# Chapter 5 Review Questions

1. What is the difference between constructs and operational definitions? How do operational definitions allow constructs to be studied scientifically?
2. What is the difference between random error and bias? Which is more serious—and why? What technique might eliminate bias but not random error?
3. What is the difference between subject bias and observer bias? Which bias is mostly likely to affect a study in which participants fill out rating scales?
4. What is the difference between social desirability bias and obeying demand characteristics? How would making responses anonymous affect these biases?
5. What is the difference between reliability and validity?
6. What is the relationship between reliability and validity?
7. What are the basic differences between the different reliability indexes?
8. How do you assess and, if necessary, improve your measure’s reliability?
9. How would you make the case that your measure has construct validity?

# Answers to Chapter 5 Review Questions

1. **Question 1:** What is the difference between constructs and operational definitions?How do operational definitions allow constructs to be studied scientifically?

Constructs, such as intelligence and anxiety, cannot be seen objectively. They are hypothetical characteristics (theoretical concepts) that can only be inferred from observable events. Operational definitions, are recipes that researchers use to objectively, but indirectly, manipulate and measure constructs.

**Question 2:** What is the difference between random error and bias? Which is more serious—and why? What technique might eliminate bias but not random error?

 Bias is much more serious than random error because (a) bias ruins objectivity and (b) random error, although it makes individual scores unreliable, tends to have less of an effect on group averages because random error (a) tends to balance out and (b) can be accounted for by statistics: the science of randomness. Put another way, bias introduces systematic error (often to support the hypothesis that is supposedly being tested) whereas random error introduces unsystematic error.

Some techniques can reduce both random error and bias. For example, you might

train raters to be more consistent and less biased,

standardize procedures to make observers more consistent and more accurate, and

you might limit with observers do by using instruments, multiple-choice tests, or other ways of simplifying the observer’s task to reduce the observer’s opportunities to make any kind of error.

 Other techniques reduce one kind of error but not both types. For example,

Using blind techniques can eliminate observer bias (observers can’t bias the results toward the hypothesis if they don’t know which condition the participant is in) but blind techniques do not eliminate random error due to observers being inconsistent (e.g., by being inattentive, careless, sloppy, or moody).

Using multiple raters, on the other hand, would reduce the effects of random observer error (because random observer errors would have more opportunities to balance out [random errors by one observer might be canceled out by random error by the other observers]) but might not reduce bias (all the raters could be biased to support the hypothesis).

**Question 3:** What is the difference between subject bias and observer bias? Which bias is mostly likely to affect a study in which participants fill out rating scales? How would you reduce subject bias? How would you reduce observer bias?

Observer bias is due to the observer “seeing” what the observer expects or hopes to see; subject bias is due to the participant responding in a way that will help or impress the researcher.

There should be no observer bias in a rating scale study (we should agree about whether the participant marked a “2” or a “4”). There could, however, be substantial subject bias if the participant—rather than responding honestly—responded in a way that the participant thought would either help the researcher “prove” the hypothesis or impress the researcher.

You could reduce subject biases by making participants blind or using placebo treatments so participants did not know whether they had received the treatment, and by hiding the observer, and using unobtrusive measures so participants did not know they were being observed

Making observers blind and standardizing procedures can reduce observer bias.

**Question 4:** What is the difference between social desirability bias and obeying demand characteristics? How would making responses anonymous affect these biases?

Although social desirability bias and obeying demand characteristics are both subject biases, they are different. In social desirability bias, the participant is trying to make a good impression. So, making responses anonymous greatly reduces social desirability bias (because participants know they can’t impress you with their responses if you won’t know that they made that response). In obeying demand characteristics, the participant is trying to give you the response that will help you prove your hypothesis. So, making responses anonymous will not stop participants from obeying demand characteristics (because they can still help you “prove” your hypothesis).

**Question 5:** What is the difference between reliability and validity?

Reliability is the degree to which the measure is not affected by random error, so a reliable measure will produce stable, ***consistent*** scores. For example, if your height was measured reliably, your recorded height would be the same each time you were measured.

The (construct) validity of a measure is the degree to which the measure measures what it claims to measure. So, a valid measure of political beliefs is not valid if it is called an intelligence test.

In short, you can think of reliability as consistency and validity as accuracy.

**Question 6:** What is the relationship between reliability and validity?

A valid measure will be reliable, but a reliable measure will not necessarily be a valid measure. In other words, reliability is a necessary first step toward validity. An example will make this clear.

