Sara Mills Mid-term Exam

PSYC 557 March 30, 2009

*1. We hypothesize that there is a strong positive relationship between self-esteem (the predictor) and career success (the outcome). To examine this relationship, we have people complete a self-repot measure of self-esteem and we record their salaries. We find that the correlation his zero. With respect to the diagram below, briefly describe three possible linkages that could have produced this zero correlation. In other words, describe that “could have gone wrong” with respect to these linkages.*

One of the linkages in our model that could be faulty is the link between the construct of self-esteem and the self-esteem measure. That is, the questions on the self-esteem measure may be, in fact, measuring something other than our construct of self-esteem. Likewise, a second linkage that could be faulty is the link between career success and salary. Perhaps salary is not the best measure of career success, or not what people with high self-esteem would equate with career success. A third linkage in our model that could be faulty is the linkage between our construct of self-efficacy and the construct of career success. Perhaps no such relationship exists and our original premise was incorrect.

*2. Provide an example (other than the one I discussed in class) of a relationship between a construct and a measure that supports the following statement, “How we interpret scores at the construct level does not depend on the scale of measurement of the measure. What’s important is the relationship between the construct and the indicator (i.e., measure).”*

One example of the above statement, which says that how we interpret scores does not depend on the scale of measurement but rather the relationship between the construct and the measure, would be when studying the effects of mothers’ blood sugar levels on fetuses. A difference of 5 points in the blood sugar level of mothers in the average blood-sugar range may not have any differential effects on a fetus. However, a 5-point difference in a mother’s blood sugar level at the high end of the range may have a very large impact on the developing fetus. The difference of 5 points, itself, is not what matters. What makes the most difference is where within the range of blood sugar levels those 5 points occur.

*3. We examine the relationship between anger and aggression. We do this in two ways. First, we randomly assign participants to one of three anger conditions (high, medium, low anger) and then measure their aggression. Second, we measure pre-existing individual differences in anger and in aggression.*

1. *Describe one limitation of each approach.*

One limitation of the first approach, the experimental approach, is that it is difficult to generalize the findings of the experiment to other situations because the conditions of the experiment are so precisely controlled. One limitation of the second approach, the correlational approach, is the lack of such control over the context in which people respond, and the limit that sets on being able to determine causality.

1. *Describe what would constitute “systematic variance” and “error variance” in each of these designs.*

With experimental design, systematic variance would be the variance between scores *across* groups while error variance would be the variance across participants *within* one group. With correlational design, systematic variance is the differences in true ability between individuals. Error variance, then, is all other sources of variance that are not related to true ability, such as the variance resulting from the particular measure being used, an individual’s mood on the day of testing, and so on.

*4. You are the testing expert for a trade school. The school provost informs you that she would like you to create a measure of student morale. List and describe the steps that you would follow in describing how you would proceed (up to, but not including, the point of data analysis).*

Step 1: What is the question or phenomenon? The provost has indicated that she would like to study student morale. I would speak with her about what, exactly, it is that she is talking about when she speaks of “student morale,” and how she would like to be able to use the information from the measure.

Step 2: Define the construct. After clarifying the question or phenomenon we are studying, the next step is to define the construct. To do this, I would take into consideration what the provost is trying to measure, and look more deeply into previous literature on this topic. I would also look for pre-existing measures of student morale to see if one could be adapted for the provost’s purposes. A lot of thought must go into defining the construct because it drives the types of items that will be included on the measure.

Step 3: Operationalize the construct. This is the stage at which items are written for the measure. To do this, we must think of all dimensions of the construct of student morale, and make sure our items cover all dimensions. Items should be written following common guidelines for item writing (see Course Notes, p. 29-30) such as clearly wording items, having each item contain only one idea, and so on. It is useful to have experts in the field evaluate the items to ensure that they do cover all of the dimensions of student morale and are appropriate for the construct.

Step 4: Collect and analyze data for pilot testing. Pilot testing is important to give us information about the statistics associated with the measure (e.g., correlations, reliability), and information about the measure from participants (e.g., which items are confusing, appropriateness of items). Participants in pilot testing should be as close as possible to the target sample.

Step 5: Item revision. Based on the results of the pilot testing, the measure will be revised. These revisions can include deleting poor items, rewording or revising confusing items, or selecting items to give the measure the desired properties. For example, in this case we may want to select items that will give us a measure that distributes scores across a normal curve. Another option might be to select items that discriminate between individuals with a certain level or student morale from those that do not.

Step 6: Administer the measure. After the measure has been revised following pilot testing, it is ready to be administered to the intended sample.

