Research and Evaluation Methods Doctoral Comprehensive Exam

Written Portion

August 2011

Instructions:

Congratulations on qualifying to sit for the Research and Evaluation 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 6 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. If you choose, you may write your responses on the 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 HAND IN.

You have 5 hours to complete Part I. There is a suggested time for each item to help you pace yourself, but you do not have to adhere to these time limits. Work at your own pace, but be sure to complete all 6 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.

Part I is an open book exam. Feel free to use any textbooks, notes, or other class resources in responding to questions, but remember to pace yourself! Internet searches are an exception, and we ask that you do not access the Internet while taking this exam.

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!

REMP Doctoral Comprehensive Exam: August 24, 2011

Part 1: Research Design and Statistics (35 points, 4 hours)

Research Design

1. The CollegeReady Test Preparation Company advertises that the average score gain in SAT verbal scores from students who take their course is 80 points. Let us assume that they are truthful: For all the students who have taken the CollegeReady prep course after the first time they took the SAT, the average score after the course is indeed 80 points higher. There are thousands of students, so we can disregard sampling errors.
   (a) Is 80 points a good estimate of the effect of the course on scores? Why or why not? Use concepts of internal and external validity in your explanation.
   (b) What do you think is the best way to estimate the effect of CollegeReady's course on SAT scores? Design an experiment to get a better estimate. Discuss which of the common threats to validity your design protects against relative to the way CollegeReady computed their estimate.
   (c) Do you think the result from your experiment would yield an estimate of the effect that is smaller, about the same, or larger than CollegeReady's estimate? Why?

   [5 points, estimated time is 30-45 minutes.]

2. There is an often cited relation between children's reading ability and how much their parents read to them.
   (a) Is this a correlational or causal relation?
   (b) How would you design a study to test the correlation between these two variables?
   (c) How would you design a study to test the causal relation between these two studies?

   [4 points, estimated time is 30 minutes.]

3. Describe the sampling distribution of sample mean differences. What is the mean of this distribution? What is the standard deviation of this distribution? How is this distribution used to test hypotheses about group differences?

[5 points, estimated time is 30 minutes.]

4. The National Academy of Sciences (NAS) is interested in determining the effects of different methods of teaching science on science achievement. Based on an extensive review of research in this area, they identified four acceptable approaches: 1) classroom lecture, 2) classroom lecture coupled with informal hands-on tasks, 3) classroom lecture coupled with formal laboratory exercises, and 4) a combination of all three (classroom lecture, informal hands-on tasks, and formal laboratory exercises).

To determine which approach was best, they recruited 40 students who were roughly equivalent based on science achievement and randomly assigned each to one of the four methods. After a semester of instruction an achievement test was administered to all students. NAS has hired YOU to analyze the data from this study and tell them which method is best. The future of our nation's science curriculum is in your hands. The data appear in Table 1. Note that there are 10 different students in each teaching method (column).

Table 1
Science Achievement Test Scores for Students Taught with Different Methods

Classroom Lecture	Lecture and Hands-on Tasks	Lecture and Formal Lab.	Lecture, Hands-on, and Lab.
68	41	71	85
69	81	75	75
70	75	69	80
65	70	70	83
64	55	68	82
50	60	75	80
63	71	81	85
45	68	75	71
60	59	78	87
65	64	70	83

Prepare a brief research report for NAS. In your report:
(a) describe the experimental design,
(b) state and test any relevant hypotheses, and
(c) using an appropriate multiple comparison procedure determine whether: 1) adding hands-on tasks to "lecture only" affects performance, 2) adding hands-on tasks to "lecture + lab" affects performance, and 3) approaches including a laboratory component are different from those not including a lab component. Your report should also include a discussion of effect sizes.

NOTE: YOU MUST SHOW ALL WORK (TYPED OR HAND-WRITTEN), BUT YOU MAY USE SPSS OR EXCEL TO CHECK YOUR WORK OR ASSIST WITH CALCULATIONS. DO NOT ATTACH OUTPUT AS YOU MUST SHOW ALL YOUR WORK.
[7 points, estimated time is 45-60 minutes.]

5. Regression is one of the most common statistical methods used in educational and psychological research. However, the application and interpretation of a regression analysis is complex.

(a) What are the assumptions of linear regression?
(b) For what purposes is it necessary for these assumptions to hold?
(c) How do you test the assumptions of linear regression?
(d) What can you do if you violate the assumptions?

[6 points, estimated time is 45 minutes.]

6. Multivariate Analysis of Variance (MANOVA) is often overlooked as too complicated and multiple ANOVAs are conducted instead. Alternately, multiple ANOVA analyses are used to follow the MANOVA.

