Psychometric Methods Doctoral Comprehensive Exam

May 2012

Instructions:

Congratulations on qualifying to sit for the Psychometric Methods, Educational Statistics, and Research Methods Doctoral Comprehensive Exam. This exam consists of three parts over a three-day period. This is Part I of the exam, which focuses on Research Design and Statistics. There are six questions in this part of the Exam. Some questions have multiple parts. Please answer all questions. All responses should be typed or neatly written by hand. You may use the printer in room 149A to print out anything you type for this exam. Please let the proctor know if you need paper, pencils, pens, or other material.

PLEASE BE SURE YOUR NAME IS ON ALL PIECES OF PAPER THAT YOU SUBMIT.

Day 1 involves five total hours of testing time. Part 1A (Research Methods & Statistics I) is from 9:30 am – 12:00 pm. Part 1B (Statistics II – Multivariate II) is from 1:00 pm to 3:30 pm. Work at your own pace, but be sure to complete all the questions within the five-hour period. Please note the number of points each question is worth is listed for each question. All questions over the 3-day period add to 100 points. Part I is worth a total of 35 points.

We expect all work to be your own, and that you will not look at others' work, or ask questions of anyone except the faculty proctor. Copying someone else's answers or other forms of cheating will result in failing the exam. When you hand in the exam, you will be asked to sign a statement confirming that the work you are turning in is your own.

The content of this exam closely follows what you have learned in class and so the questions should align well with your knowledge and experience.

Good luck!

Topic/Day 1: Research Designs and Statistics (35 Points)

Morning Session

1A: Research Methods

Item 1 (7 points)

A university research group is interested in investigating the relationship between two constructs in elderly populations, "Mental Agility" and "Physical Agility." The researchers decide to conduct a study with 20 residents at a nearby community center. The age range of the 20 residents is 72-79 years and four of the residents are male.

To investigate the relationship between the two constructs of interest, the researchers administer an IQ-test to the residents and also as them about their exercise habits. Specifically, they ask the subjects how they would rate their exercise activity on a five-point scale from 1 = "not very active" to 5 = "highly active."

The researchers correlate the scores from the IQ test with the activity ratings and find a correlation of 0.31, which is statistically significant at alpha = 0.10. They claim that their findings prove that physical activity causes the elderly to have a high degree of mental agility.

As an independent researcher you were asked to comment on the validity of the conclusions drawn by the primary researchers.

(a) Identify at least four threats to the internal validity in this study and explain their underlying reasons.
(b) Describe an alternative research design that would be more appropriate to investigate the link of mental agility and physical activity in the elderly. Provide a thorough justification for the structure of this alternative design.

Item 2 (4 points)

SelfControl Laboratories has developed medication they believe will reduce the frequency of outbursts among students with Turrets Syndrome. They have identified a pool of Turrets patients for a test of their medication. They have a placebo pill they can use in the experiment if they wish (that is, a similar-looking pill that is known to have no medical effects). They are aware that different participants may have different baseline frequencies of outbursts, and that male patients may experience them more frequently than female patients.

Describe possible experiments using the following designs they could use and note particular strengths and weaknesses of each.

(a) Two groups, posttest only
(b) One group, pretest and posttest
(c) Two groups, pretest and posttest
(d) Two groups, stratified by sex, pretest and posttest

1A: Statistics I

Item 3 (6 points)

1. Discuss the relationships among population, sample, and sampling distributions.
2. Given a positively skewed population distribution μ= 35, and σ^2= 100, describe the sampling distribution of the mean based on samples of size 40, derived from this population in terms of its shape, location, and spread.
3. Draw the population and sampling distributions described in (2).