*5. We collect data from a sample of 4 individuals on their self-reported health and happiness. The data are below.*

|  |  |  |
| --- | --- | --- |
| *Person* | *Health* | *Happiness* |
| *1* | *3* | *4* |
| *2* | *2* | *5* |
| *3* | *3* | *2* |
| *4* | *1* | *4* |

1. *Show how you would compute the variance of each variable AND the covariance between the two variables by inserting the appropriate values into the appropriate formulae.*
2. *Solve the formulae from part a. You can use a computer program like Excel to compute these statistics using the values from part a if you would like. Also, be sure to label the sums of squares and sums of cross products in the formulae.*
3. *Assume that this is a two-item test (i.e., each of the two variables is an item on the test). Compute the coefficient alpha reliability for the test by inserting the appropriate values into the appropriate formulae.*
4. *Discuss two ways to increase the reliability of this test by making reference to the formula for coefficient alpha.*

One way to increase reliability is to maximize the covariance of items while minimizing item variances. This would make the numerator larger and the denominator smaller in the coefficient alpha formula. Another way to increase the reliability of the test would be to increase the number of items on the test. This also increases the numerator in the coefficient alpha formula.

*6. Explain why missing data is more of a problem with formative measures than with reflective measures.*

Reflective measures assume that there is an underlying construct, and we can measure indicators that reflect that construct. For example, the construct of effective teaching may be made up of behavior management skills, subject-matter knowledge, and pedagogical content knowledge. We would expect that differences in teacher effectiveness will show up when we measure those indicators. If some data is missing, we still have other pieces of information to help us measure the underlying construct.

When using formative measures, the construct is formed from the indicators. For example, a construct such as standard of living is not one thing in and of itself. Rather, it is a composite of indicators, such as income and home ownership rate, that makes up the standard of living construct. If one of these indicators is missing, you are no longer measuring standard of living because standard of living is, by definition, that particular combination of indicators.

*7. Describe two consequences of having a test in which all of the items are nearly perfectly positively correlated with each other.*

If all of the items on a test are nearly perfectly positively correlated with each other, that means that most of the p values of the items on the test are equal to (or nearly equal to) .5. One of the consequences of this maximum covariance between items is that the test will have high reliability. However, another consequence of this maximum covariance is that the test will not correlate well with other measures of the same construct.

*8. You are the psychometrician for Fletcher school district. The superintendent, Mr. Underhill, just heard about this great procedure called computer adaptive testing (CAT). He is thinking that maybe the school district should start using CAT for its advanced placement (AP) test. He is confused though about exactly how CAT works, and he does not want to implement CAT without understanding it.*

*Your job is to write him a memo describing CAT. Specifically, he would like you to (1) explain how CAT (and IRT) “works.” “How is it possible that we could compare scores when different students respond to different items and a different number of items?” he wonders. (2) He also would like you to outline the steps you would follow in developing a CAP for the AP test. Be sure to write the letter at the appropriate level of complexity. Given that he is a school administrator, he has a basic understanding of item analysis. However, he is not a testing expert.*

Computer adaptive testing (CAT) is based on item response theory. Typically, when we give tests, we calculated means and standard deviations to help us understand how students performed. We transform our students’ scores into z-scores so we can compare their performance across tests. All of these measures are calculated based on the students in our testing sample. If we test other students, our numbers will change.

With item response theory, on the other hand, we can measure students’ ability independent of the sample of students taking the test. Each item on the test has its own item characteristics, and we can use these to predict someone’s ability without having everyone take the same items or the same number of items. We can do this because, through the test development process, we find out the mathematical model for the probability at which individuals along the ability continuum will get the item correct, and we find out the standard error of measurement along this ability continuum. We would have this information for each possible item on the test.

Computer adaptive testing works like this: We will have large pool of possible items at different difficulty levels. First, a student will receive a question that about 50% of students will get correct (based on the item characteristics we determined during test development). If the student gets the item correct, he or she will get an item at a slightly higher difficulty level. If the student gets the item incorrect, he or she will get a slightly easier item. Therefore, each item given to the student will be determined by whether or not he or she got the previous item correct. In this way, we can hone in on the students’ true ability level. Each student will get his or her own sequence of questions from the test pool based on how he or she answers prior questions. The testing ends when additional items administered do not significantly change the standard error.

How would we go about developing such a test? The first step will be to develop a large pool of test items and administer them to a large pool of people. A good rule of thumb is to have more than 80 items and more than 2000 participants for test development. With data from those test-takers, we can calculate the properties of each item. We will then select items to include on the CAT based on these properties. We will select items at a range of difficulty levels that can provide us information across the range of ability levels. If we want our test to spread students out on a normal curve, then we will select items that provide us information about all of the space under that curve.

References

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