(a) Describe, conceptually, what a MANOVA is, and how it works.
(b) Describe when using MANOVA is more appropriate than using multiple ANOVAs.
(c) Describe why following a MANOVA with multiple ANOVAs may not be appropriate.
(d) Describe the appropriate follow-up techniques to a significant MANOVA.

[8 points, estimated time is 45 minutes.]

THIS IS THE END OF PART I OF THE EXAM. BE SURE YOU HAVE HANDED IN RESPONSES TO ALL 6 QUESTIONS.

REMP Doctoral Comprehensive Exam: Part II, August 25, 2011

Instructions: Part II (Day 2)

This is Part II of the REMP Doctoral Comprehensive Exam. This section focuses on Measurement Theory and Applications. There are 6 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. If you choose, you may write your responses on the 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 HAND IN.

You have 5 hours to complete Part II. There is a suggested time for each item to help you pace yourself, but you do not have to adhere to these time limits. Work at your own pace, but be sure to complete all 6 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.

The first 5 questions on the exam are "open book," and you may use any textbooks or other resources that you feel are helpful. Internet searches are an exception, and we ask that you do not access the Internet while taking this exam.

Question 6 is "closed book," which means you will have to put away any resource material when responding to this question. Question 6 will be distributed to you after you have completed the other 5 questions. You may request to complete question 6 before any or all of the other questions as long as you put away all resource material.

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!

Part 2: Measurement Theory and Practices (40 points, 4 hours)

Classical Test Theory

1. Reliability is of fundamental concern in assessment, and for good reason.
   (a) Define what reliability means generally.
   (b) Reliability can be assessed in multiple ways, and the method that is used depends upon the nature of the data and the testing context. Describe the four ways that reliability can be assessed, in which situations should each type of reliability be used, and with each of the four methods, include a description of what reliability means within each situation.
   (c) If a reliability estimate for a test is low, how can the reliability be improved? Are there circumstances where reliability could be low, and might be expected and even acceptable?
   (d) Non-psychometricians often report coefficient alpha as a measure of reliability of an instrument and as long as that value is at least 0.80, they believe they are finished their psychometric work. What is wrong with this approach to thinking about score reliability?
   (e) There is a formula for establishing test score reliability if a test is lengthened or shortened. Upon what important assumption or assumptions is the formula based? There is also a formula for adjusting test score reliability for group homogeneity. Again, under what assumption or assumptions is the formula based?
   (f) Score reliability tends to vary from group to group, especially if score variability is different. But what about the standard error of measurement—does it vary across groups, and so do we need to adjust the standard error of measurement too as a function of test score variability? Please explain.

[9 points, estimated time is 60 minutes.]

2. Some questions about the classical test theory (CTT) model follow. Please answer all questions.

(a) What is the CTT model and what are the assumptions that are made?

(b) How viable do you think the model is, and the assumptions, for analyzing current test data with large scale state assessments?

(c) What are the strengths and limitations of the CTT model?

(d) The CTT test statistics for item discrimination and difficulty are often criticized. Why? Are these criticisms valid? What is the value of these statistics?

(e) Any ideas for why why ETS may have introduced delta values as a substitute for p-values in conducting item analyses and subsequent work with item statistics?

(f) What problems would be encountered if to-day's measurement specialists tried to use CTT to build state tests, identify bias, equate scores, build item banks, etc. [Hint: think of the advantages of IRT over CTT.]

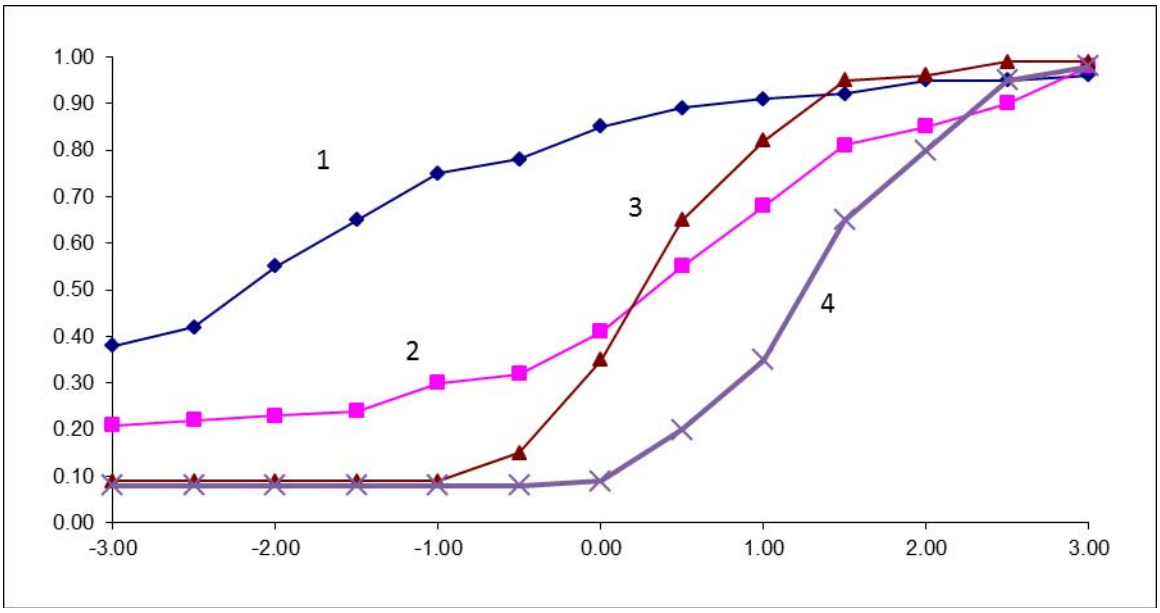
[10 points, estimated time is 60 minutes.]

