### Example assessment questions for model quantization

Question

Why do we use quantization with neural networks? (Select all that apply)

1. It’s more accurate
2. **It saves computational time**
3. **It saves memory**
4. **It’s faster on architectures without floating point units**

Question

Quantization looks like noise, and neural networks are good at handling noise.

1. **True**
2. False

Question

You would use TensorFlow Lite over TensorFlow in order to: (Select all that apply)

1. Enable more Ops
2. **Enable support for quantization**
3. Gain greater accuracy
4. **Maintain a smaller memory footprint**

Question

Quantizing a 32-bit float model into an 8-bit integer model would decrease memory requirements by 16x.

1. True
2. **False**

Explanation: It would decrease memory requirements by 4x.

Question

Quantization works well for neural network weights that have a very wide and flat distribution over a large range of values.

1. True
2. **False**

Explanation: Quantization works best for weights the have a narrow distribution over a small range of values.