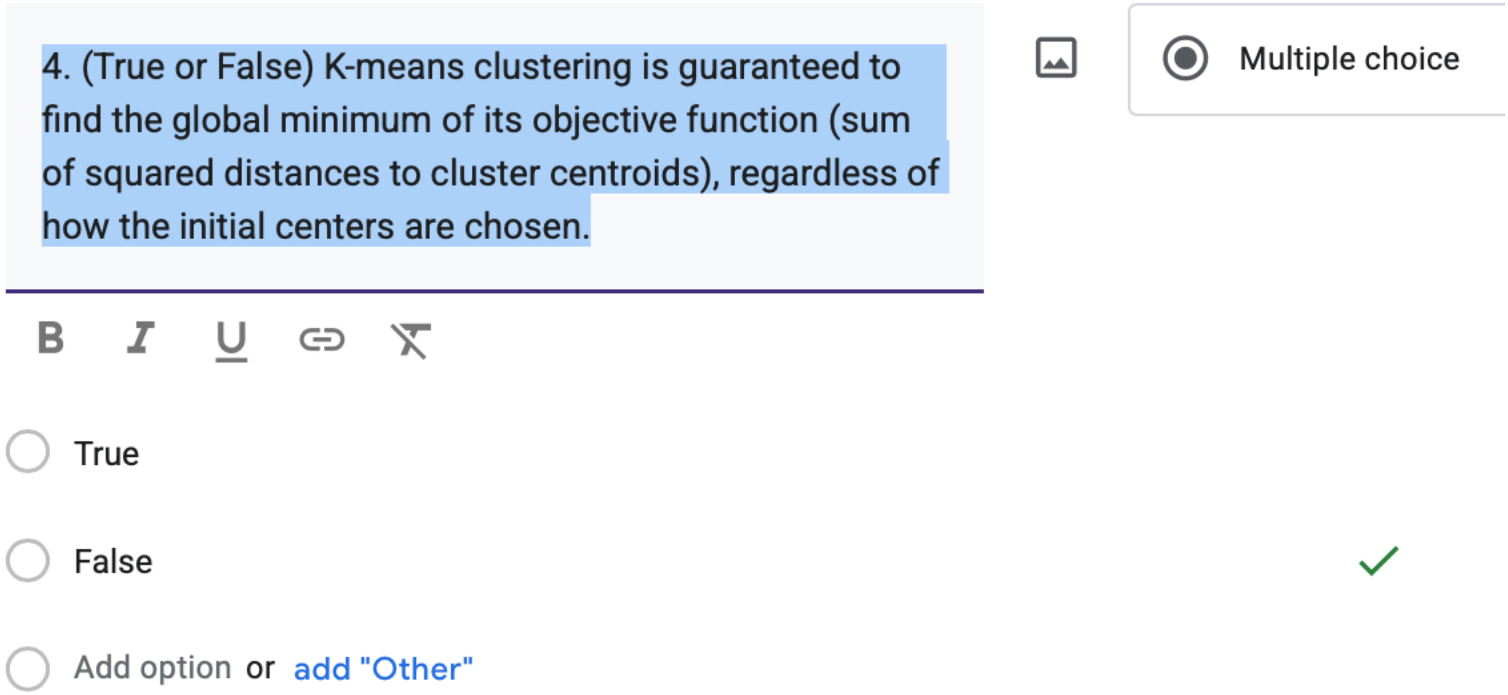
3. We run K-means with K = 3, using the red dots as initial centers. Which statements are true? (Check all that apply)

□ Each final center will end up roughly at the middle of one blob. ✓

□ With different random initial centers, K-means might end in a different clustering re... ✓

□ If we use K = 5, one blob could get split into several clusters. ✓

□ Add option or add "Other"



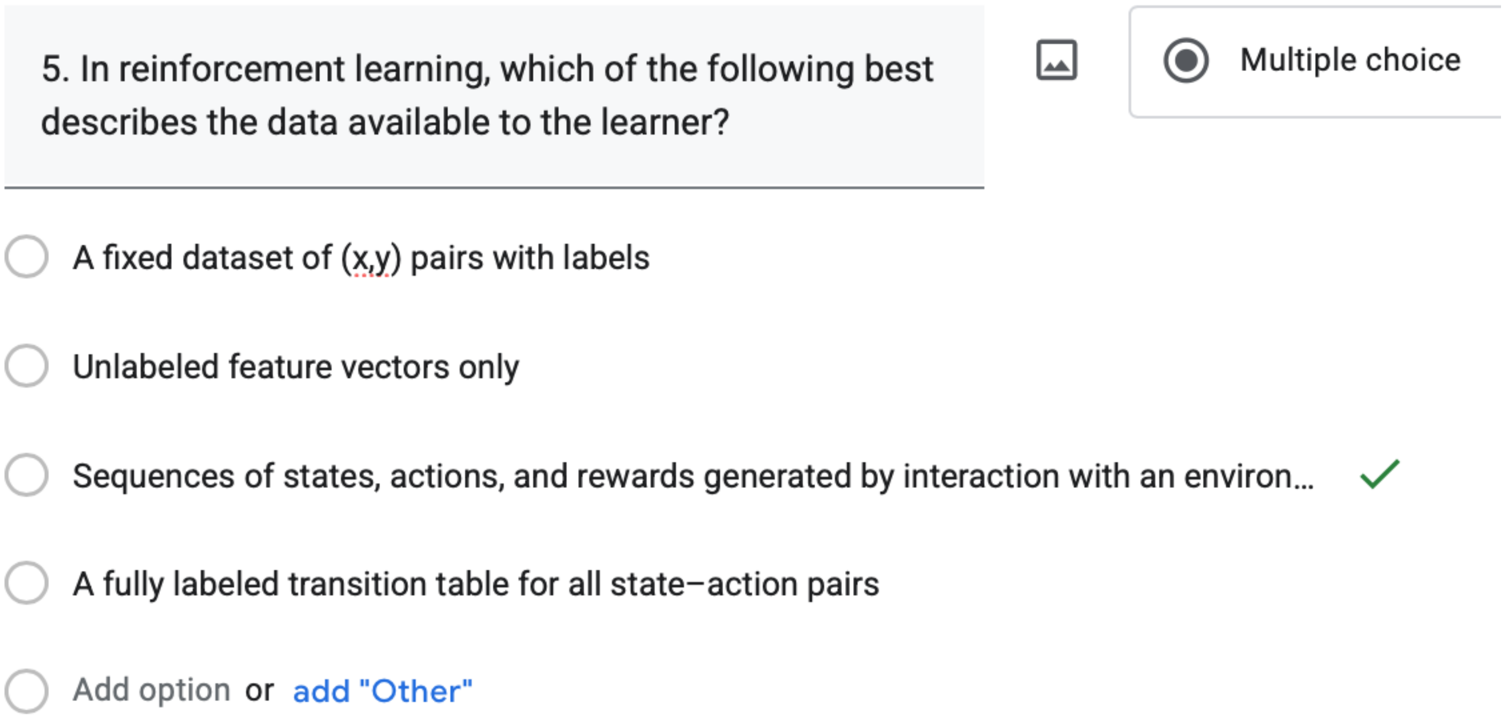
4. (True or False) K-means clustering is guaranteed to find the global minimum of its objective function (sum of squared distances to cluster centroids), regardless of how the initial centers are chosen.

● Multiple choice

○ True

○ False ✓

○ Add option or add "Other"



5. In reinforcement learning, which of the following best describes the data available to the learner?

● Multiple choice

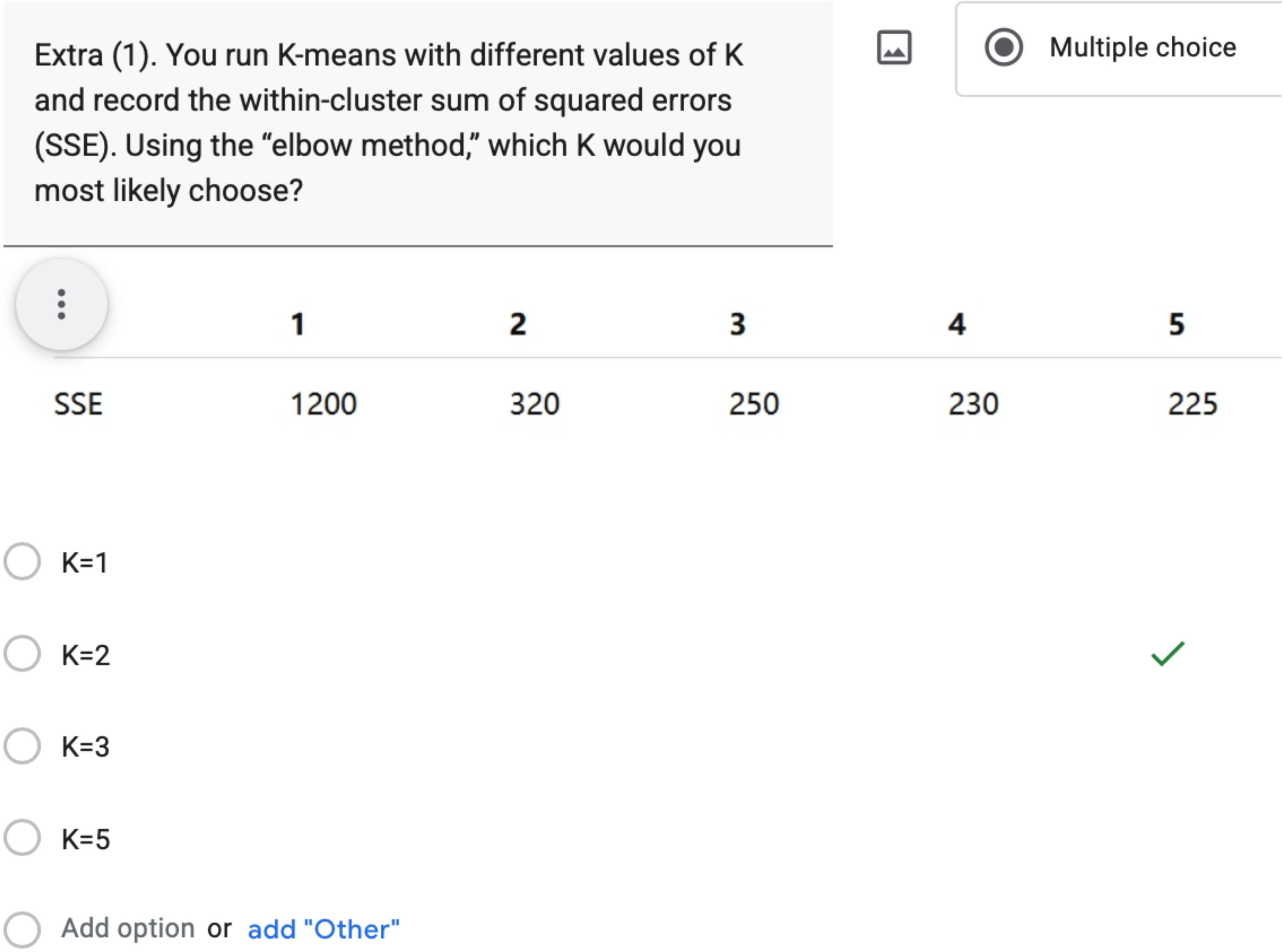
○ A fixed dataset of (x,y) pairs with labels

○ Unlabeled feature vectors only

○ Sequences of states, actions, and rewards generated by interaction with an environ... ✓

○ A fully labeled transition table for all state–action pairs

○ Add option or add "Other"



Extra (1). You run K-means with different values of K and record the within-cluster sum of squared errors (SSE). Using the “elbow method,” which K would you most likely choose?

● Multiple choice

	1	2	3	4	5
SSE	1200	320	250	230	225

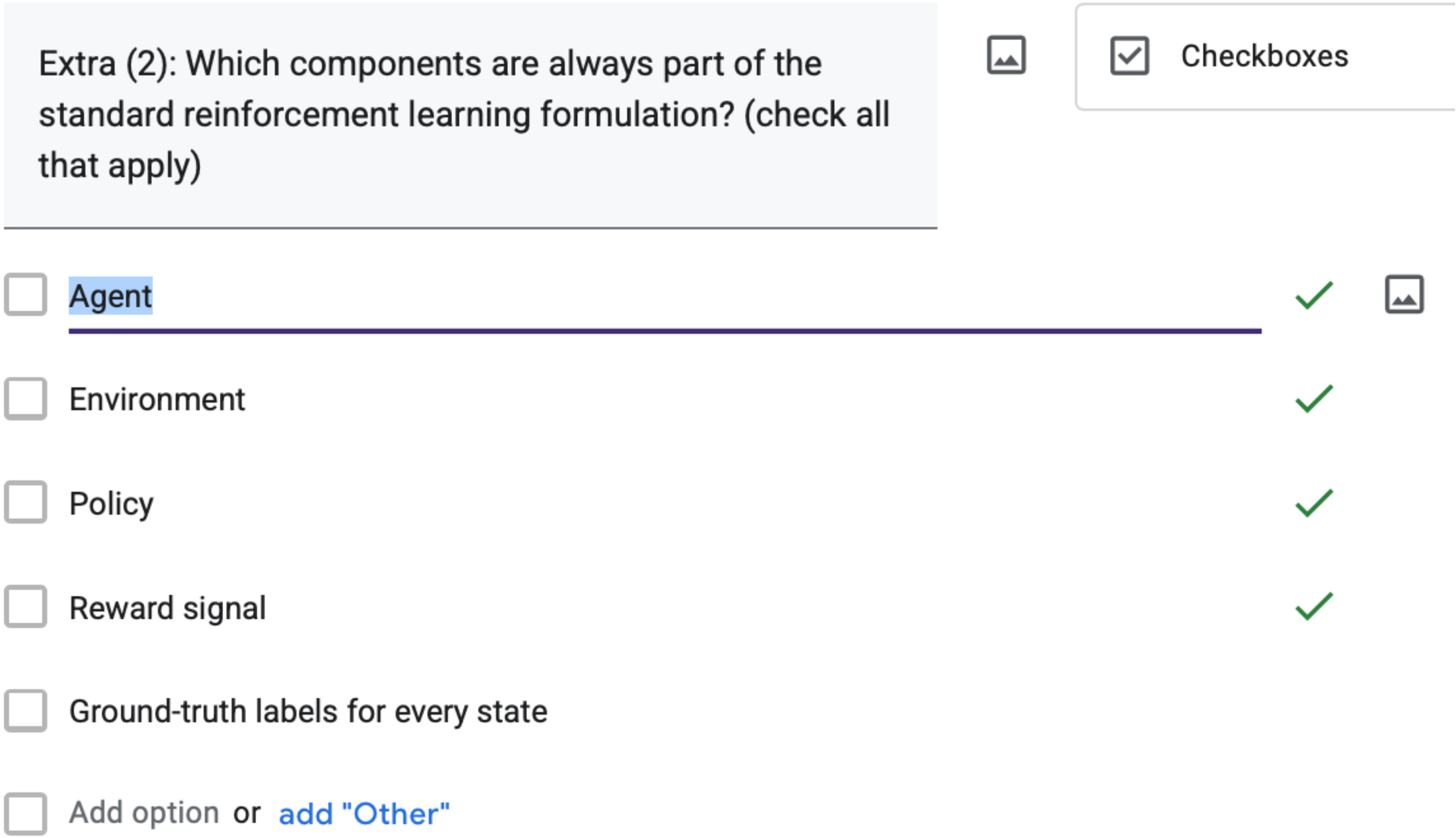
○ K=1

○ K=2 ✓

○ K=3

○ K=5

○ Add option or add "Other"



