Post-lecture Notes and then Questions I.5 – Threats to Construct Validity

One of the many differences between reliability and construct validity is whether there is an "excuse" for either of these things to be low. In the case of test/retest reliability – the type of reliability that applies to all measures, regardless of whether the measure comes from a raw score, a summary score, or a condensed score – there is no (acceptable) excuse. Reliability really needs to be at least +.70 or the measure should be avoided. (Even stronger: none of the better psych journals will publish research using a measure with a test/retest reliability below +.70.) In contrast, there are some reasonable "excuses" for a convergent validity score to be less than +.70, or for a discriminant validity score to be greater than +.20 or less than –.20.

The "excuse" for a score below +.70 on a test for convergent validity is that the target theoretical construct is *ad hoc*, instead of real, and you failed to activate the construct before attempting to measure it. The "excuse" for a score more than .20 away from zero on a test for discriminant validity is that the target theoretical construct is distributed, instead of unitary, and the second construct to which you are comparing your measure happens to involve at least one of the same basic elements.

Note, however, that the existence of these excuses should not be seen as a get-out-of-jail free card that allows you to ignore any failures of convergent or discriminant validity. In the case of a failure of convergent validity, you need to show that if you alter your measure (and/or add some extra instructions, etc), you can "wake up" the construct and now find a strong correlation between the new and old measures. Similarly, when you have a failure of discriminant validity, you need to show that at least part of your measure for the target theoretical construct is not correlated with at least part of the measure for the other theoretical construct. For example, you can break your measure up into sub-scales and show that only one of the sub-scales is correlated with the measure for the second construct.

If all of this seems a bit complicated, you are not the first to have this reaction. But it is not OK to just throw up your hands and walk away. Remember: the operational definition of the important theoretical construct is what allows us to make testable predictions from theories; these predictions are how we test theories (since we're empirical scientists), so without the operational definitions, psychology grinds to a halt. Even more: if we take short-cuts with regard to construct validity, we could easily end up making the wrong predictions for our theories which could cause us to reject a theory that is correct or retain a theory that is wrong. Therefore, doing all of the background work that is necessary to "validate" a measure is critical. It needs to be done before we move on to testing theories.

With that said, there are at least two other (less popular) approaches to construct validity that are both, in part, reactions against the idea that there should ever be "excuses" for a failure to verify convergent or discriminant validity. The first of these alternatives follows the same story as the standard approach right up until you reach the issues of real vs *ad hoc* and unitary vs distributed. This approach simplifies matters by removing the "excuses" by saying: only real and unitary constructs should ever be studied. Unfortunately, it also makes it impossible to study some things that are of great interest to psychologists.

The other way of avoiding these issues is to start from a completely different place when it comes to developing operational definitions. Instead of starting with a construct that you want to be able to estimate, start with the data and use them to infer the existence of constructs. This is done by asking a large number of questions of a large number of people under a variety of situations. The first two

requirements – large number of items and large number of people – are to ensure that we end up with measures of all of the relevant constructs. The last requirement – a variety of situations – is to ensure that only those constructs that are active in many situations will be tapped. Once you have all this data, you conduct a fancy analysis that boils all of the variables (i.e., the items on the test) down to a relatively small number of "factors" (hence one of names for this: factor analysis). The factors – by the magic of stats – can be forced to be uncorrelated with each other, so that they each correspond to a distinct theoretical construct. Thus, this approach automatically has very high discriminant validity. And, as long as you included many items related to each of the factors, the odds of covering all aspects of each of the constructs is high, which implies pretty good convergent validity, as well.

The factor-analytic approach has shown itself to be useful in at least one sub-areas of psychology: personality. This approach produced what has come to be the most-popular way to measure personality: the Big Five (OCEAN) model, according to which all personalities vary in five ways: openness, conscientiousness, extraversion, agreeableness, and neuroticism.

In case you're now wondering why this last approach isn't taken by everyone, one reason is suggested in the last sentence of the paragraph before the last … the sentence that begins with: "as long as you included many items…." It turns out that the specific items included in the huge data-set can be very important. In some cases, adding or deleting some items can cause the factor analysis to produce a completely different set of theoretical constructs.

What are two main differences between the reliability of a measure and construct validity of a measure?

In general, what are the three main threats to construct validity?

What do you need to do in each of the two phases of validating a new measure of some construct?

Some example multiple-choice questions for this week:

1. The reliability of a measure _____ .
   (A) completely depends on what it is being used for
   (B) 70% depends on what it is being used for
   (C) 49% depends on what it is being used for
   (D) does not depend on what it is being use for

2. The idea that some important constructs are actually collections of overlapping subunits _____ .
   (A) can be used as an "excuse" for low reliability
   (B) can be used as an "excuse" for low convergent validity
   (C) can be used as an "excuse" for low discriminant validity
   (D) can be used as an "excuse" for any problem with any kind of validity

The most important difference is that reliability only concerns the data produced by the measure, while construct validity involves the relationship between the data and the theoretical construct being estimated. The second difference is that there is only one (important) kind of reliability (which must be +.70 or better), while there are two sides to construct validity (one of which must have +.70 or better correlations, the other must have correlations within .20 of zero).

In general, your measure might lack convergent validity (by not covering all of the target construct), it might lack discriminant validity (by including things that are not parts of the target construct), or the use of your measure might trigger some kind of reactivity (and, therefore, not measure what you want it to). All threats to construct validity usually come down to one of these three things.

In the first, convergent-validity phase, you make sure that your measure has +.70 or better correlations with all other (good) measures of the same construct. This might require that you add some more items to your measure, which might force you to use a condensed score. In the second, discriminant-validity phase, you make sure that your measure has correlations with .20 of zero with measures of other constructs. This might require that you refine, focus, or eliminate some items.

The correct answer to the first question is D. What you are using the measure for has nothing to do with reliability.

The correct answer to the second question is C. Distributed constructs that overlap other distributed constructs make it impossible to pass the test for discriminant validity. Construct validity and the details of the constructs have nothing to do with reliability, so the answer it isn't A. They don't stop you from covering all of the construct, so the answer isn't B. And there's no excuse for all problems, only a few excuses for specific problems, so it isn't D.