**Sara Mills May 10, 2010**

**PSYC 557 Final Exam**

*1. Briefly describe the domain sampling model and the concept of parallel tests AND explain their roles in the Classical Test Theory (CTT) formulation of reliability. That is, describe how these concepts relate to CTT’s conceptualization/treatment of true score variance and error variance. One way that might help to think about this is in terms of the role of these concepts in coefficient alpha reliability.*

 In the domain sampling model, a test is conceived as a sample of all the possible information contained within the domain of interest. For example, if you were designing a test of introductory psychometric knowledge, the questions on the test would reflect a random sample of the universe of knowledge of introductory psychometrics (the domain). You could write another test of introductory psychometric knowledge the contained a different random sample of the domain. In fact, you could write an infinite number of tests of introductory psychometric knowledge using an infinite number of random samples of items from the domain. These tests would be reliable in so much as they adequately sampled the domain of introductory psychometric knowledge. Reliability is the proportion of total score variance that is true score variance. Error variance, in this model, is due to the differences in the content of the tests.

 The concept of parallel tests is a specific case of the domain sampling model. In Classical Test Theory, when two tests are considered parallel the following assumptions are made: (1) true scores on parallel tests are equal; and (2) the error variances on the parallel tests are the same (Kaplan, 2010, p. 70). If someone scores 95 on Test A, then, and another person scores 95 on Test B, we can assume that they have the same degree of knowledge about introductory psychometrics. When tests are truly parallel, the correlation between the tests is equal to the reliability of either test.

 However, in the real world we often do not have truly parallel tests to use to compute the reliability of a measure. Instead, under the domain sampling model, each item on the test is treated as a parallel item. Reliability, then, is the average correlation among parallel items.

*2. A colleague of yours was not wise enough to take psychometrics (Loser). As a result, he/she thinks that there is only way to conceptualize and estimate reliability - to compute coefficient alpha. Briefly describe to him/her why there is no “one” estimate of reliability. In doing so, be sure to describe the type/nature of reliability that coefficient alpha (versus other reliability indices) assesses.*

 There are several methods for computing reliability, and the method you choose depends on the construct you are testing and the purpose of the test. Basically, the various methods differ in terms of what is considered true variance and what is considered error variance. For example, you may be interested in a characteristic that is stable over time and you want to make sure your test gives you the same result time after time. A test designed for early detection of HIV is an example of this. Then you would be interested in the “coefficient of stability” (Kaplan, 210, p. 74). That is, you want to show that your test will accurately detect the presence of HIV each and every time it is given. Therefore, you would want to compute test-retest reliability by giving the test to a sample of people, then giving it to the same sample at another time, and correlating the scores. With test-retest reliability, and change in an individual’s scores from testing time 1 to testing time 2 is error.

Other times, you may be interested in computing a “coefficient of equivalence” (Kaplan, 210, p. 74). A coefficient of equivalence measures individuals at one given point in time, as well as group factors, and is tied to the content of the test. Error is due to tests having different items. One way to determine the coefficient of equivalence is to give parallel test and calculate the correlation between them. However, we often do not have two truly parallel tests. Another coefficient of equivalence is split-half reliability. To compute split-half reliability you split the items on the test in half and calculate the correlation between the two halves of the test. This assumes that the two halves of the test are parallel to each other, which is important to consider when you decide how to split the test (e.g., split by odd and even numbered items, first half and second half of the test). A formula called the Spearman-Brown Correction can tell you the reliability of the whole test, rather than half of the test, which is what you get with split-half reliability. Coefficient alpha, which is a well-known measure of reliability, is another coefficient of equivalence. Alpha is essentially the average of all the possible split-half reliabilities for a given measure.

3. *That same ignorant colleague from above has a question for you. After negotiating a hefty sum in exchange for your continued help, you agree to answer the question. He/she wants you to briefly explain the concept of correcting for unreliability in an observed correlation and to describe when it is and is not appropriate to make this correction.*

 Because of measurement error, observed correlations are often not accurate. There may be issues that occur at the time the test is given, or errors that occur because researchers believe that measurement error is not an issue for a particular measure, or errors that occur for a whole host of other reasons (see Schmidt and Hunter, 1996, for numerous example). Because these types of measurement errors inevitably occur, measures of reliability are not accurate and must be corrected. This is a particular concern in psychological testing where researchers are attempting to measure underlying traits through observed scores on tests. Due to error of measurement in the tests, the measure of the underlying trait is not accurate and it is appropriate to make corrections (Schmidt & Hunter, 1996). Smaller samples sizes also distort measurement and making corrections in these situations in appropriate. It is perhaps unnecessary to make corrections when there is a large sample because correlation measures are more precise with larger samples.