Extra (2): Which components are always part of the standard reinforcement learning formulation? (check all that apply)

☑ Checkboxes

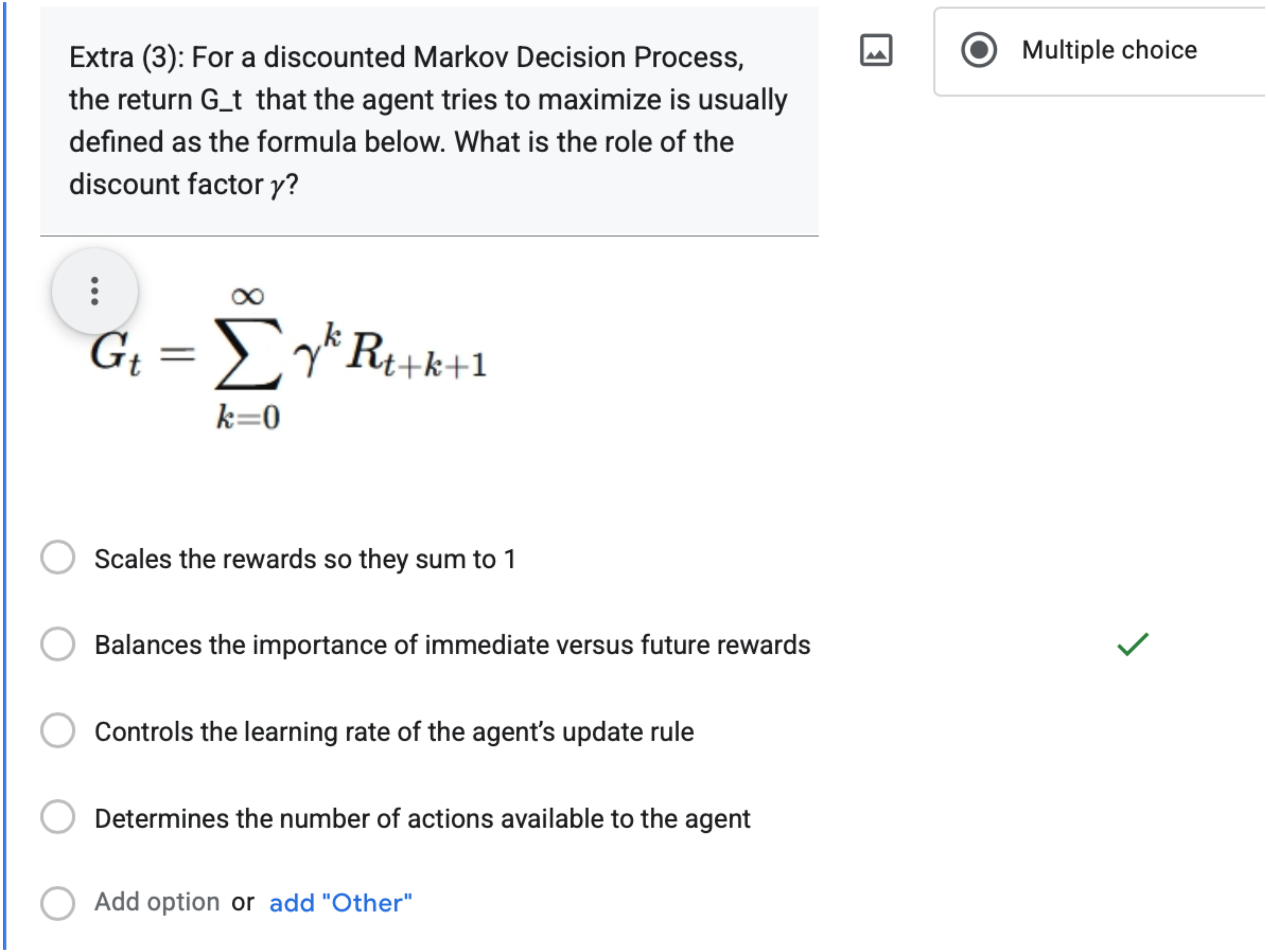
□ Agent ✓

□ Environment ✓

□ Policy ✓

□ Reward signal ✓

□ Ground-truth labels for every state



Extra (3): For a discounted Markov Decision Process, the return G_t that the agent tries to maximize is usually defined as the formula below. What is the role of the discount factor γ?

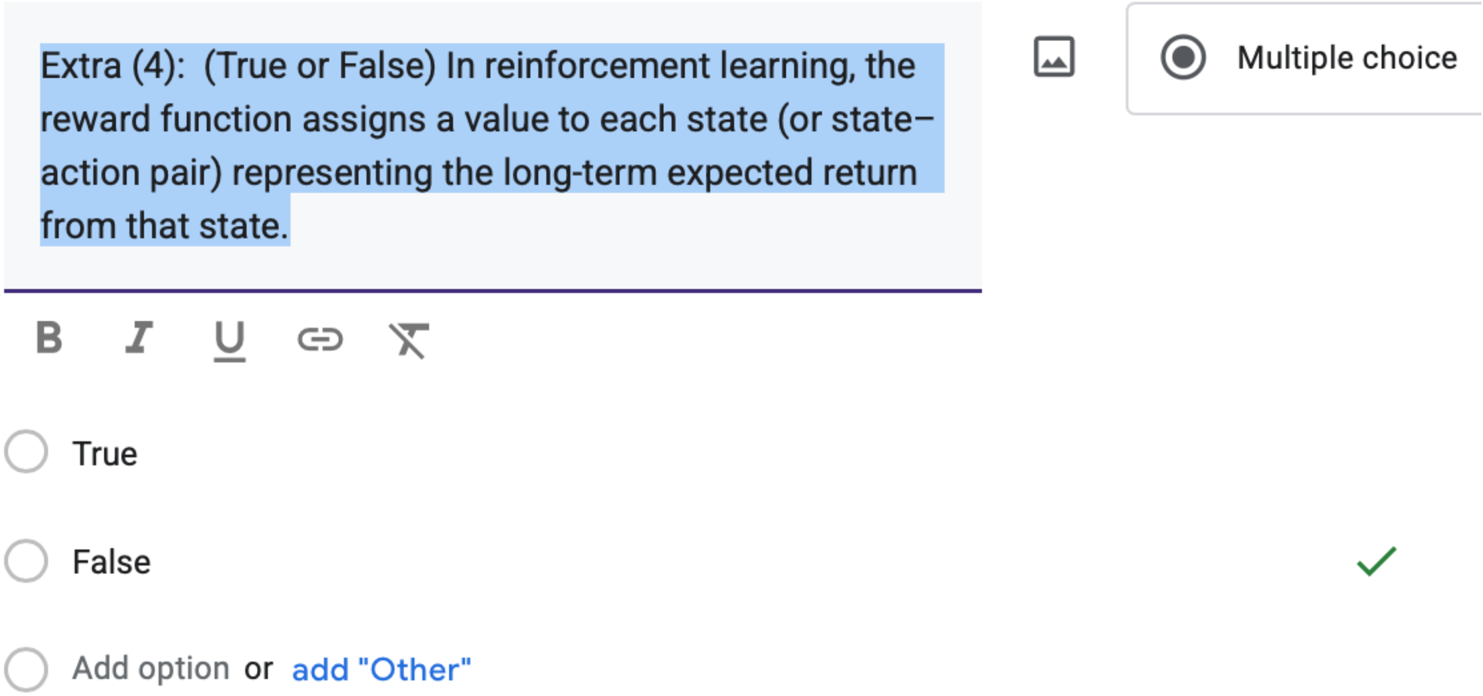
$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

○ Scales the rewards so they sum to 1

○ Balances the importance of immediate versus future rewards ✓

○ Controls the learning rate of the agent’s update rule

○ Determines the number of actions available to the agent



Extra (4): (True or False) In reinforcement learning, the reward function assigns a value to each state (or state–action pair) representing the long-term expected return from that state.

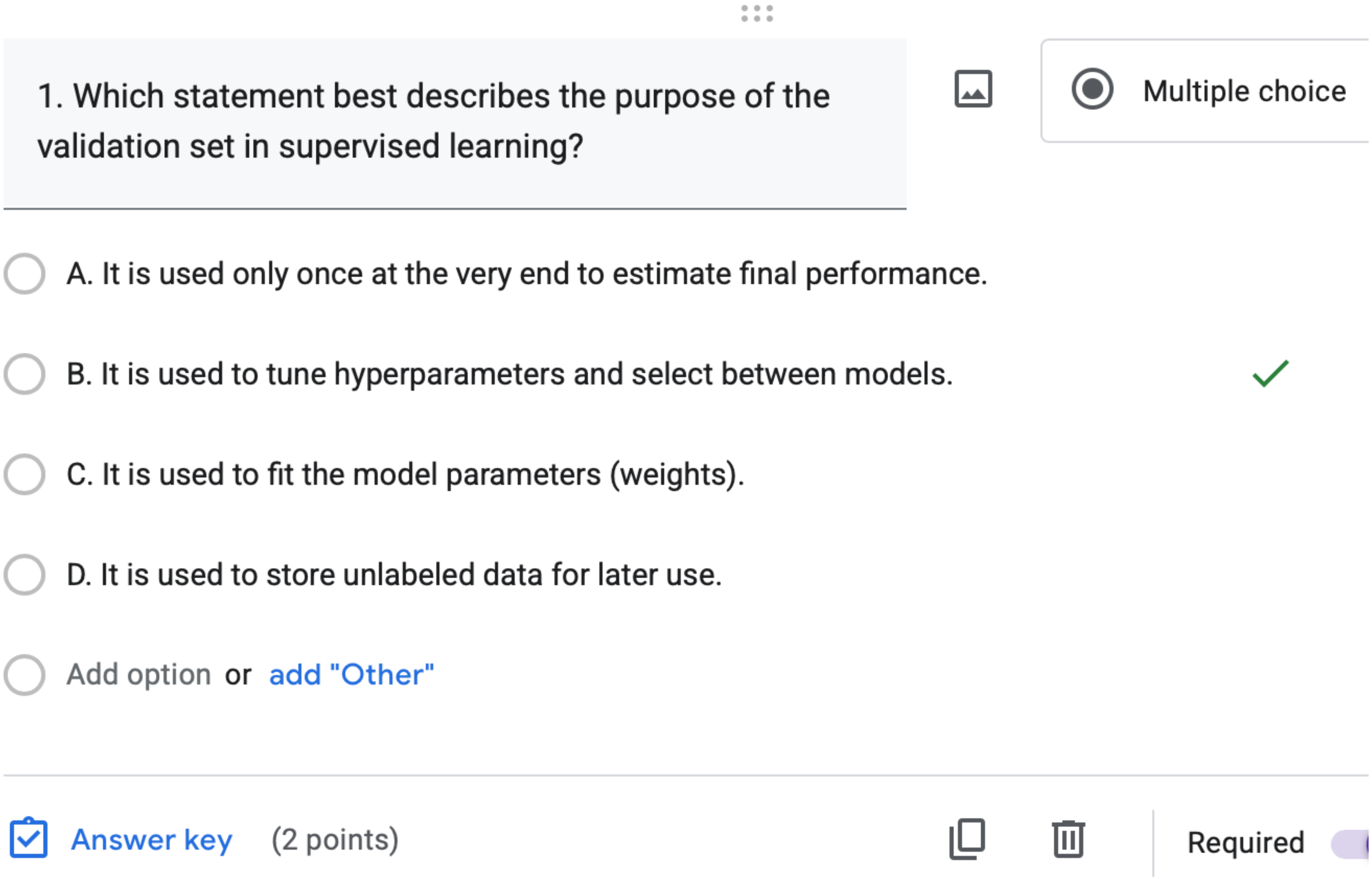
● Multiple choice

○ True

○ False ✓

○ Add option or add "Other"

Q14



1. Which statement best describes the purpose of the validation set in supervised learning?

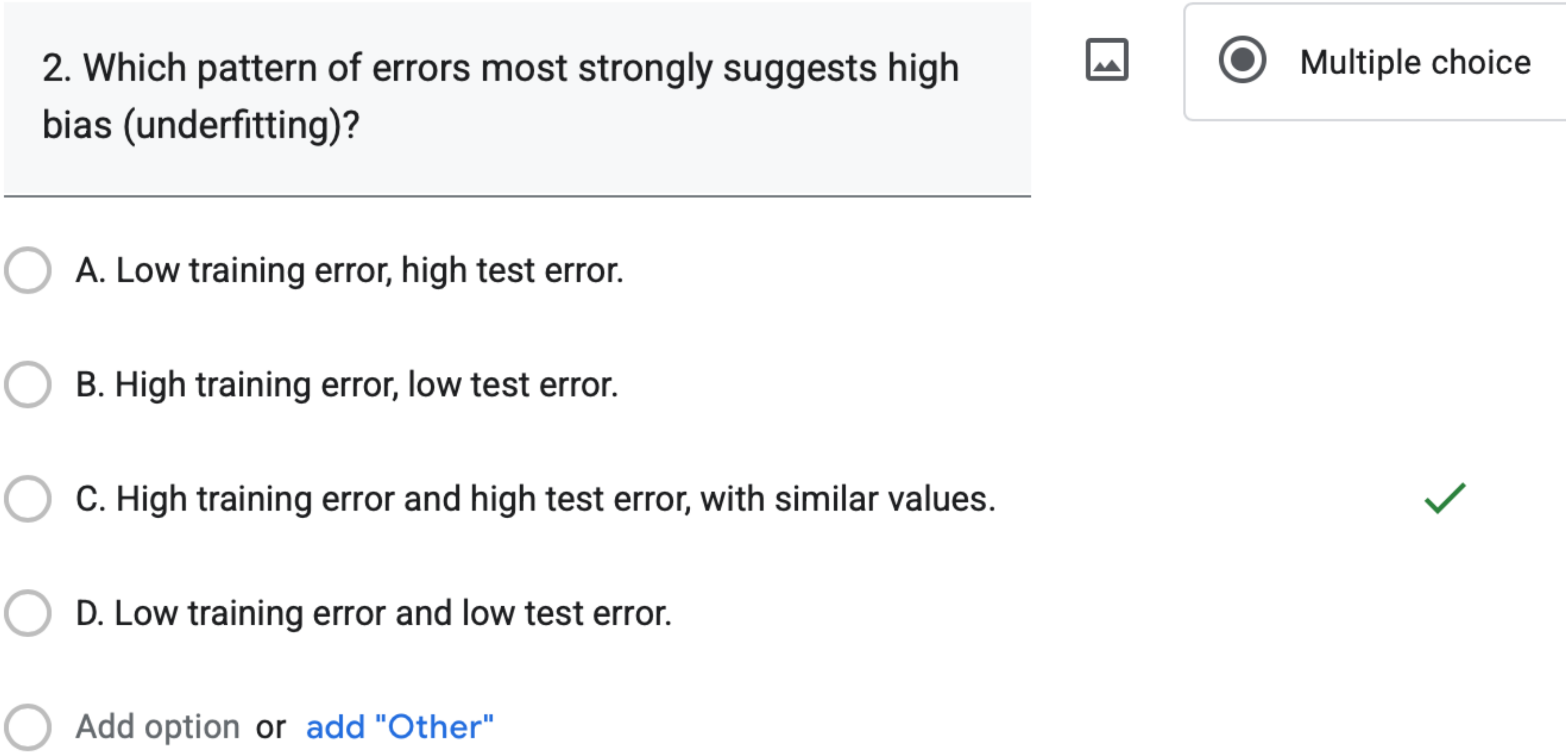
○ A. It is used only once at the very end to estimate final performance.

