

Psychometric Methods Doctoral Comprehensive Exam

May 2013

Instructions:

This exam consists of two parts over a two-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. Remember that the exam is closed-book. No books, notes, handouts, etc. can be brought with you to days 1 and 2 of the exam.

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. 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)

Research Design

Microsoft developed an educational video game designed to help students learn algebraic concepts. They would like to hire you to determine the effectiveness of this game for improving students' learning about algebra. In addition to paying you for your services, they have arranged for 1,000 students from 50 different classrooms across the U.S. (20 students per classroom) to participate in the study.

- (a) Design a study to evaluate the utility of Sony's game for helping students learn algebra. What type of research design is your study?
- (b) Explain how your study protects against the major threats to internal validity (at least 3 threats) discussed by Campbell and Stanley (1963).
- (c) What are the limitations of your study?

The recommended length for this item is approximately three to five single-spaced pages.

Item 2 (4 points)

A university mathematics instructor developed and incorporated an extensive computer-assisted instructional (CAI) program into her course on introductory calculus. Wanting to assess the effectiveness of her approach, she decides she will randomly assign students to her and another instructor's section of the course and compare the mean scores on the final exam at the end of the semester (the posttest-only control group design). She hypothesizes a significant difference between the means in favor of his approach.

What rival explanations represent a threat to this design? And how can she avoid them?

The recommended length for this item is approximately two to three single-spaced pages.

1A: Statistics I

Item 3 (6 points)

The animal learning course in the psychology department requires each student to train a rat to perform certain behaviors. The student's grade is partially determined by the rat's performance. The instructor for this course noticed some students are very comfortable working with the rats and seem to be very successful training their rats. The instructor suspects these students may have previous experience with pets that gives them an advantage in the class. The instructor decides to conduct a study and gives the entire class a questionnaire at the beginning of the course. One question determines whether each student currently has a pet of any type at home. Based on the responses to this question, the instructor divides the class into two groups and compares the rats' learning scores for the two groups. The data are as follows:

Rats' Scores for Students with Pets	Rats' Scores for Students without Pets
n = 30	n = 30
$\bar{x} = 78$	$\bar{x} = 66$
$s^* = 15.5$	$S^* = 15.8$

* Assume equal variances

- State the null and alternative hypotheses for this study.
- Conduct a statistical test of the null hypothesis ($\alpha = .05$) using either the p -value or critical value approach. State your conclusions.

Afternoon Session

1B: Analysis of Variance, Regression, Multivariate Analysis

Item 4 (7 points)

A researcher collected data comparing four interventions using a between-group design and a quantitative dependent variable. The researcher is interested in conducting all pairwise comparisons using two-sample t -tests and a full alpha (e.g., 0.05) for each comparison. Do you agree with the researcher's method of conducting the multiple comparisons? Please fully explain why you agree or disagree with the analyses. If you disagree, please explain how the analyses should be conducted.

Item 5 (5 points)

Multicollinearity is a concern in many cases of regression where there are multiple independent variables

- a. What is multicollinearity conceptually?
- b. Why is multicollinearity problematic?
- c. How do we assess multicollinearity?
- d. How do we proceed with a regression analysis in the presence of multicollinearity?

Item 6 (6 points)

Linear Independence and Invertibility of a Matrix.

- a. What does it mean for a set of random variables to be linearly independent?
- b. What information does linear independence provide?
- c. How do you assess linear dependence?
- d. How can the idea of linear dependence be applied to the regression or ANOVA context?
- e. What does it mean for a matrix to have an inverse?
- f. How is linear independence related to the invertibility of the matrix?

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 Principles of Assessment, Measurement Theory and Applications. There are six 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. Remember that the exam is closed-book. No books, notes, handouts, etc. can be brought with you to days 1 and 2 of the exam.

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, 2B, and 2C (Principles of Assessment, and Classical Test Theory & Item Response Theory) is from 9:30 am – 12:00 pm. Part 2D (IRT Test Development & 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 questions within the five-hour period. Please note the number of points each question is worth is listed for each question. 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: Principles of Assessment, Measurement Theory and Applications (40 Points Total)

Morning Session

2A: Principles of Assessment

Item 1 (5 points)

One of the most confusing aspects of the assessment field concerns the particular reporting scales that are used. The problems begin with the use of two fundamentally different paradigms: norm-referenced testing and criterion-referenced testing. The problems continue with what many may consider to be a confusing array of reporting scales for both norm-referenced and criterion-referenced tests. Here are some practical questions to answer:

- (a) What are the main similarities and differences between criterion-referenced testing and norm-referenced testing?
- (b) What are the implications for the reliability and validity assessment of these two kinds of tests?
- (c) Since two types of testing (NRT and CRT) introduce some confusion, do we need both?
- (d) Distinguish among age norms, grade equivalent scores, percentile scores, normalized z-scores, and stanines.
- (e) What do you think is the most problematic aspect of criterion-referenced testing? Please explain your answer.