Afternoon Session

1B: Analysis of Variance, Regression, Multivariate Analysis

Item 4 (7 points)

A school principal was interested in providing supportive feedback for the teachers in her school. She wanted to know if they would prefer daily supervision, where she would observe their classroom every day for 5 minutes, or monthly supervision, where she would observe their classroom for an hour (continuous) once per month. There were 18 teachers in her school. She randomly assigned them to one of two observation conditions: daily or monthly. For the nine teachers in the daily group, she sat in on their classes for 5 minutes a day for the entire semester. For the nine teachers in the monthly group, she sat in on their classes for one continuous hour per month for the entire semester. At the end of the semester, she met with each teacher and provided feedback on his/her teaching. Following the interview, each teacher completed an anonymous survey that measured satisfaction with the principal's feedback. A higher score on the survey indicated higher satisfaction. The data for each group were:

Daily Group	Monthly Group
4	9
12	6
10	11
20	4
16	7
18	6
22	15
11	4
19	5

Using your knowledge of statistics and hypothesis testing, test whether there is a statistical difference between the two evaluation methods with respect to teacher satisfaction with feedback. In describing your conclusions, be sure to show all work, including your choice of alpha, critical values, and effect size. Give the principal some advice based on the results of your analysis of these data.

Item 5 (5 points)

A team of researchers conducted a study and analyzed the data using a linear regression model. They submit the results to a journal for publication and a reviewer notices that the results do not seem to replicate what has been shown in the literature, and questions the authors about the validity of their findings. In particular, the authors gave no information regarding checking the assumptions of linear regression, and the reviewer requests evidence that the assumptions were met.

a. What are the assumptions of linear regression?
b. What are the consequences of violating the assumptions?
c. How do you test the assumptions of linear regression?
d. What can you do if you violate the assumptions?

Item 6 (6 points)

Consider the following multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

a. Write this equation in the form of $\underline{y} = X\underline{\beta} + \underline{e}$
b. What are the common assumptions regarding the distribution of errors in regression?
c. Why do we need to make the assumptions in part b?
d. If we make the assumptions in part b, what is the distribution of y? [Include shape and relevant parameters.]
e. If you recall that the estimate of $\underline{\beta}$ is $\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$, what is the distribution of b?

Psychometrics Methods Doctoral Comprehensive Exam: Part II

Instructions: Part II (Day 2)

This is Part II of the Psychometrics Doctoral Comprehensive Exam. This section focuses on Measurement Theory and Applications. There are **four** questions in this part of the Exam. All questions have multiple parts. Please answer all questions. All responses should be typed or neatly written by hand. You may use the printer in room 149A to print out anything you type for this exam. Please let the proctor know if you need paper, pencils, pens, or other material.

PLEASE BE SURE YOUR NAME IS ON ALL PIECES OF PAPER THAT YOU SUBMIT.

Day 2 involves five total hours of testing time. Part 2A (Classical Test & Item Response Theory) is from 9:30 am – 12:00 pm. Part 2B (Equating & Validity) is from 1:00 pm to 3:30 pm. You have five hours to complete Part II. Work at your own pace, but be sure to complete all four questions within the five-hour period. Please note the number of points each question is worth is listed for each question. All questions over the 3-day period add to 100 points. Part II is worth a total of 40 points.

We expect all work to be your own, and that you will not look at others' work, or ask questions of anyone except the faculty proctor. Copying someone else's answers or other forms of cheating will result in failing the exam. When you hand in the exam, you will be asked to sign a statement confirming that the work you are turning in is your own.

The content of this exam closely follows what you have learned in class and so the questions should align well with your knowledge and experience.

Good luck!

Topic 2: Measurement Theory (40 Points Total)

Morning Session

2A: Classical Test & Item Response Theory

Item 1 (10 points)

In 2012, it is very clear that graduates of a program in psychometric methods who aspire to be successful, require knowledge of both CTT and IRT models and their applications. In this question we would like you to focus only on CTT models and methods. These methods are methods that are used all over the world, including test agencies such as ETS and ACT .

Identify at least 10 equations introduced in the two-semester CTT course that today are part of basic testing practices. For each equation,

(a) write it down if you can remember it,
(b) list any assumptions that are associated with the equation, and
(c) identify at least one situation where each equation is used or you think it might be used.

If you can't remember the particulars of the equations, that is fine, but explain as best as you can what each equation is in words, and how it can be used, offer any assumptions you can remember about the equation, and explain how the equation is or could be used in applied testing settings.