○ B. It is used to tune hyperparameters and select between models. ✓

○ C. It is used to fit the model parameters (weights).

○ D. It is used to store unlabeled data for later use.

☑ Answer key (2 points)



2. Which pattern of errors most strongly suggests high bias (underfitting)?

● Multiple choice

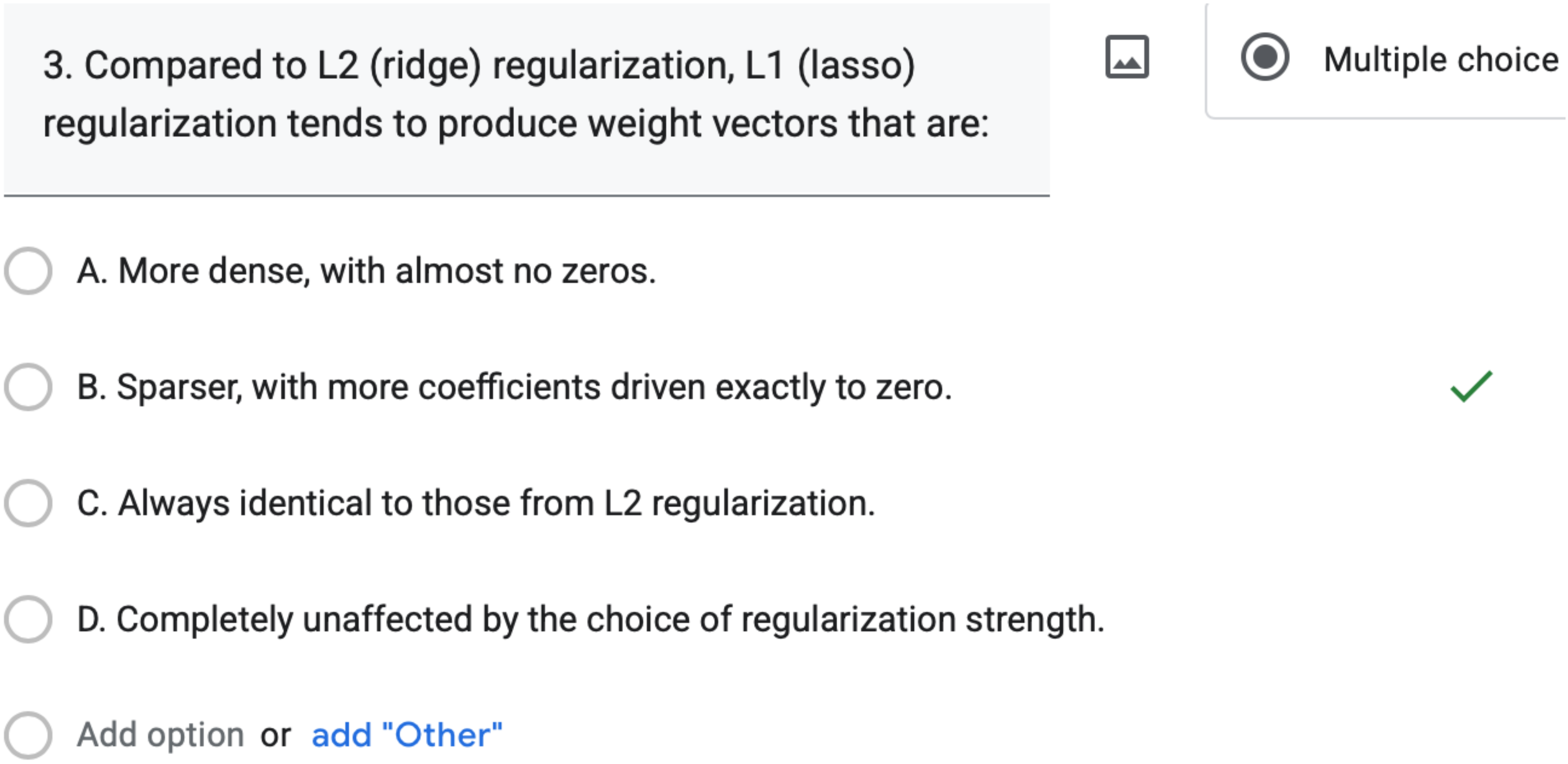
○ A. Low training error, high test error.

○ B. High training error, low test error.

○ C. High training error and high test error, with similar values. ✓

○ D. Low training error and low test error.

○ Add option or add "Other"



3. Compared to L2 (ridge) regularization, L1 (lasso) regularization tends to produce weight vectors that are:

● Multiple choice

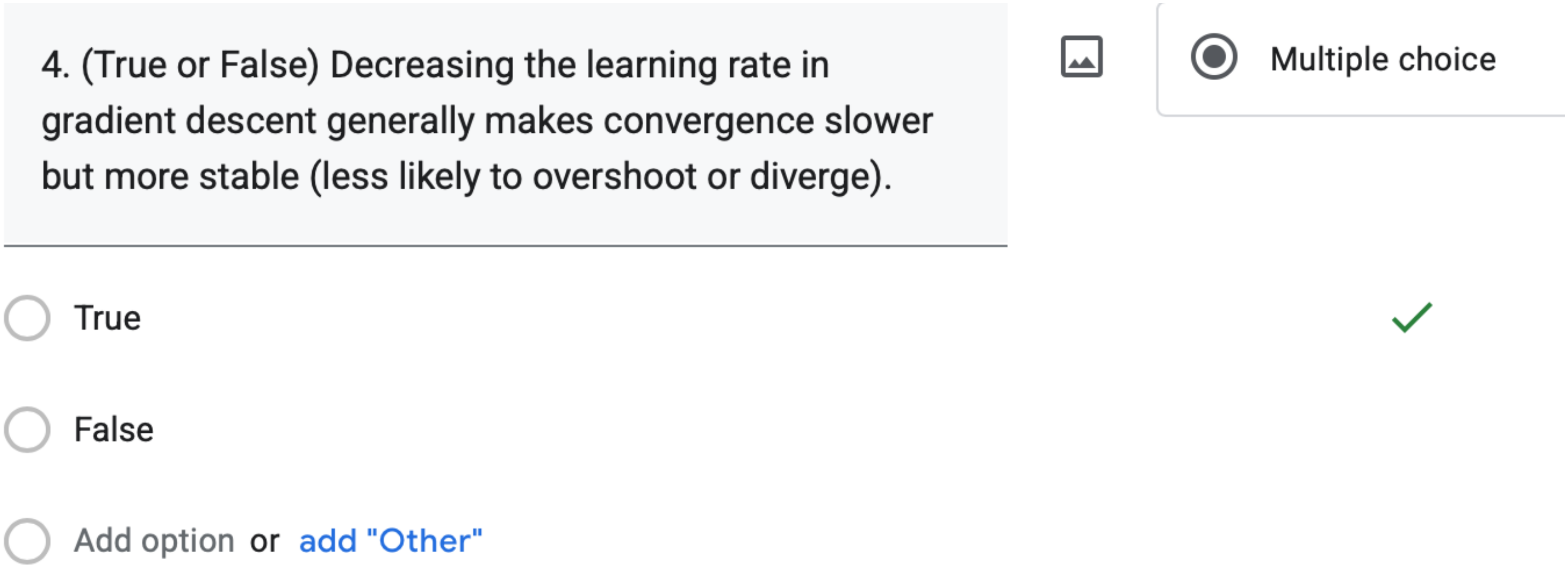
○ A. More dense, with almost no zeros.

○ B. Sparser, with more coefficients driven exactly to zero. ✓

○ C. Always identical to those from L2 regularization.

○ D. Completely unaffected by the choice of regularization strength.

○ Add option or add "Other"

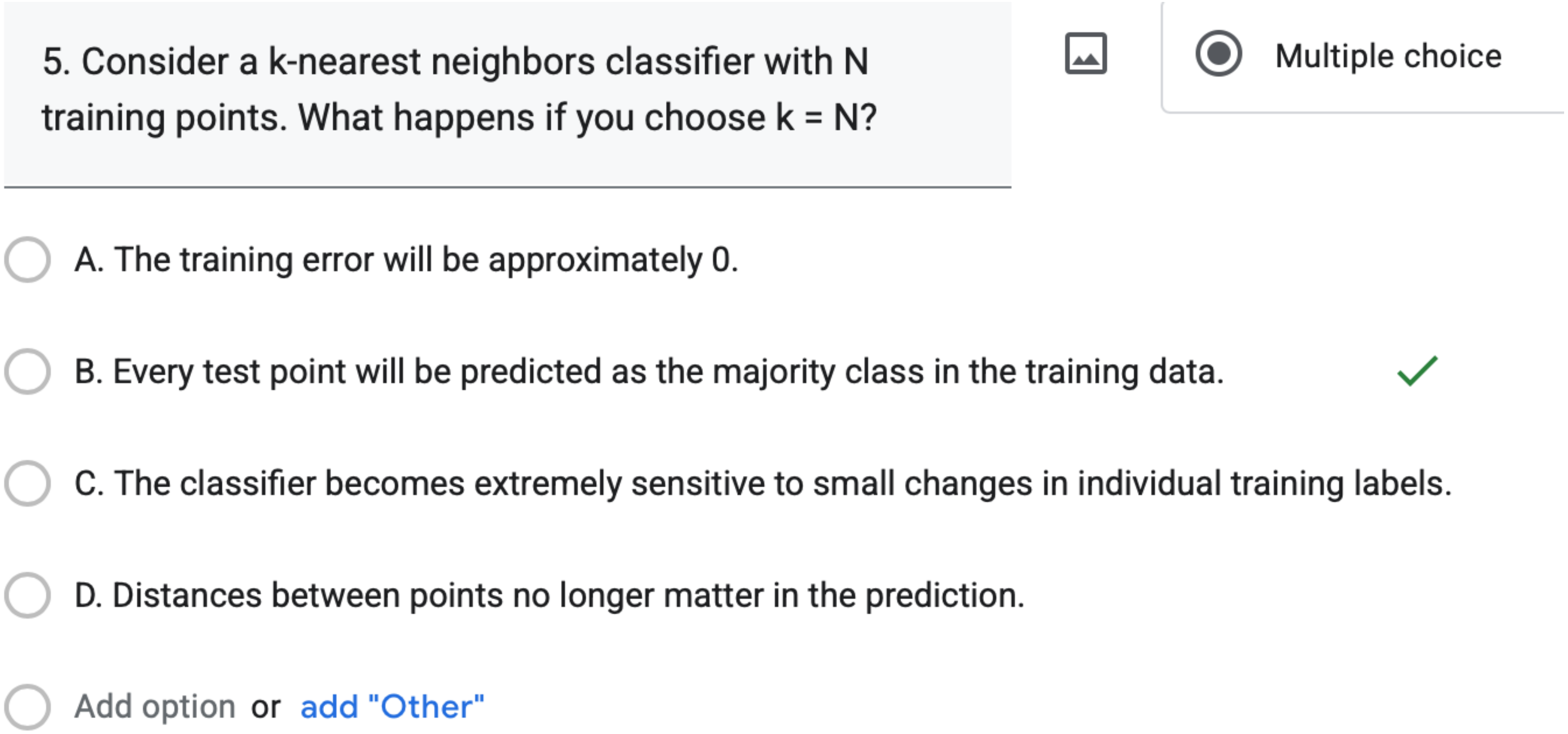


4. (True or False) Decreasing the learning rate in gradient descent generally makes convergence slower but more stable (less likely to overshoot or diverge).

