Introduction to Political Science Comprehensive Exam Assessment

Jamie DeLeeuw, Ph.D.

Monroe County Community College

7/5/12

Segments of this document include material from the following sources (either copied directly or modified):

1)http://www.washington.edu/oea/services/scanning_scoring/scoring/item_analysis.html

2)http://www.samuelmerritt.edu/files/IT/ParSystem/2011%20ParScore%20v7.0%20Training%20Workbook%2020110801.pdf

Introduction to Political Science Comprehensive Exam Assessment Results

In order for a test to be *valid*, as defined as measuring what it is intended to, it has to be reliable. While there are different types of reliability, here *reliability* refers to test questions tending to "pull together"; students who answer a given question correctly are more likely to also answer other questions correctly. Low reliability means the questions tend to be unrelated to each other in terms of who answers them correctly; the resulting test scores reflect peculiarities of the items or the testing situation rather than knowledge of the subject matter. Reliability coefficients theoretically range in value from zero (no reliability) to 1.00 (perfect reliability). The Kuder-Richardson Reliability coefficient (KR_{20}) reflects three characteristics of the test:

- The intercorrelations among the items: The greater the relative number of positive relationships, and the stronger those relationships are, the greater the reliability.
- The length of the test: Tests with more items tend to have higher reliability, all other factors being equal.
- The content of the test: Generally, the more diverse the subject matter tested and the testing techniques used, the lower the reliability.

High reliability is essential in situations where a single test score is used to make major decisions, such as professional licensure examinations. Because classroom examinations are typically combined with other scores to determine grades, the standards for a single test need not be as stringent. The following general guidelines can be used to interpret reliability coefficients for classroom exams:

Reliability (KR_{20})	Interpretation
.90 and above	Excellent reliability; at the level of the best standardized tests.
.80 - .90	Very good for a classroom test.
.70 - .80	Good for a classroom test; in the range of most. There are probably a few items that could be improved.
.60 - .70	The test needs to be supplemented by other measures (e.g., more tests, papers) to determine grades. There are some items that could be improved.
.50 - .60	Suggests need for revision, unless it is quite short (ten or fewer items). The test definitely needs to be supplemented by other measures.

.50 or below	Questionable reliability; this test should not contribute heavily to the course grade as it needs revision.

The *item discrimination index* assesses how well an item discriminates between students who performed well overall on the exam – those who knew the material – and those who didn't. It takes the bottom 27% of test scorers, calculates the percentage who answered the item correctly, and subtracts this figure from the percentage of the top 27% of test takers who answered correctly. If 90% of the highest scorers and 65% of the lowest scorers got a particular item correct, the discrimination index would be 25. Typically if the discrimination is below 25% the item should be revised, unless it is a question that one expects virtually all students to know. For this exam assessment, the *point biserial correlation coefficient* (PBCC) is primarily used as the item analysis indicator, as it takes into account all test takers, as opposed to only the top 27% and bottom 27% of scorers. A higher coefficient indicates that students who responded correctly to the item are the ones who performed well on the test as a whole, and those who answered incorrectly are the students with lower exam scores.

PBCC	Interpretation
.30+	Very good item
.20 -.29	Reasonably good item (subject to improvement)
.09 -.19	Marginal item (needs improvement)
< .09	Poor item (reject or improve)

Method

The reliability analysis for the Introduction to Political Science (POLSC-151) cumulative final exam was comprised of four traditional and four online sections, totaling 119 students. Sixty-three percent of the sections were from Winter 2012 and the remainder were from Spring 2012. While there are a greater number of traditional sections offered to students relative to online sections, an equal number of online sections were included given the online sections' smaller class sizes. Due to having to manually transfer Winter 2012 scantron responses to new forms to complete the analysis, these sections were randomly sampled whereas the Spring 2012 exams that were directed to me were included in the analysis.

Results

The exam consisted of 50 multiple-choice questions and the mean and median test scores were 75.5% and 76.6%, respectively. Overall the exam was very reliable, KR_{20} Reliability Coefficient = .80; students tended to perform consistently – whether good, average, or poor – across the exam, indicating that test scores truly reflect knowledge of the material. On question 1, 99% of students identified the correct answer. By discrimination index standards, typically this question would be considered overly easy and indiscriminate of the highest and lowest test scorers; however given that it is psychologically beneficial to testers to begin with a few easy questions, it is well-placed. Question 9 has a very high correct response rate and could be moved to an earlier position for the aforementioned reason. In terms of item improvement, the PBCC indicates that #50 is the worst item on the test (-.02); correct responses have no relationship with overall test performance. Items 15, 21, 41, and 45 have coefficients between .08 and .15 and could use improvement or be entirely replaced. Questions 46 and 48 are also technically in the "needs improvement" category at .19.

The high reliability of the exam set the groundwork for establishing validity. If exam questions are linked to a variety of course objectives, as opposed to coming from only a few chapters or objectives, *content validity* is established. When creating the cumulative exam the political science department aligned each exam question with one or more of the 17 course objectives; each objective was represented at least once, and on average, four times. While some objectives are broader and more intricate than others, requiring more question representation, *Objective 2: Describe the political ideology spectrum* could be better represented, as content was tapped by only one question. *Objective 9: Describe the role of individual participation in the political process* occurred the most frequently (11 times), but was typically one of multiple objectives identified per question.

Exam validity could also be improved by converting several existing questions into applied questions; in turn the KR_{20} would predictably decrease slightly since descriptive questions are different from applied questions. Adding a short answer section comprised of applied scenarios is another alternative, and I believe this is already being considered by the department. It is advisable to replace exams over time due to validity issues related to test questions becoming accessible to students. Once an exam is considered sufficiently reliable and valid, it can serve as a standard of comparison (criterion) for future tests. For instance, a sample of students could take the “old” test (or perhaps a subset of questions) as well as the new version, to determine whether students scored similarly on both exams. If scores are similar, the new version is considered to have *criterion validity*.