Concept Question 2-3: Why does delaying either of two signals delay their convolution?

This can be shown algebraically (property #5 in Table 2-3). A simpler reason: Delaying the input to an LTI system delays the output by the same amount (time-invariance). The output is the convolution of the input and the impulse response, so delaying one signal in a convolution (call it the input) delays the output by the same amount.