○ True ✓

○ False

○ Add option or add "Other"



5. Consider a k-nearest neighbors classifier with N training points. What happens if you choose k = N?

● Multiple choice

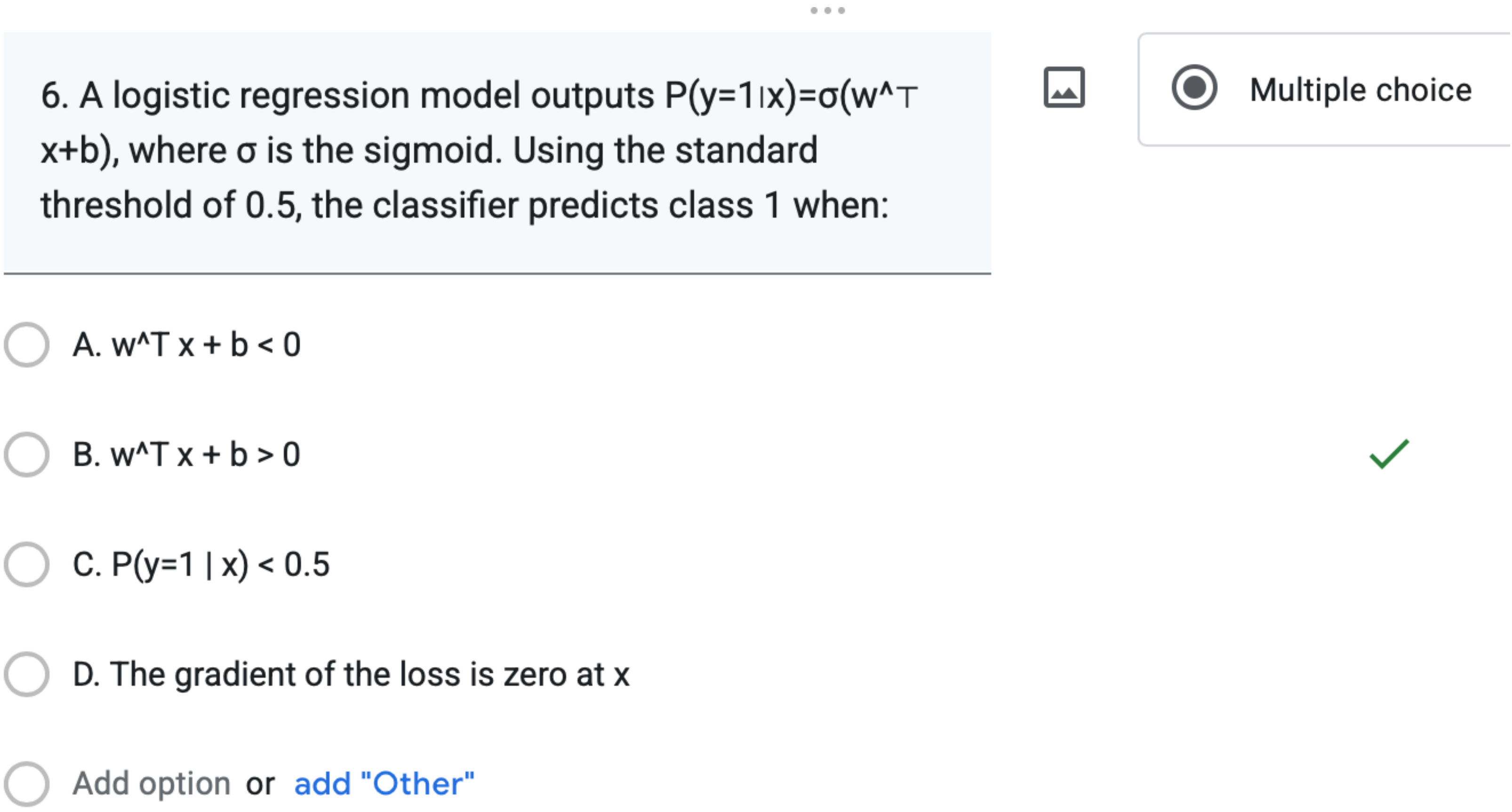
○ A. The training error will be approximately 0.

○ B. Every test point will be predicted as the majority class in the training data. ✓

○ C. The classifier becomes extremely sensitive to small changes in individual training labels.

○ D. Distances between points no longer matter in the prediction.

○ Add option or add "Other"



6. A logistic regression model outputs $P(y=1|x)=\sigma(w^T x+b)$, where σ is the sigmoid. Using the standard threshold of 0.5, the classifier predicts class 1 when:

● Multiple choice

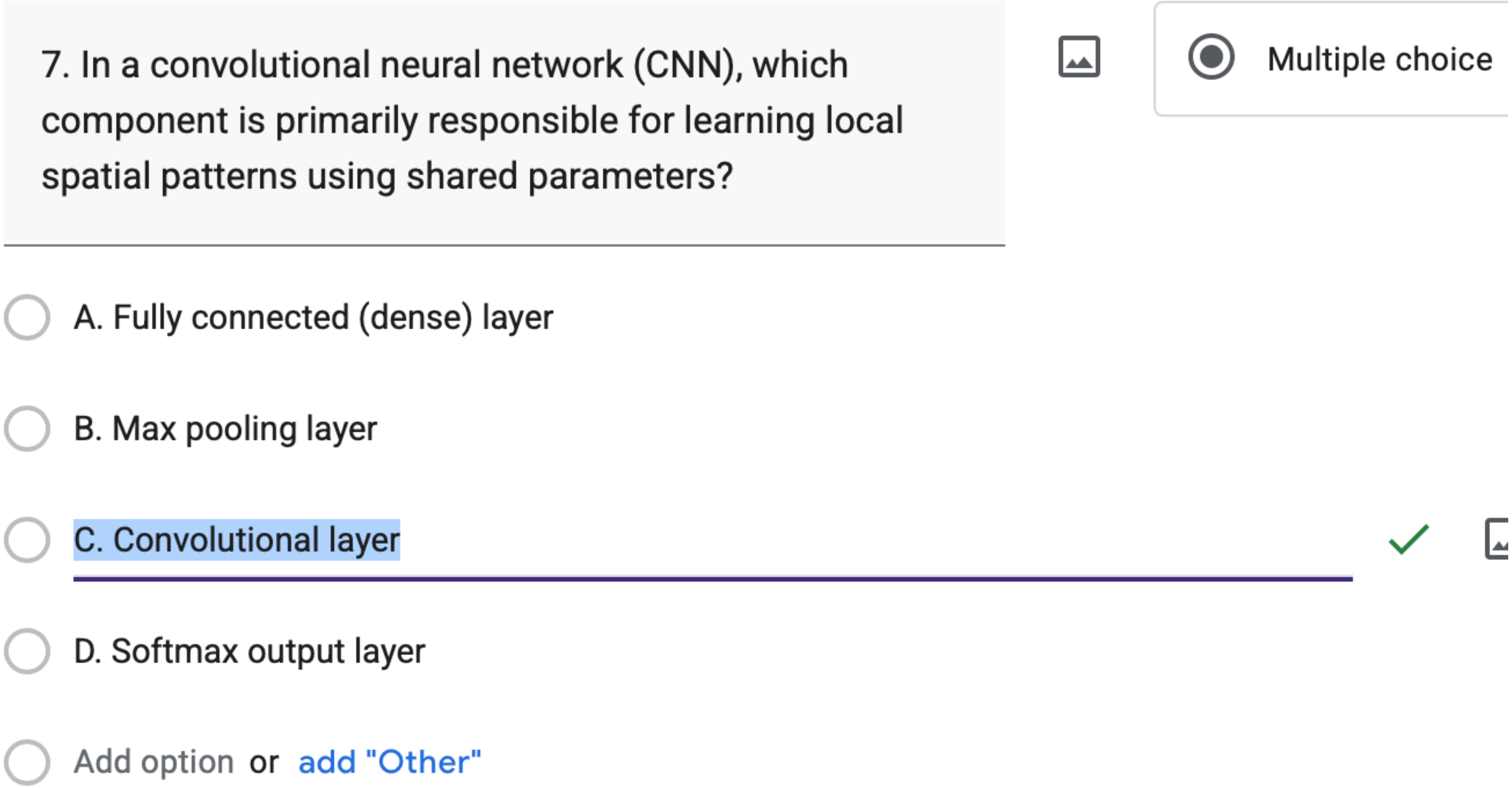
○ A. $w^T x + b < 0$

○ B. $w^T x + b > 0$ ✓

○ C. $P(y=1 \mid x) < 0.5$

○ D. The gradient of the loss is zero at x

○ Add option or add "Other"



7. In a convolutional neural network (CNN), which component is primarily responsible for learning local spatial patterns using shared parameters?

● Multiple choice

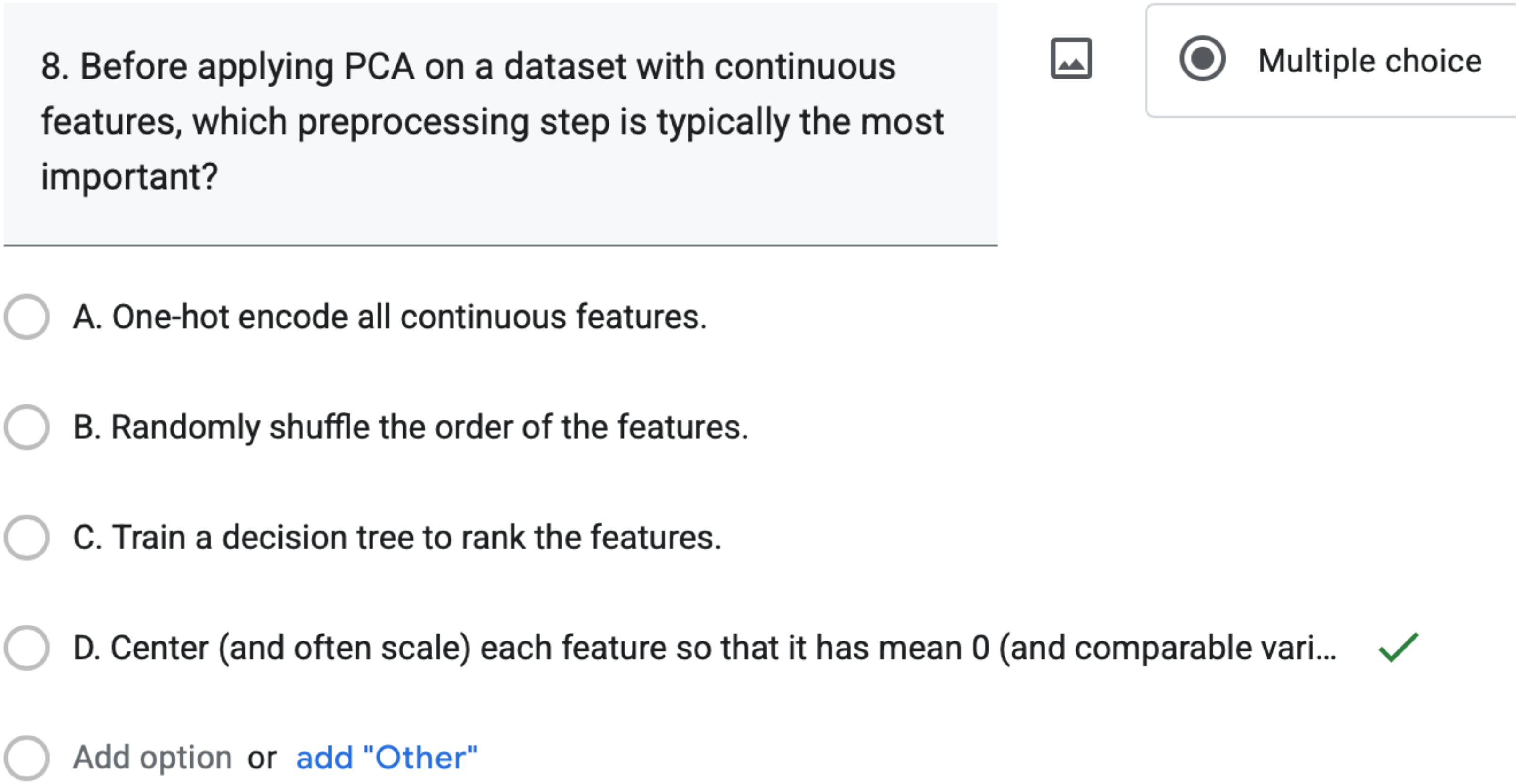
○ A. Fully connected (dense) layer

○ B. Max pooling layer

○ C. Convolutional layer ✓

○ D. Softmax output layer

○ Add option or add "Other"



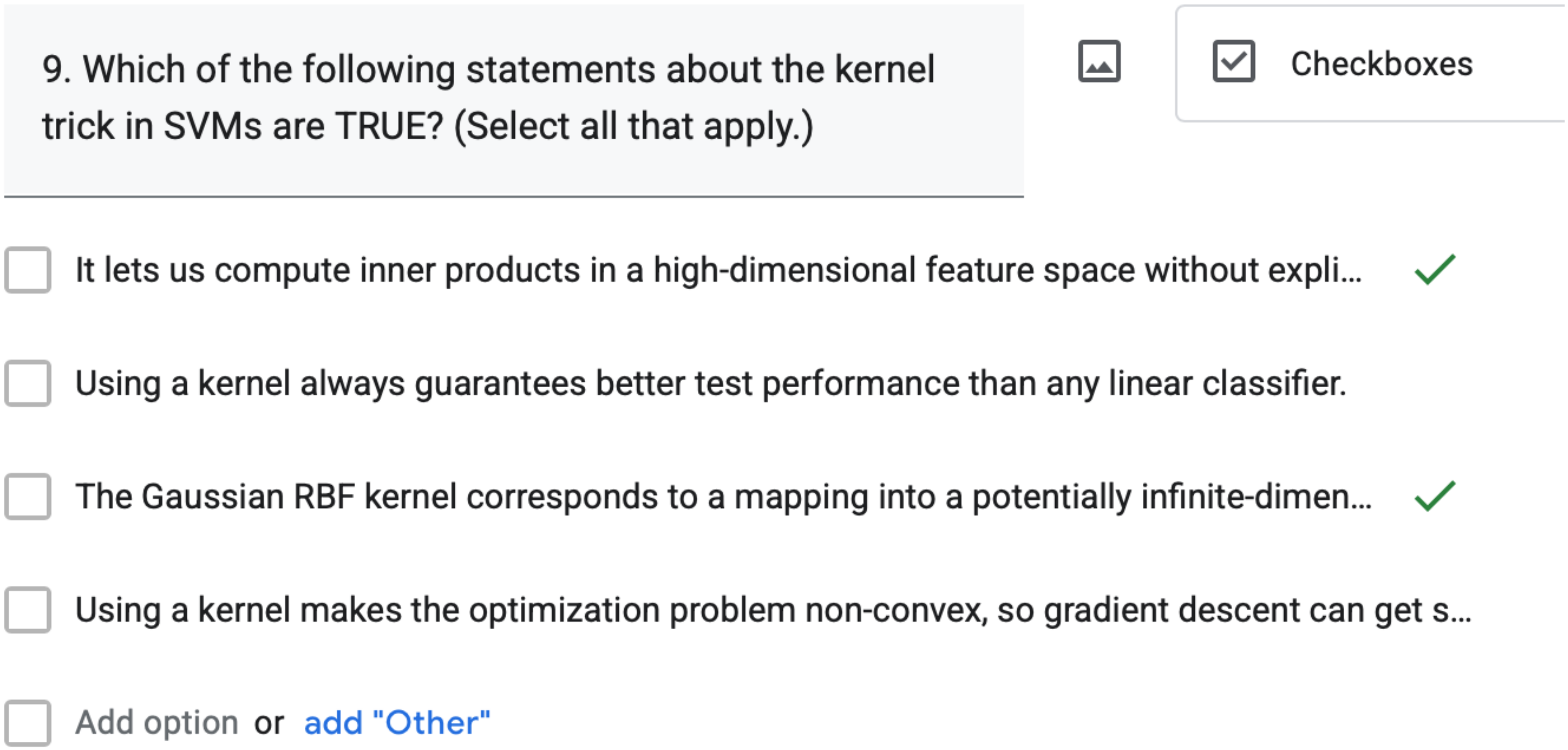
8. Before applying PCA on a dataset with continuous features, which preprocessing step is typically the most important?

○ A. One-hot encode all continuous features.

○ B. Randomly shuffle the order of the features.

○ C. Train a decision tree to rank the features.

