Post-lecture Notes and then Questions I.4 – Measuring the Unobservable

Start with the two most important definitions that are related to "measuring the unobservable":

> Operational Definition – *a statement that maps one or more empirical measures onto one or more theoretical constructs*
>
> Construct Validity – *the extent to which the measure provides an exhaustive and selective estimate of the target theoretical construct*

Note that a given measure, such as mean response time or BDI score, does not have construct validity. Instead, a measure only has construct validity with regard to some unobservable theoretical construct, such as average mental processing time or level of depression, respectively. In other words: measures don't have construct validity; operational definitions have construct validity.

Next, keep in mind that construct validity – like all types of validity – is not an all-or-none thing. This is why the definition of construct validity starts with "the extent to which…" instead of "whether or not…."

Now start to "unpack" the details of construct validity. It has two components which correspond to the "whole truth" and the "nothing but the truth" components of swearing the oath before testifying in court. Each of these has its own definition:

> Convergent Validity [whole truth] – *the extent to which the measure is correlated with other measures of the same (or similar) underlying construct(s)*
>
> Discriminant Validity [nothing but the truth] – *the extent to which the measure is **not** correlated with measures of different (and dissimilar) underlying constructs*

The beauty and usefulness of these two components of construct validity is how they move us from a definition that said what we were hoping for (i.e., an exhaustive and selective measure) to a concrete way of assessing whether or not we have succeeded. Convergent validity requires that the measure be correlated with all other measures of the same thing; discriminant validity requires that the measure not be correlated with any measure of a different thing. The specific values that we aim for in psychology is +.70 or better correlations with other measures of the same thing (although we're not as strict about this as we are for reliability), and correlations within .20 of zero for measures of other things (which, sadly, doesn't get checked anywhere near as often as it should).

So now you know what you need to do in order to "validate" a new operational definition: go and collect lots of data that will allow you to calculate values for your new measure of whatever you're interested in, plus values of several old measures of what you're interested in, and at least a few (old) measures of some things that you are not interested in.

With regard to how we got to where we are now, it's really the happy middle-ground between two different approaches. The first approach was very rationalist (or "arm-chair") in nature: here it was said that in order to have high levels of "content validity," the measure should be exclusive and exhaustive. Those are two very important ideas, but, unfortunately, the way that content-validity folks went about validating their measures was to simply argue about it; that's a weak form of rationalism and psychology is an empirical science. The second approach was much more empirical: in order to have "predictive

validity," the measure should correlated with the known consequences of the construct. This time the strength was in the use of objective measure – namely, correlations – while the problem was the use of "known consequences" which all depend on theory (and, therefore, can be argued about and could, therefore, be wrong). Modern construct validity, with its two components of convergent and discriminant validity, takes the best from both of the previous approaches. It agrees that good measures should be exclusive and exhaustive, and it agrees that we should use empirical data (in the form of correlations) to determine whether or not we have succeeded.

When it comes to testing theories, why is it important that an operational definition have convergent validity?

When it comes to testing theories, why is it important that an operational definition have discriminant validity?

If you are in the process of "validating" a new measure of some unobservable construct and you find correlations with other measures of the same thing that are weaker than +.70, what, in general, will you probably have to do to fix your measure?

If you are in the process of "validating" a new measure of some unobservable construct and you find correlations with other measures of different things that are stronger than ±.20, what, in general, will you probably have to do to fix your measure?

Example multiple-choice questions for this material (taken from Spring, 2011, Exam 1):

1. An alternative name for "operational definition" is _____ .
   (A) linking hypothesis
   (B) indexing function
   (C) *both of the above*
   (D) *neither of the above*

2. Which of the following is **not** an operational definition of anxiety?
   (A) How often the subject crosses and uncrossed his or her arms.
   (B) How much less food the person eats at a meal than usual.
   (C) How nervous the subject is.
   (D) How tightly the muscles in the subject's legs are co-contracted.

If your operational doesn't have convergent validity, then it doesn't measure all of the thing that you're interested in. If your theory works by making predictions about the missing part in particular, then your data will not match the predictions, but not because the theory is definitely wrong; you could be getting data that don't match the predictions because, in effect, you don't have the right data.

If your operational definition doesn't have discriminant validity, then it's measuring things that you aren't actually interested in. This could (again) cause you to get data that don't match the predictions of your theory, but, this time, it's because the "extra stuff" that is also included in your measure is causing the difference. As above, you could be getting data that don't match the predictions because, in effect, you don't have the right data.

If you keep getting correlations below +.70 when testing the convergent validity of a new operational definition, then you've probably forgotten to include some aspect of your target theoretical construct. You will need to expand your measure. If it's a questionnaire, then you'll probably need to add more questions. If it's some sort of observational measure, then you'll need to code some more behaviors. Example: a measure of depression that doesn't include some questions about energy level doesn't really cover all aspects of depression.

If you keep getting correlations that are farther than .20 from zero when testing the discriminant validity of a new operational definition, then you've probably included some things that you shouldn't. You will need to contract (or better focus) your measure. If it's a questionnaire, then delete some items. If it's observational, then stop coding something. Example: a measure of anxiety (which is fear of future, negative events) that includes questions about past injuries, etc, may end up classifying people with dangerous jobs as always being anxious, even when they aren't. Danger is separate from anxiety and you should try to keep danger out of your operational definition of anxiety.

The correct answer to the first question is C, because "operational definition," "linking hypothesis," and "indexing function" all mean the same thing. You'll see "operational definition" most often in clinical and social; you'll see "linking hypothesis" in sensation and perception, plus some physiological psychology; you'll see "indexing function" most often in cognitive neuroscience.

The correct answer to the second question is C. While C might sound like a better operational definition of anxiety than one or more of the other options, what makes C an unacceptable operational definition is that it has not converted an unobservable thing into one or more objective and replicable observable things. Nervousness is not directly observable; it must be inferred from something else, which means it is not (yet) in an objective form. The other three options are all in directly observable and, therefore, objective and replicable terms; that makes them acceptable. (Watch out for this kind of thing. What might seem to be a lousy operational definition might actually be the best possible objective and replicable measure. Be especially on the look-out for "fake" operational definitions that simply map one unobservable thing onto another unobservable thing, which gets us nowhere.)