Item 2 (12 points)

Help your boss make the case for using item response theory models. He just returned from a meeting where he heard a very muddled presentation about the strengths and weaknesses of CTT and IRT. Here are questions he needs to answer to straighten everyone out at the next meeting:

a. What are the shortcomings, if any, of classical test theory?
b. What is item response theory?
c. Can particular IRT models handle both dichotomous as well as polytomous response data? Explain/justify your answer.
d. What sample sizes and test lengths are required to carry out IRT analyses?
e. What are some of the popular IRT models that are currently being used by the states and national testing agencies? (Identify at least four models.)
f. What assumptions are common to these models and how can these assumptions be checked with real data?
g. What is the difference between building tests with classical models versus IRT models?
h. How can DIF be studied with IRT models?
i. What are the merits, if any, of simulation research?
j. What software should we be using to do the IRT analyses?

Afternoon Session

2B: Validity and Score Equating

Item 3 (8 points)

Compare and contrast how Messick (1989) and the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) define validity, and describe the advice they provide for test validation. Your answer does not need to include exact quotes, but your answer should describe (a) how Messick defined validity, (b) how the *Standards* defined validity, and (c) the frameworks they proposed for validating the use of a test.

The expected length of your response is approximately four to eight single-spaced pages.

Item 4 (10 points)

Think about the task of linking two tests together, both more or less measure the same construct and are approximately of the same difficulty. Assume that a non-equivalent anchor design (NEAT) is the only option for linking the tests. How would you respond to the following questions?

a. Under what conditions might the NEAT design be the only equating data collection option?
b. What are considerations in choosing an anchor set for this design?
c. What might be the advantages or disadvantages of using an internal versus an external linking design?
d. More or less, how does the Stocking-Lord equating method work? (You don't need to offer equations but describe in words.)
e. In many texts, the Stocking-Lord method is recommended among the equating methods based on IRT. Any idea why this might be true?
f. When looking at a b-plot in applying the NEAT design, how might outliers be identified?
g. If asked, what would you say are the differences between horizontal and vertical equating?
h. Do tests need to be of the same length and of the same score points to be linked? Please explain.

Psychometric Methods Doctoral Comprehensive Exam: Part III (25 points)

Instructions: Part III (Day3)

You have applied for a job with a testing company and the company has a novel way to select from among their four finalists. Their plan is to provide each finalist with a set of data and request various classical and modern analyses. The candidate who provides the best set of solutions will be hired. (This happened with one of our UMass students this year and I am pleased to report that she was hired.) **Take 24 hours if you want prior to submitting your work**.

You will be given a dataset with the following characteristics: 40 MCQ and five CR items (scored 0 to 4) to make up a 60 point test. There are some missing data in the candidate responses.

a. The data are not complete in the sense that students omitted answers to some of the questions. For this exercise, convert the data so that missing responses are treated as zero scores. But what do you think are the consequences of this decision? How much missing data will be treated as zero scores in the data? Does this worry you? Do it anyway.

b. Carry out a classical item analysis (just look at p and r values) and see if any of the items appear flawed. If you run into convergence problems simply delete the problematic item or items. Please don't invest time in trying to save these problematic items.

c. Do at least one analysis to address the test dimensionality question—eigenvalue plots would be one approach, a SEM analysis could be another.

d. Using PARSCALE because the test consists of both multiple choice items and constructed response items, fit the 3p model to the 0-1 items and the GR model to the polytomous response items and consider model fit. It would be best to use Tie's Resid Plots software (but any software you have will be acceptable) to get the analysis completed correctly—you could look at residuals, standardized residuals, distribution of standardized residuals, prediction of score distributions, and even use some of the chi-square statistics that the software produces. What did you find? Do you think the 3p/GR models would provide a good fit?

e. Using only the binary scored items in the test, what does the test information function look like? If I told you that the cutscores were going to be set at -1.0, 0.5, and 2.5, how do you think the test would function in sorting candidates into these four performance categories?

f. Suppose you were told to add another 9 items to the original test of 60 points for the purpose of linking. Would you use both MCQ and CR items? If so, how many of each would you use and why? Where would you place these linking items in the test and why?

g. How did the item parameter estimation analysis go? Did you have any problems?

When you have finished Part III of the exam, please print all of your work and give it to the proctor, or simply slide your work under Ron Hambleton's door. Good luck!