○ D. Center (and often scale) each feature so that it has mean 0 (and comparable vari... ✓



9. Which of the following statements about the kernel trick in SVMs are TRUE? (Select all that apply.)

☑ Checkboxes

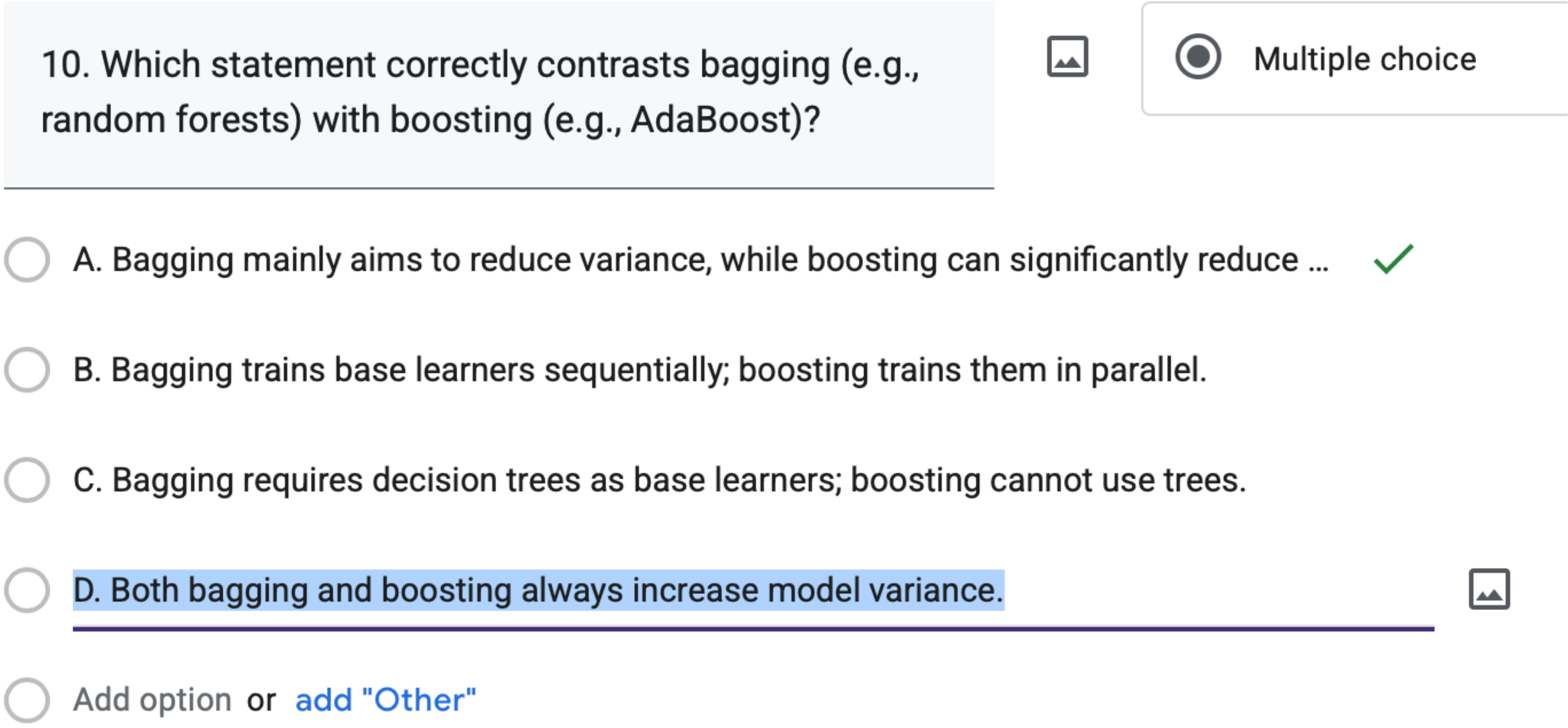
□ It lets us compute inner products in a high-dimensional feature space without expli… ✓

□ Using a kernel always guarantees better test performance than any linear classifier.

□ The Gaussian RBF kernel corresponds to a mapping into a potentially infinite-dimen… ✓

□ Using a kernel makes the optimization problem non-convex, so gradient descent can get s…

□ Add option or add "Other"



10. Which statement correctly contrasts bagging (e.g., random forests) with boosting (e.g., AdaBoost)?

● Multiple choice

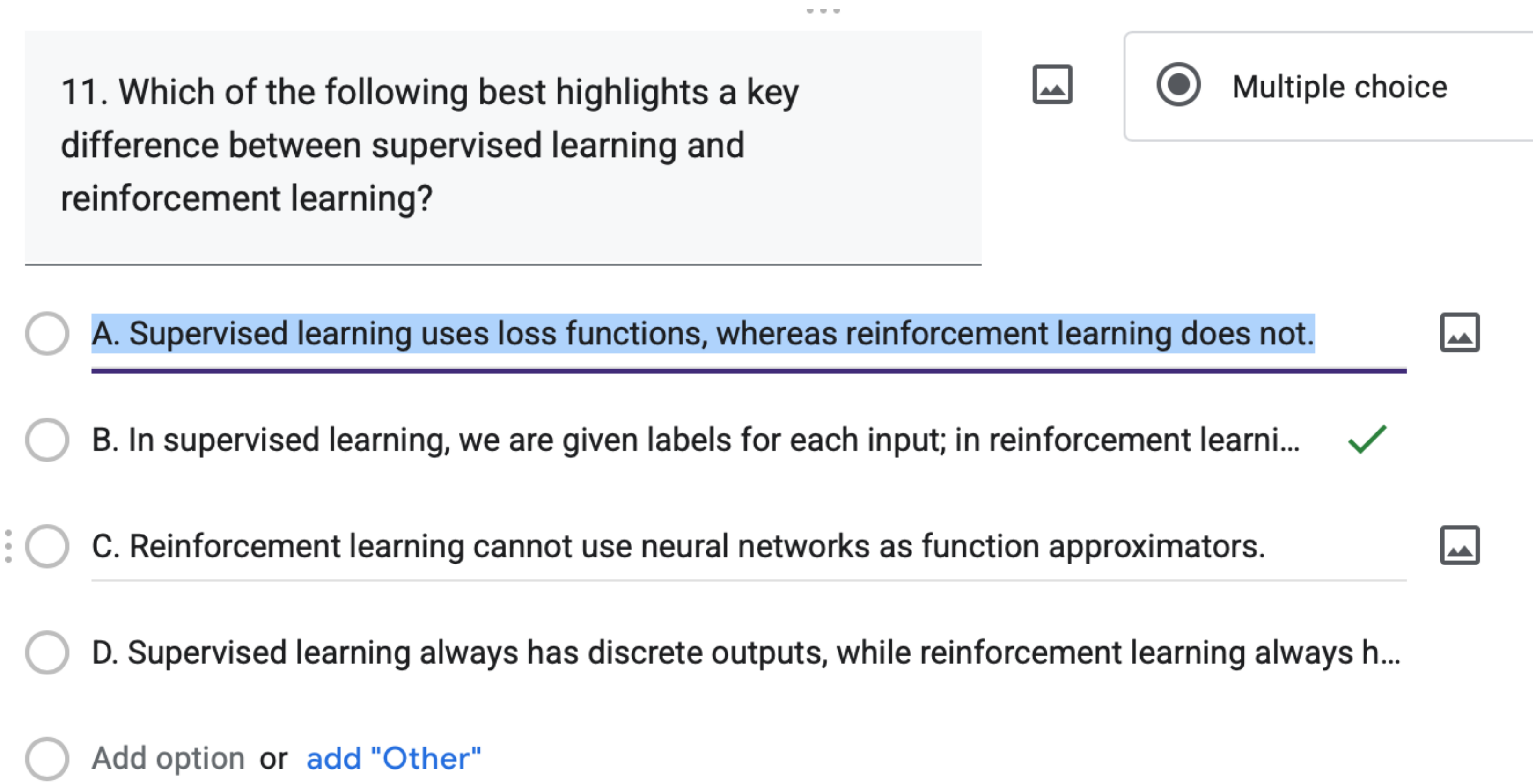
○ A. Bagging mainly aims to reduce variance, while boosting can significantly reduce ... ✓

○ B. Bagging trains base learners sequentially; boosting trains them in parallel.

○ C. Bagging requires decision trees as base learners; boosting cannot use trees.

○ D. Both bagging and boosting always increase model variance.

○ Add option or add "Other"



11. Which of the following best highlights a key difference between supervised learning and reinforcement learning?

● Multiple choice

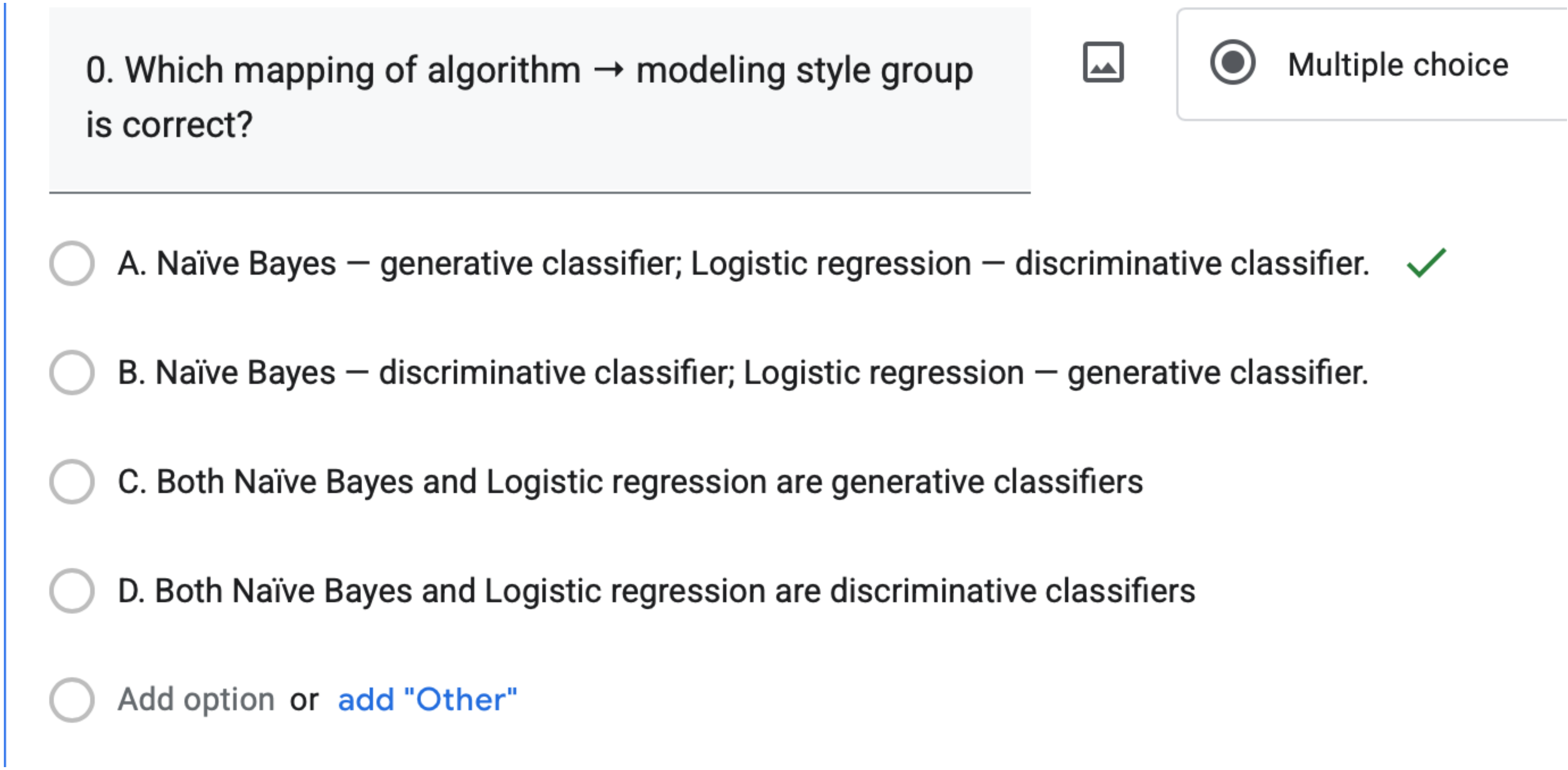
○ A. Supervised learning uses loss functions, whereas reinforcement learning does not.

○ B. In supervised learning, we are given labels for each input; in reinforcement learni... ✓

○ C. Reinforcement learning cannot use neural networks as function approximators.

○ D. Supervised learning always has discrete outputs, while reinforcement learning always h...

Q15



0. Which mapping of algorithm → modeling style group is correct?

● Multiple choice

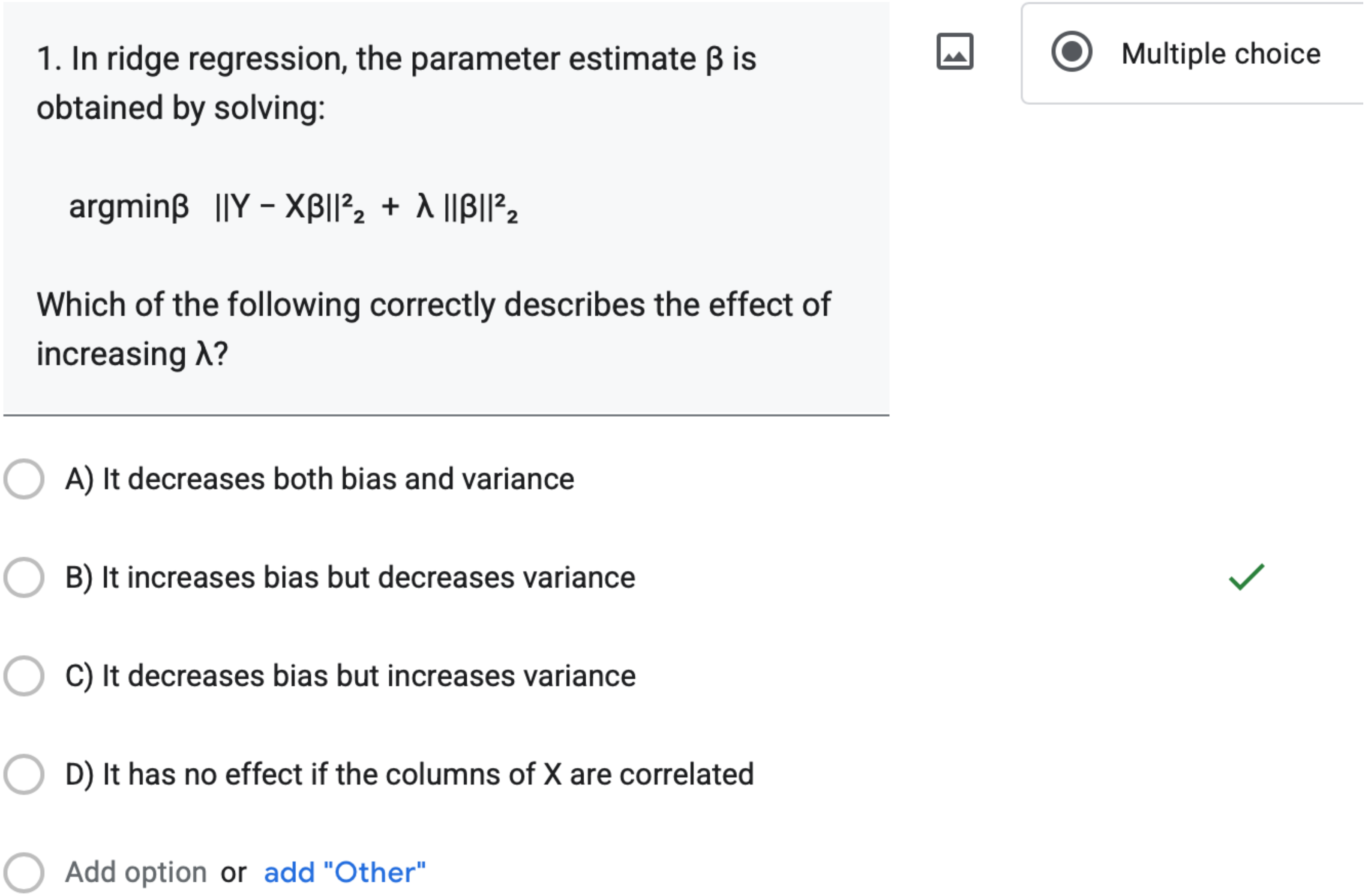
○ A. Naïve Bayes — generative classifier; Logistic regression — discriminative classifier. ✓

○ B. Naïve Bayes — discriminative classifier; Logistic regression — generative classifier.