If a measure is not consistent (i.e., if it is not reliable), it cannot be consistently right (valid). For example, if your height was measured twice today and the measurements were 6 feet and 4 feet, we would know that at least one of the measurements was off and that the average measurement was off by at least a foot.

If a measure is consistent (reliable), it ***could*** be consistently right (valid). If we measure you consistently at 5’6”, and you are 5’6”, we would be right. If, however, you are actually 6’2”, we would be consistently (reliably) wrong.

**Question 7:** What are the basic differences between the different reliability indexes?

Some, such as test-retest reliability, assess the stability of the scores over time. That is, they look at the degree to which the measure is unaffected by all sources of random error.

Interobserver reliability focuses on only one source of random error: random error due to the observer. Obviously, there would be no need for interobserver reliability for a multiple-choice test.

The various measures of internal consistency (split half reliability, Cronbach’s alpha, Kuder-Richardson reliability, inter-item correlations) look at the degree to which participants’ responses to the questions seem to be consistent. If participants are answering the questions inconsistently (e.g., if, on a test of introversion, participants respond to question 1 like an introvert but respond to question 2 like an extrovert), this inconsistency could be due to the participant being inconsistent or it could be due to some problem with the questions (e.g., the questions are poorly worded or the questions don’t all measure the same construct). Obviously, if your measure has only one question, you would not use a measure of internal consistency.

**Question 8:** How do you assess and, if necessary, improve your measure’s reliability?

If you are measuring a construct that is stable over time (adult height or IQ), then participants who take your test today and six months from now should get basically the same score each time. Test-retest reliability can tell us to what extent subjects are getting the same score each time. Typically, you expect a high test-retest reliability coefficient (between .90 and 1.0).

But what if you got a low test-retest reliability coefficient? For example, what if it was below .60? Then, your measure is being affected by something other than the stable construct.

What is this something else? Inconsistent, erratic, **random error!**

What do you do about this problem due to inconsistent, unstable random error?

 1. Replace the measure with a more reliable measure.

 2. See if the random error is due to inconsistencies due to the observer.

Calculating inter observer reliability will tell you if this is a problem. If

inter observer reliability is low, you could use multiple raters or apply any of the remedies suggested in Table 5-1 (p. 136).

 3. See if you can do anything to reduce any inconsistencies in

how the measure is administered. In technical terminology, try to

**standardize** the administration of your measure.

4. If your measure has many items, you could look at the measure’s internal consistency. Throwing out items that seem to disagree with the other items will increase your measure’s internal consistency and may increase your measure’s overall reliability.

**Question 9:** How would you make the case that your measure has construct validity?

You would start by showing that your measure—your operational definition (the recipe, a concrete set of steps or procedures that you will follow to get a score for each subject) is objective. However, you don't just assume that your measure is objective. You try to see if observer bias is a threat to your measure. If it is, you either modify your measure or you try to get objectivity by making your observers "blind." Next, you show that your measure is reliable (see answer to Question 7 for how you would do that). Then, you might try to establish that

Your measure has **content validity:** it has items that measure all the relevant dimensions of your construct and there are enough items for each dimension. (According to experts, it has the right stuff/content).

 Your measure has **internal consistency:** all the items seem to be measuring the same thing (answers to the questions seem to agree with each other). The evidence for this is that participants respond to all the items in a similar way. For example, participants who strongly agree with item 1, should also strongly agree with item 2, and item 3, etc. Note that just because all the items are measuring the same thing does not necessarily mean that they are all measuring the construct you want to measure. To measure internal consistency, researchers may use a variety of indexes, such as mean inter-item correlations, median inter-item correlations, split half reliability, Cohen’s alpha, and Kuder-Richardson reliability.

Your measure has **convergent validity:** the measure correlates with other indicators of the construct. For example, people who score high on your measure should also score higher on other measures of the construct, high scorers should do more of the behaviors associated with your construct than low scorers, and people who are known to be high on your construct should score higher on your measure than people known to be low.

Your measure has **discriminant validity.** Thus, people who score high on your IQ measure should **not** also score high on outgoingness, modesty, social desirability, moodiness, etc. If you can show that you're not measuring the wrong thing, it helps build the case that you may be measuring the right thing. Notice that when you were establishing the reliability of your measure, you were showing that it was not being unduly affected by random error; with discriminant validity, you are trying to show that your measure is not being unduly affected by a construct other than the one you want to assess.