EPSY 5221: Principles of Educational & Psychological Measurement Validity Exercises

Consider each scenario below regarding validity coefficients (correlations).

1. Will the validity coefficient (a) increase, (b) remain the same, or (c) decrease?

2. Justify your thinking.

A predictive criterion-related validity correlation is reported as .54 over a three-month period. The time is extended to six months.

*Decrease. Time introduces more substantial changes to the characteristics being measured as well as additional measurement error. Both of these effects result in lower correlations.*

The standard deviation of predictor scores is 5.4. We repeat the study with a more variable sample (same sample size), and the standard deviation increases to 7.4.

*Increase. All else equal, more variability increases correlations, as they are dependent on variability in both measures.*

The sample size in the validity study is doubled (twice as many participants).

*Remains the same. There is no association between sample size and correlations or variance. Sample size has a direct effect on the precision of a correlation (statistical significance), but not the magnitude of the correlation.*

Random scores are included in the predictor test.

*Decrease. Random (due to low motivation, careless responding, unintended errors, etc.) scores do not correlate with anything. So more randomness decreases correlations.*

Instead of 25 items in the criterion measure (the one we are trying to predict), 50 items of equal quality are used.

*Increase. More items is analogous to more observations and greater information – also leading to greater variability.*

Each of the following research scenarios has a problem. Identify the problem and a remedy.

1. A high school math teacher makes up a test for identifying mathematically gifted high school students and gives the test to 100 students in the school. On the basis of high test scores, students are permitted to enroll in a calculus course at the local junior college. At the end of the semester the test developer correlates test scores with grades in the college course. The teacher finds no significant correlation.

*Because the teacher used a test to select students and selected the highest scoring students, no correlation could be a result of regression to the mean – some of the students scored high on the test due to measurement error, so the scores were inflated. Another reason would be if the students get the same grade in the college course – leading to no correlation. Finally, if the students that were selected all had the same test score, again there would be no correlation with grades – both of the last two reasons are due to range-restriction.*

*To do this validation study well, you would randomly select high school students and allow them to take the college course, then correlate pre-test scores with grades.*

2. An industrial psychologist is developing a test to screen applicants for clerical jobs in a given corporation and asks job supervisors to rate present workers on a scale with categories: above average, average, and below average. The psychologist then tests workers in the above average category and those in the below average category and compares their group means. The psychologist finds a significant difference in favor of the above average group.

*The problem here is that the test is intended to be a screening test. If the psychologist compares above- v. below-average employees (those that have already been selected) and the test differentiates between these two groups, it suggests that the test can be used to distinguish between above- and below-average employees. It does not provide evidence that the test is useful for selection – that it can distinguish between employees who are successful and those who are not likely to be successful.*

3. A developmental psychologist devises a test for social maturity in children. The test is administered to 50 students at a preschool, and parents and teachers are notified of their scores. Teachers are requested to observe the children carefully for the next three months and to keep anecdotal records on each child’s behavior that would be useful in assessing level of social maturity. On the basis of these records, the psychologist ranks the children on the level of social maturity and rank orders them on the basis of their test scores. These ranks are correlated by the Spearman rank order correlation coefficient, and it is quite high.

*Since teachers are aware of the test scores, their interactions with children could be influenced by their perceptions of child maturity because of the test score. This interferes with the nature of the records kept by the teachers.*

*To do this well, the teachers (and parents) should not know the maturity scores of the child during the period of the study (teachers should be blind to the measure).*

4. A group intelligence test was validated by correlating it with the individually administered Stanford Binet scale. Later an educational psychologist develops a shortened form of the group test consisting of vocabulary words only. Scores on the long form and the short form of the group test are correlated and found to be high and positive. The test developer now proposes using this shortened form of the group test to replace the Stanford Binet, which requires much more time to administer for each individual.

*The connection between the shortened form of the group test and the Stanford Binet is indirect – since these scores are never correlated. Because they have a common test to which they are related (the group test), does not mean that the part that is correlated with the common test is the part that is common between the short form and the Stanford Binet. These scores should also be correlated to assess the magnitude of the relation.*

5. A professor at the University of Minnesota wants study retention of information among college students. The primary question is: Do students benefit more from using flash cards than traditional class notes. After seeing an ad in the Daily about the study, 150 students signed up to participate; they were randomly divided into two groups. One group of students received a set of class notes with the information on it to study; the other group received flash cards containing the same information. Following a two-hour study period, both groups were administered a test by the researchers. The score on the test was used as a measure of retention rate.

*There are a number of limitations in the study design, including recruitment solely through the MN Daily (limited reach), the requirement of students self-selecting into participation, and the use of course content for which students were unlikely to be familiar. The inference is about retention rate, but the experiment was unassociated with students’ coursework and their study habits. Retention is also facilitated by being motivated to learn the content and having a reason (or multiple reasons) to study. Retention was defined over a two-hour period and this is less likely to be useful in practice. We care about retention over longer periods.*