○ C. Both Naïve Bayes and Logistic regression are generative classifiers

○ D. Both Naïve Bayes and Logistic regression are discriminative classifiers

○ Add option or add "Other"



1. In ridge regression, the parameter estimate β is obtained by solving:

$$\text{argmin}\beta \quad ||Y - X\beta||^2_2 + \lambda ||\beta||^2_2$$

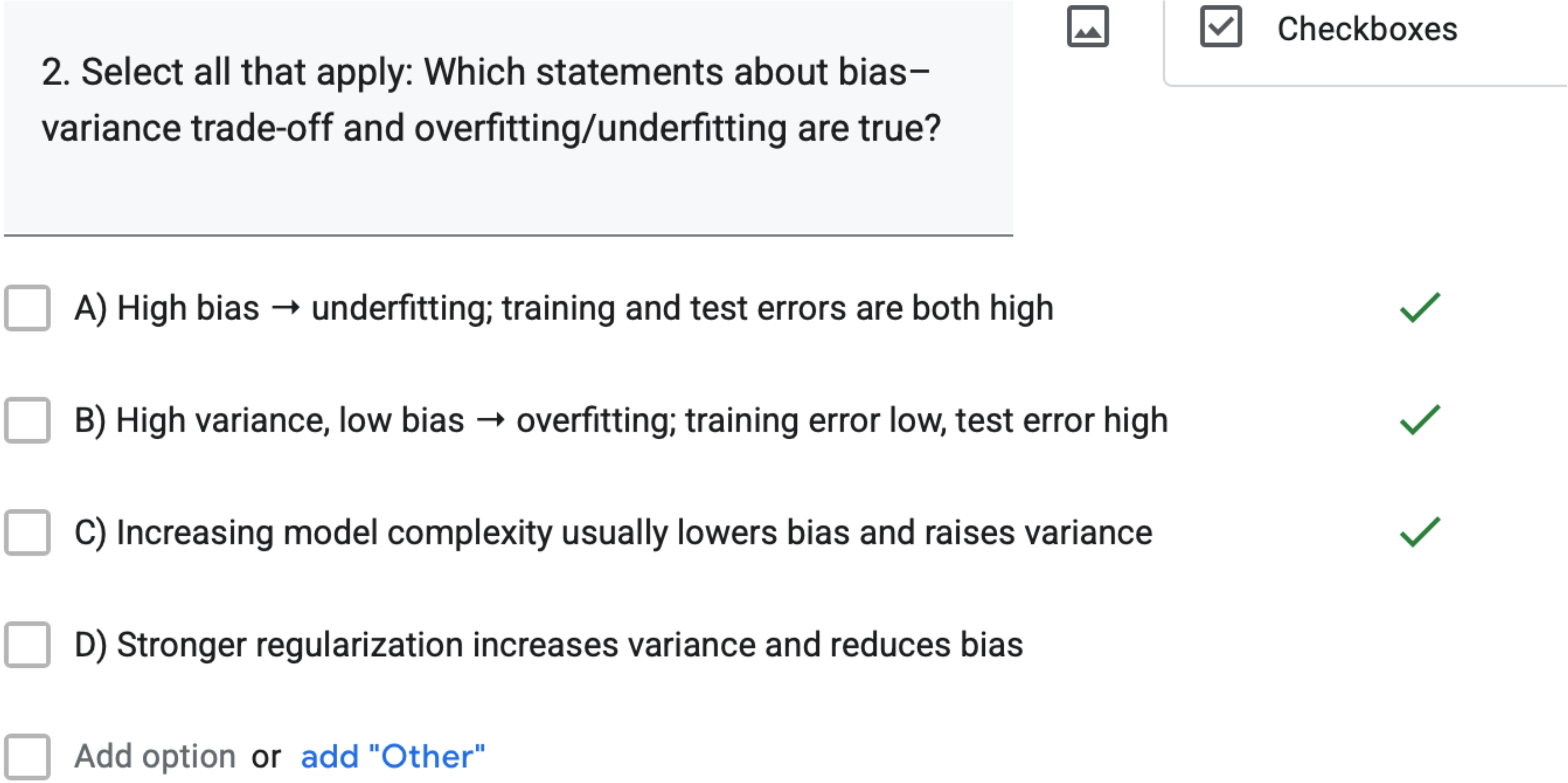
Which of the following correctly describes the effect of increasing λ?

○ A) It decreases both bias and variance

○ B) It increases bias but decreases variance ✓

○ C) It decreases bias but increases variance

○ D) It has no effect if the columns of X are correlated



2. Select all that apply: Which statements about bias–variance trade-off and overfitting/underfitting are true?

☑ Checkboxes

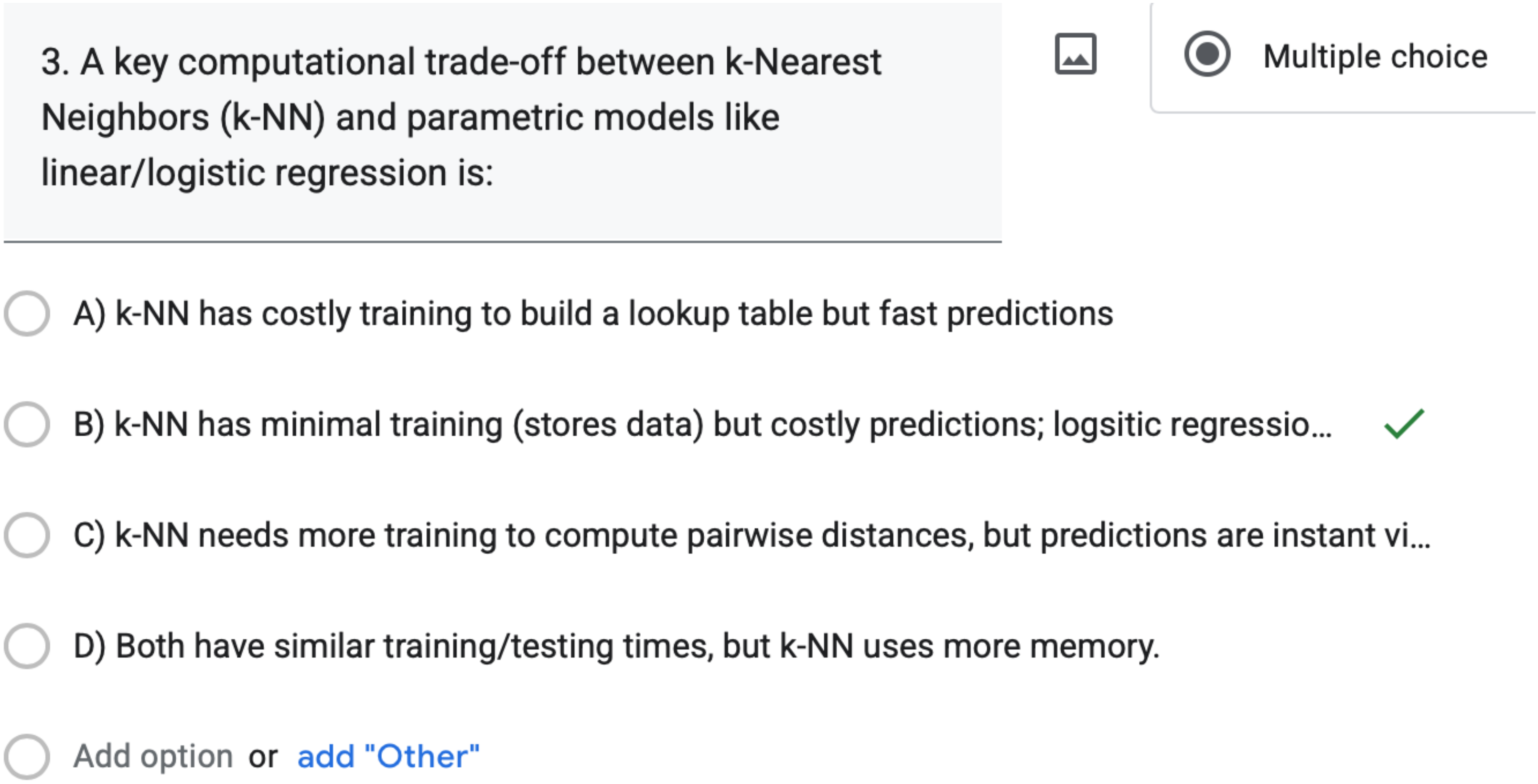
□ A) High bias → underfitting; training and test errors are both high ✓

□ B) High variance, low bias → overfitting; training error low, test error high ✓

□ C) Increasing model complexity usually lowers bias and raises variance ✓

□ D) Stronger regularization increases variance and reduces bias

□ Add option or add "Other"



3. A key computational trade-off between k-Nearest Neighbors (k-NN) and parametric models like linear/logistic regression is:

● Multiple choice

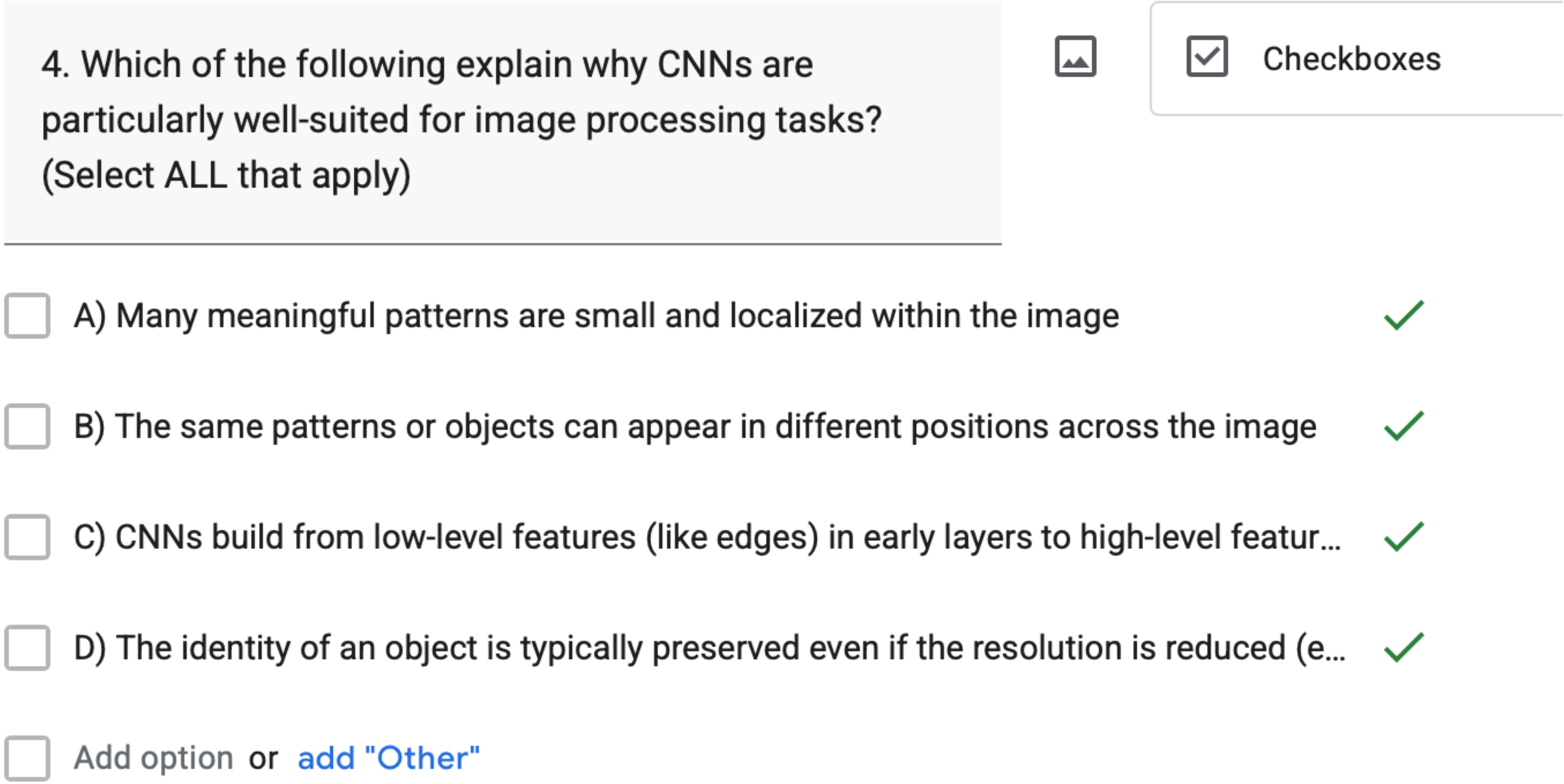
○ A) k-NN has costly training to build a lookup table but fast predictions

○ B) k-NN has minimal training (stores data) but costly predictions; logsitic regressio... ✓

○ C) k-NN needs more training to compute pairwise distances, but predictions are instant vi...

○ D) Both have similar training/testing times, but k-NN uses more memory.