2B: Classical Test Theory

Item 2 (7 points)

Think back to your study of classical test theory (CTT), and recall the many important models and results that have guided test development and evaluation practices for about 100 years. You may not remember most of the details of the equations introduced over two semesters and that's fine (you can look them up when they are needed during your career), but beginning with the classical test model, recall as many of the results from CTT that you can that impact the practice of good measurement. For example, the basic model has forced measurement specialists to think a lot about measurement errors and how they might be minimized. You learned that an unbiased estimate of the true score mean is the test score mean. You learned about the basic statistic for reflecting measurement error, called the standard error of measurement. See if you can come up with at least 15 more results (i.e., formulas) and explain how they have impacted on educational measurement practices (explain the formula briefly and offer one sentence per formula).

The recommended length for this item is approximately two to three single-spaced pages.

2C: Item Response Theory

Item 3 (8 points)

Please read the scenario below and answer the five questions that follow.

During a job interview for a psychometric position at a state department of education an interviewer states, “We want a new person in the state department of education who can help us with our applications of classical and modern test theory (i.e., IRT) to our test development and related topics—equating, identification of item level bias, and so on.”

The interviewer then asks the following questions:

(a) Which way do you lean regarding these two modeling approaches? Do you prefer one to the other, or maybe a bit of both? Please explain why.

(b) Which of the many applications do you think IRT is best for (e.g., equating, test development, CAT, reporting, DIF detection), and why do you feel as you do?

(c) IRT methodology is still being developed—what do you think are problems that will need to be resolved to help us in the department?

(d) Who, if anyone, is working on these problems?”

(e) Tomorrow I need to go to our board and explain IRT in only 10 minutes to some policy folks. My own status will be helped if I can actually assist these policy-makers to understand IRT. They hear and read about it when discussions about NAEP, MCAS, SAT, etc. come up. In bullet form, can you give me 10 points that I should emphasize in my remarks? If you want to sketch out some graphics to use with your 10 points, they would be helpful too.

The recommended length for this item is approximately three single-spaced pages.

Afternoon Session

2D: Validity and Test Development

Item 4 (8 points)

The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) describe five sources of validity evidence that can be used to support the use of a test for a particular purpose.

- (a) Describe “validity evidence based on test content.”
- (b) Give at least three examples of how such evidence can be gathered, analyzed, and reported to support the use of a test for a particular purpose.

The recommended length for this item is approximately two to four single-spaced pages.

Item 5 (6 points)

You are being interviewed for a job with one of the smaller testing agencies, and they would like to hire someone who can handle their modern (i.e., IRT) test development process. Be sure to mention in your answers about test information, item information, test efficiency, and item banks. They have three questions:

- (a) What IRT model should they be using?
- (b) What are the main measurement concepts that arise with IRT test development? and finally,
- (c) How might test development differ if they decide to focus on a single passing score? How might their approach to item selection change?

The recommended length for this item is approximately two to three single-spaced pages.

Item 6 (6 points)

With the latest educational reform in the US, computer-based testing has become a central concept in assessment. Here are some basic questions that need to be answered:

- (a) What are the differences between computer-based testing and paper-based testing?
- (b) What are the differences between computer-based testing and computer-adaptive testing?
- (c) What are the advantages and disadvantages of IRT with computer-adaptive test designs?
- (d) With CAT designs, what are possible starting and stopping rules for assessment?
- (e) What do you see as problems to overcome in implementing a successful computer-adaptive test design?

The recommended length for this item is approximately two to three single-spaced pages.

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 two years ago 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: 30 to 40 MCQ (scored 0 to 1) to make up the total 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 BILOG-MG (and $D=1.7$) because the test consists of only MCQs, fit the 1p and 3p models to the 0-1 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 model would provide a good fit? How do the 1p results compare?
- e. Using 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. 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!