*4. Briefly explain how inter-rater reliability and inter-rater agreement differ from each other. Describe a situation in which one would be primarily interested in inter-rater reliability and a situation in which one would be primarily interested in inter-rater agreement.*

 Inter-rater reliability and inter-rater agreement are two different measures, and which is used depends on what you need to know. Inter-rater reliability is similar to the idea of parallel tests – we can compute the correlation between our raters’ scores and this tells us the reliability of the measure. If actual scores are not important, then measures of inter-rater reliability may be appropriate. These measures determine whether judges rank-order the given targets in a consistent way. For example, when interviewing students applying to the PhD program, each interviewer must fill out an evaluation form, which asks the interviewers to evaluate a candidate on 10 different characteristics using a 1-4 scale. The exact scores that candidates get are not important. Rather, their relative standing is what matters the most. In that case, inter-rater reliability would be an appropriate measure.

 If, however, the subject’s exact score is important, measures of inter-rater agreement may be more appropriate. Inter-rater agreement is calculated by finding the proportion of exact agreement between two or more raters. When calculating measures of students’ on-task behavior, for example, inter-rater agreement indicates how precise the reported percentage of on-task behavior is. It could be calculated as the proportion of times the raters agreed that a student was off task, or as a percentage of overall agreement. Inter-rater agreement is also appropriate when different raters are evaluating different subjects.

*5. The University administration has just hired you to develop a new test that it will use (in conjunction with other factors such as SAT scores and high school GPA) to make admission decisions (i.e., to determine which students will get accepted or not). The University would like to select applicants who will be “well-rounded” in college, and therefore they would like this test to predict “non-academic college success.” Your job here is to describe how you would “validate this test” for this purpose. Specifically, you will describe the steps you would take in gathering content-related, construct-related, and criterion-related validity to support the use of this test and the inferences that are made from scores on the test. Be specific in describing what you would do and why you would do it.*

 There are several steps that I would take to gather data about the content-related, construct-related, and criterion-related validity of this test. To begin, I would establish that the content included on the test is a representative sample of the entire domain of “non-academic college success.” To do this, I would ask experts to indicate how well the items on the test cover the domain, whether components of non-academic college success are missing, or whether some items do not fit within the domain. Such experts would likely be individuals within the university who work in student services or other areas that would have first-hand knowledge of what successful college students do. Such an analysis of test content by experts would provide evidence that the test covered the content it was intended to measure.

 Next, I would give the test to currently enrolled freshmen and compare the scores of students who demonstrate non-academic college success and those who do not. If the mean scores of the two groups are different, this would provide some evidence that the measure differentiates between these two groups of students. The reason for selecting freshman as the sample is twofold. First, freshmen are likely to be more similar to incoming students than students who have been in college for a longer period of time. Second, those students who have low non-academic college success scores may be more likely to drop out or transfer to other schools over time and would, therefore, be underrepresented at higher grade levels.

 A third method for establishing the validity of the test would be to run an exploratory factor analysis (EFA). Such an analysis can provide evidence that we are, in fact, measuring the construct that we think we are measuring. It may also indicate that our initial conceptualization of non-academic college success was not accurate. For example, perhaps we thought there were three subscales in our measure, but EFA reveals that we only have two dominant subscales. This information can help us more accurately define our construct. Additionally, the results of the EFA may point to items on the test that do no measure the construct we are interested in. We may then consider dropping those items from the test. We can identify the impact of test redesign on reliability through the exploratory factor analysis.

 Another powerful tool for making the case for construct validity is confirmatory factor analysis (CFA). Through the use of CFA we can evaluate how accurately our test measures non-academic college success. We can also test the hypothesized relationship between both non-academic college success and academic college success to overall success in college. With CFA we would be able to show how strong the relationship between non-academic college success and overall success in college is. We would also be able to show the strength of the relationship between academic success and overall success, as well as how closely related non-academic and academic success are. This understanding would be useful in helping the admissions team decide how to evaluate the types of evidence they have about potential students when making admissions decisions.

References

Kaplan, S. (2010). *Psychometric methods class notes*. George Mason University, Fairfax, VA.

Schmidt, F., & Hunter, J. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199-223.