○ Add option or add "Other"



4. Which of the following explain why CNNs are particularly well-suited for image processing tasks? (Select ALL that apply)

☑ Checkboxes

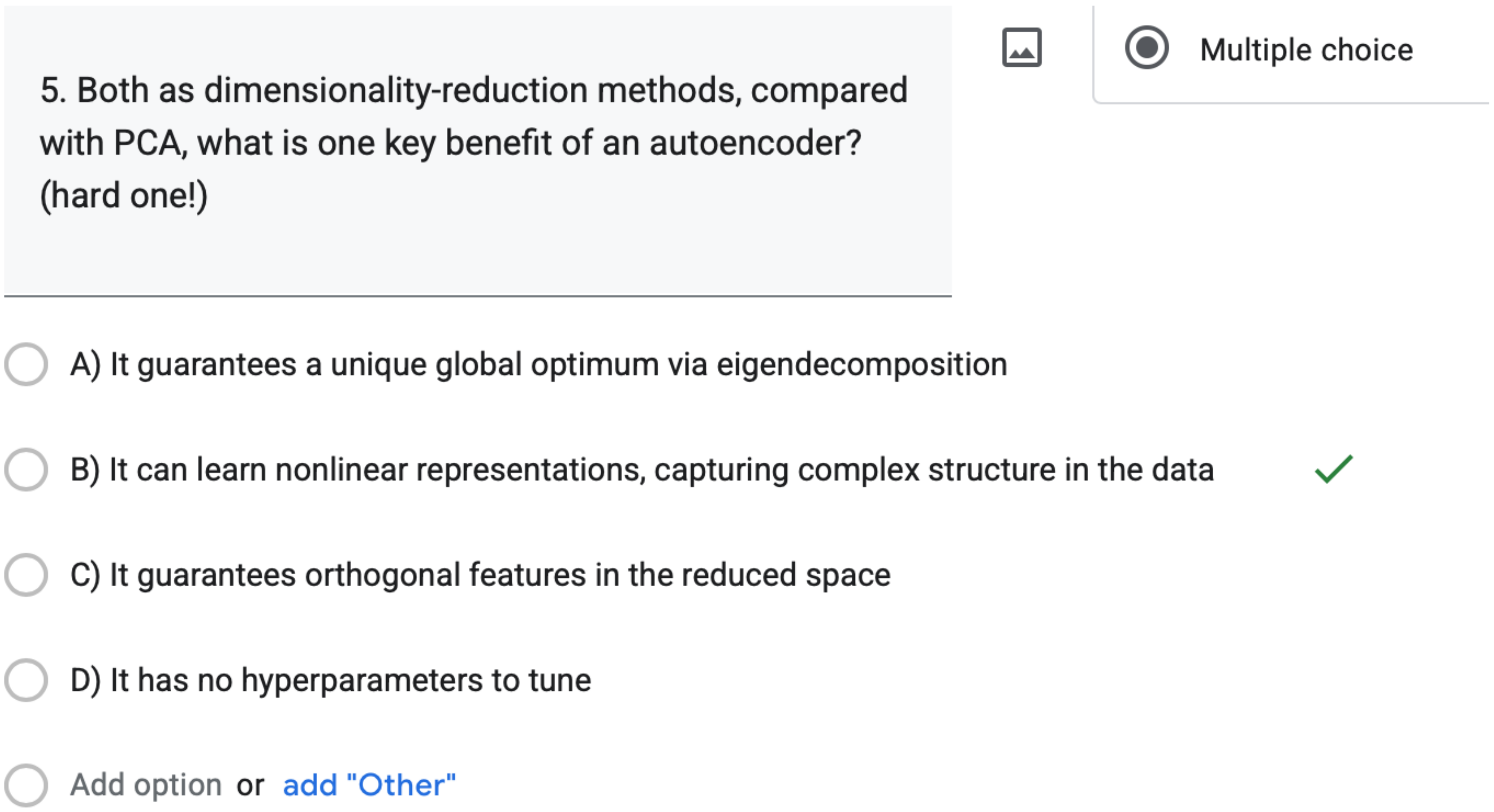
□ A) Many meaningful patterns are small and localized within the image ✓

□ B) The same patterns or objects can appear in different positions across the image ✓

□ C) CNNs build from low-level features (like edges) in early layers to high-level featur... ✓

□ D) The identity of an object is typically preserved even if the resolution is reduced (e... ✓

□ Add option or add "Other"



5. Both as dimensionality-reduction methods, compared with PCA, what is one key benefit of an autoencoder? (hard one!)

● Multiple choice

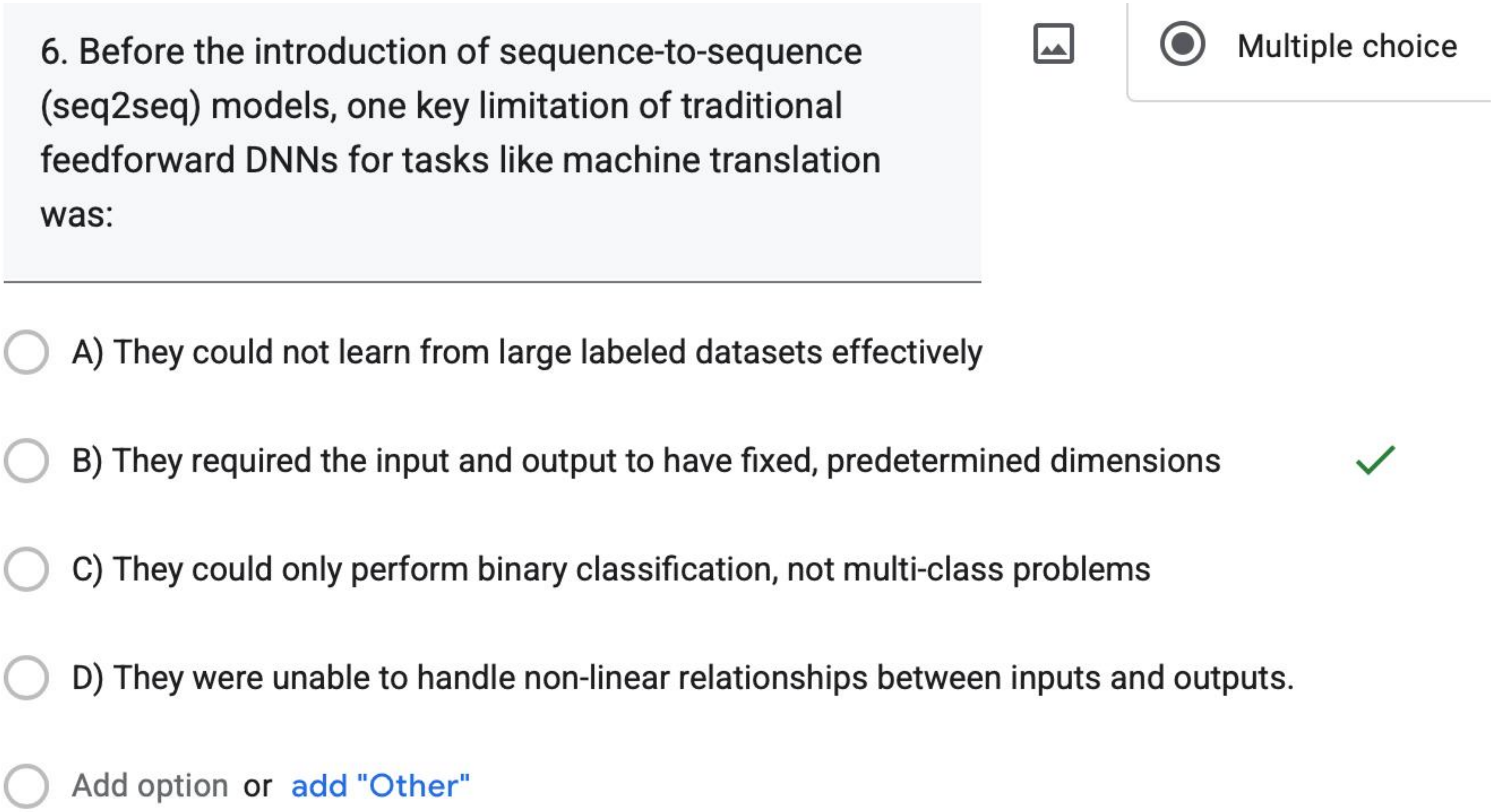
○ A) It guarantees a unique global optimum via eigendecomposition

○ B) It can learn nonlinear representations, capturing complex structure in the data ✓

○ C) It guarantees orthogonal features in the reduced space

○ D) It has no hyperparameters to tune

○ Add option or add "Other"



6. Before the introduction of sequence-to-sequence (seq2seq) models, one key limitation of traditional feedforward DNNs for tasks like machine translation was:

● Multiple choice

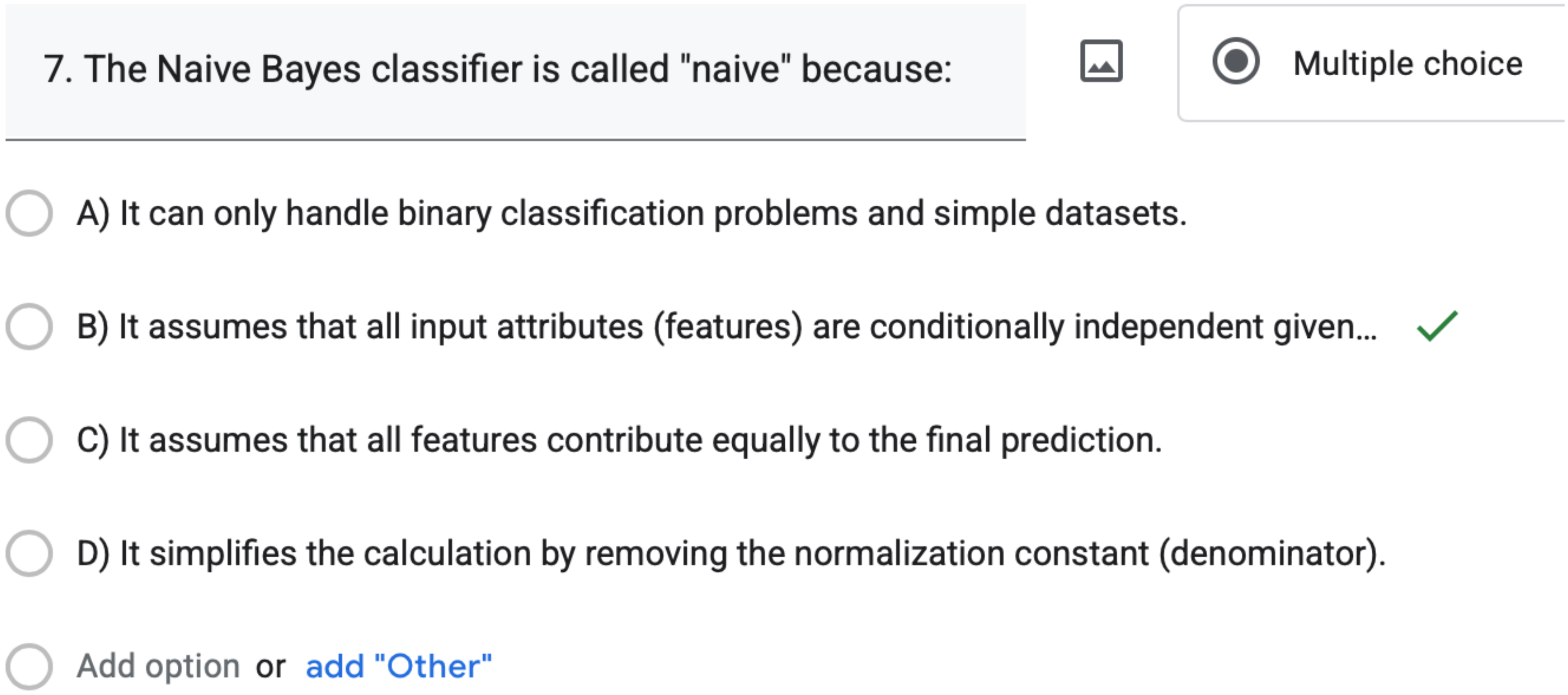
○ A) They could not learn from large labeled datasets effectively

○ B) They required the input and output to have fixed, predetermined dimensions ✓

○ C) They could only perform binary classification, not multi-class problems

○ D) They were unable to handle non-linear relationships between inputs and outputs.

○ Add option or add "Other"



7. The Naive Bayes classifier is called "naive" because:

● Multiple choice

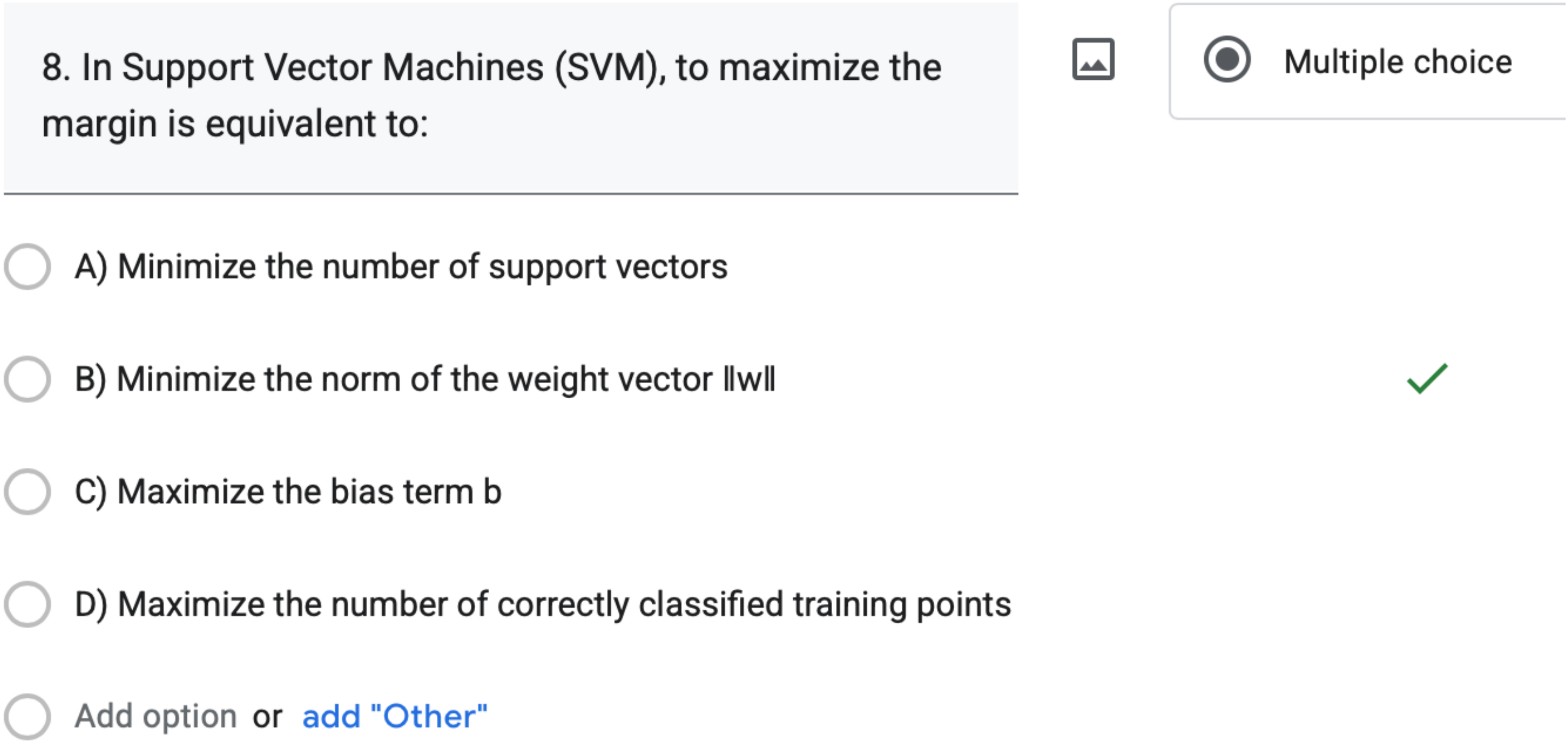
○ A) It can only handle binary classification problems and simple datasets.

○ B) It assumes that all input attributes (features) are conditionally independent given... ✓

○ C) It assumes that all features contribute equally to the final prediction.

○ D) It simplifies the calculation by removing the normalization constant (denominator).

○ Add option or add "Other"



8. In Support Vector Machines (SVM), to maximize the margin is equivalent to:

● Multiple choice

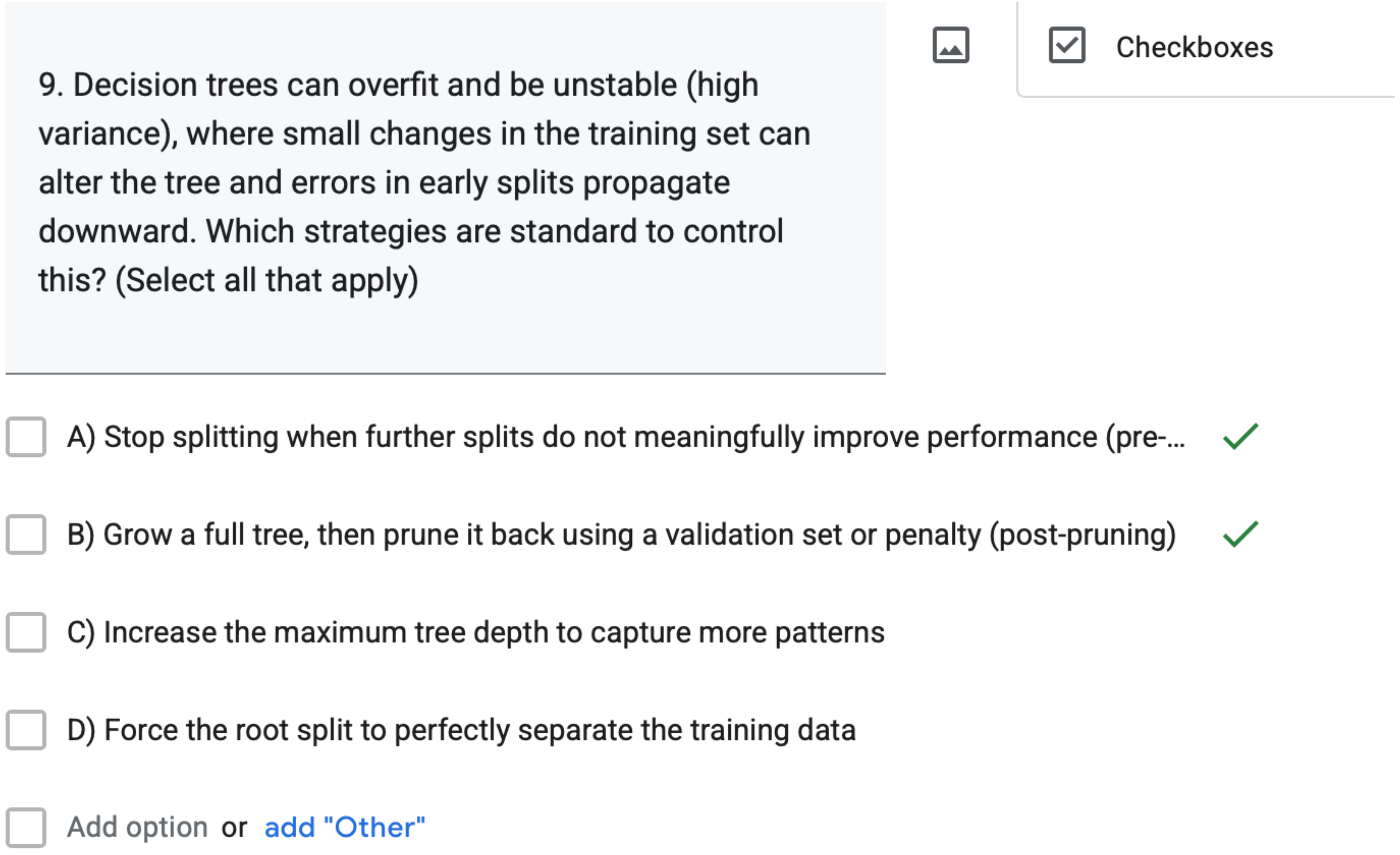
○ A) Minimize the number of support vectors

○ B) Minimize the norm of the weight vector ǁwǁ ✓

○ C) Maximize the bias term b

○ D) Maximize the number of correctly classified training points

○ Add option or add "Other"



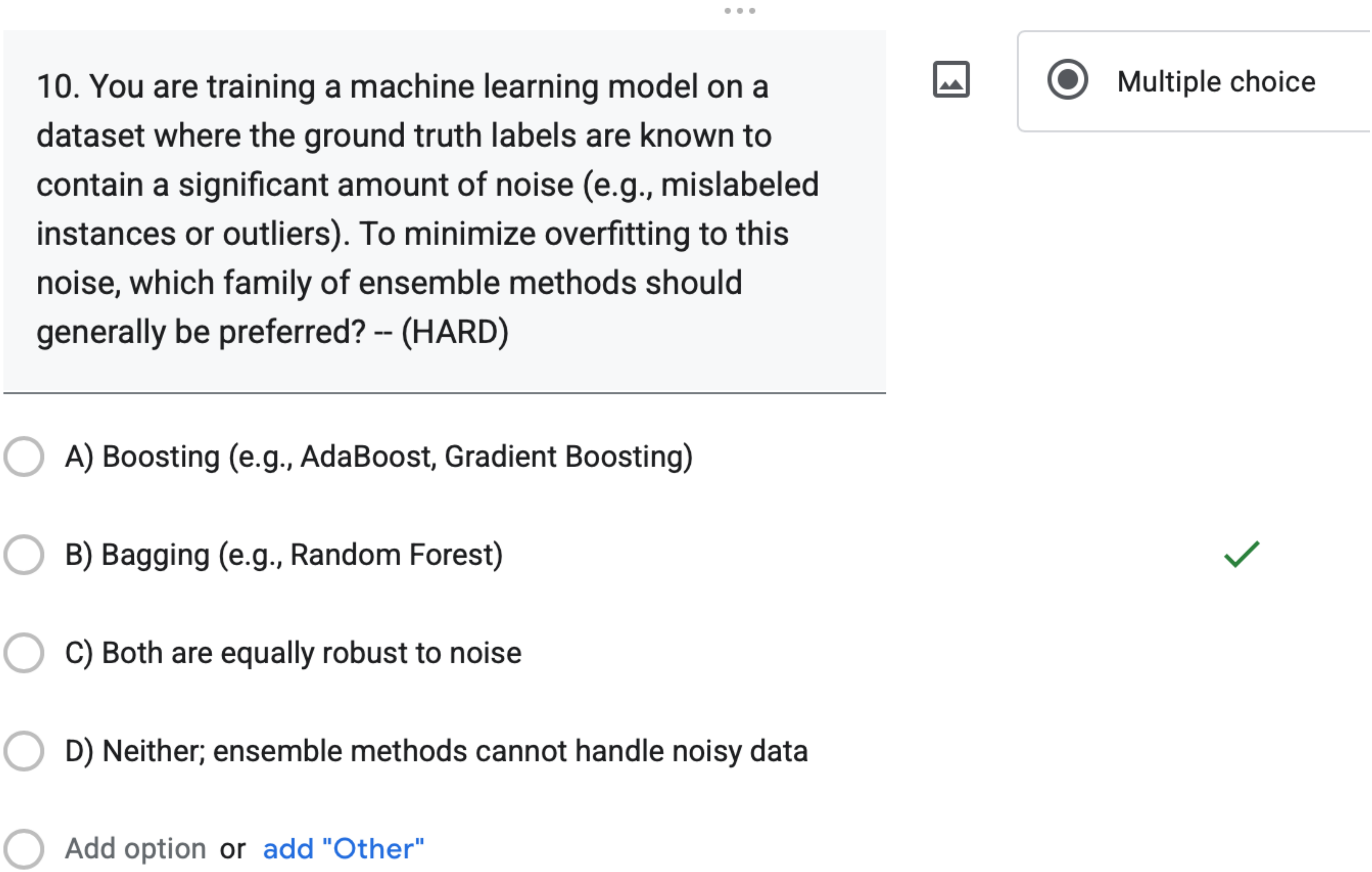
9. Decision trees can overfit and be unstable (high variance), where small changes in the training set can alter the tree and errors in early splits propagate downward. Which strategies are standard to control this? (Select all that apply)

□ A) Stop splitting when further splits do not meaningfully improve performance (pre-... ✓

□ B) Grow a full tree, then prune it back using a validation set or penalty (post-pruning) ✓

□ C) Increase the maximum tree depth to capture more patterns

□ D) Force the root split to perfectly separate the training data



● Multiple choice

10. You are training a machine learning model on a dataset where the ground truth labels are known to contain a significant amount of noise (e.g., mislabeled instances or outliers). To minimize overfitting to this noise, which family of ensemble methods should generally be preferred? -- (HARD)

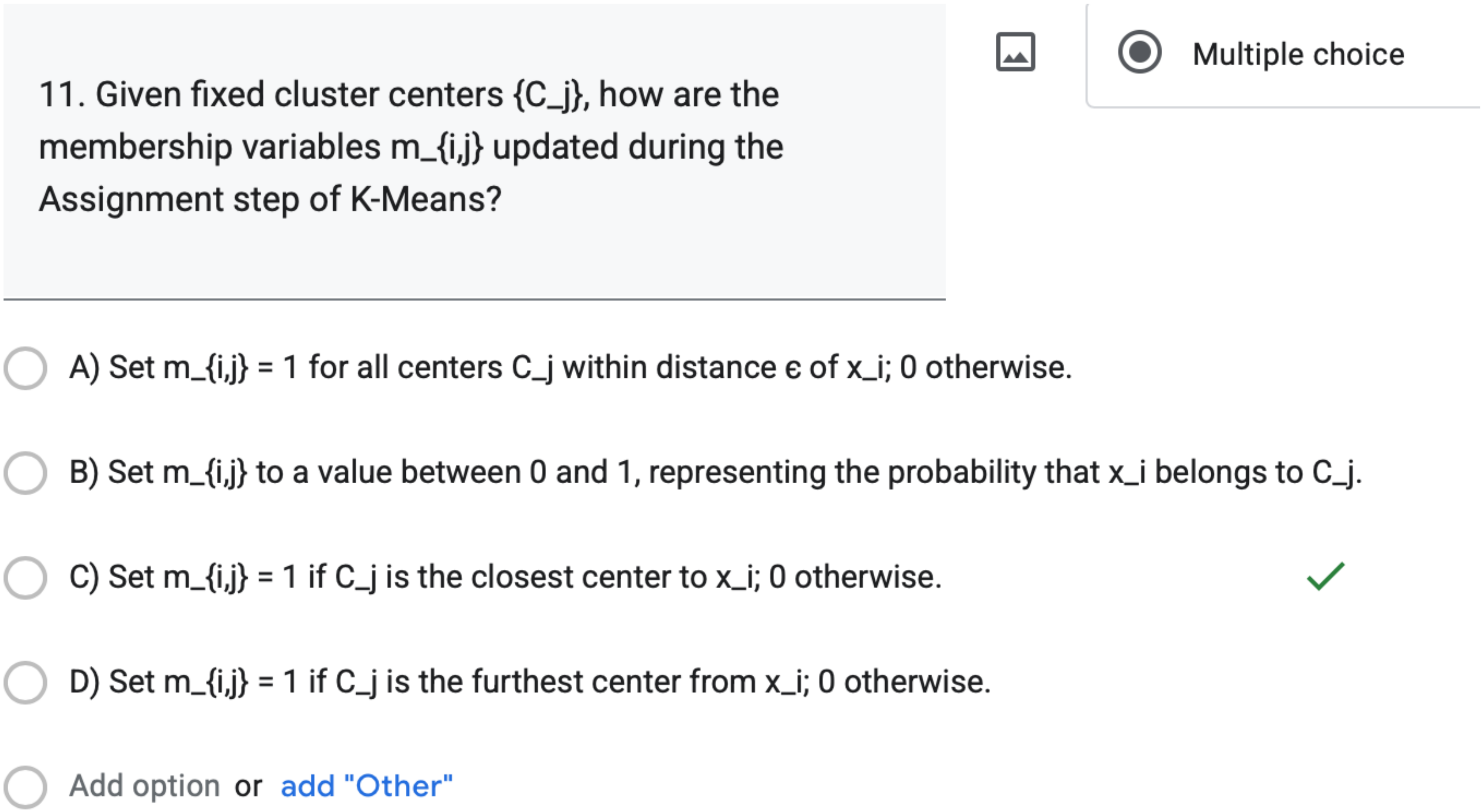
○ A) Boosting (e.g., AdaBoost, Gradient Boosting)

○ B) Bagging (e.g., Random Forest) ✓

○ C) Both are equally robust to noise

○ D) Neither; ensemble methods cannot handle noisy data

○ Add option or add "Other"



11. Given fixed cluster centers {C_j}, how are the membership variables m_{i,j} updated during the Assignment step of K-Means?

● Multiple choice

○ A) Set m_{i,j} = 1 for all centers C_j within distance є of x_i; 0 otherwise.

○ B) Set m_{i,j} to a value between 0 and 1, representing the probability that x_i belongs to C_j.

○ C) Set m_{i,j} = 1 if C_j is the closest center to x_i; 0 otherwise. ✓

○ D) Set m_{i,j} = 1 if C_j is the furthest center from x_i; 0 otherwise.

○ Add option or add "Other"

Thank you!

More in : https://qiyanjun.github.io/2025Fall-UVA-CS-MachineLearningDeep/