Item Response Theory

3. The property of item invariance is an important concept within item response theory.
   (a) What is it?
   (b) Why is it so important?
   (c) Assume that the estimates of $a_{i,}$ b_i, and c_i, (corresponding to the 3-parameter logistic model) were each estimated from two very large samples that differed in ability. What relation must hold for the pairs of estimates of $a_{i,}$ b_i, and c_i, if the property of invariance holds?
   (d) How might the assumption of item invariance be checked?

[5 points, estimated time is 30 minutes.]

4. The following plot depicts the item characteristic curves for four dichotomously-scored items. The distribution of student proficiency (denoted THETA) for population of students of interest is N(0,1).



(a) Briefly discuss, for these items, the location where most information is obtained (approximation by eye is sufficient), how well it discriminates between low and high abilities, and its usefulness in a test.

(b) Which item would provide the MOST information at the point where it is most informative? Which item provides the least information at the point where it is most informative?

(c) Suppose you were giving a computerized adaptive test (CAT) and the examinee's current THETA estimate is -2.05. Which of these items would you select to administer next, and why?

[5 points, estimated time is 30 minutes.]

5. What is the difference between classical multidimensional scaling and weighted multidimensional scaling? Please provide the formulae for both models and explain the additional information gained from a weighted MDS analysis.

[4 points, expected time is 20 minutes]

6. Describe the five sources of validity evidence stipulated in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), and discuss how they can be used to develop a validity argument for a test. Provide an example of an actual testing program and the types of validity evidence they could gather to support the use of their test.

   Your response should (a) clearly describe each of the five sources of validity evidence, (b) discuss how they can be integrated to develop a coherent validity argument, and (c) provide one example of an actual testing program and how they could realistically gather and summarize validity evidence to develop a compelling validity argument.

[7 points, estimated time is 60 minutes.]

THIS IS THE END OF PART II OF THE EXAM. BE SURE YOU HAVE HANDED IN RESPONSES TO ALL 6 QUESTIONS.

REMP Doctoral Comprehensive Exam: Part III, August 26, 2011

Instructions: Part III (Day 3)

This is Part III of the REMP Doctoral Comprehensive Exam. This section focuses on Applications of Measurement, Statistics, and Research Methods. There is a single task to complete, but it is comprehensive and asks you to complete several tasks. 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. HOWEVER, YOUR FINAL ANSWERS SHOULD BE E-MAILED TO PROFESSOR HAMBLETON AT rkh@educ.umass.edu. Please let the proctor know if you need paper, pencils, pens, or other material.

PLEASE BE SURE YOUR NAME IS ON ALL MATERIAL YOU HAND IN FOR THIS EXAM. YOUR RESPONSES ARE DUE TO PROFESSOR HAMBLETON **BY NOON TOMORROW (August 27, 2011).**

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!

Part 3: Applications of Measurement, Statistics, and Research Methods

The Massachusetts Department of Elementary and Secondary Education (MDESE) has received a not so friendly letter from a critic of the MCAS. The critic neither likes testing very much nor IRT modeling of the MCAS data. The MDOE wants to go part-way to address some of the critic's concerns and has chosen the 2011 grade 6 English Language Arts Test to illustrate the points that the test is of good technical quality and the IRT model was an inspired choice for data analysis. The MDESE wants your assistance. Their request to you goes something like this: "Fit the three-parameter model to the 0-1 items and the graded response (GR) model to the constructed response items, and then address model fit. We have some other questions too and they are listed below. The current version of the test contains 36 multiple-choice items and 4 (4-point) CR items. We are sending you the responses of several thousand students. Regarding any omits or not reached items, the expectation is that students will receive zero scores. After all, they were encouraged to answer all questions, and were provided with unlimited time to do so.

Here are the questions we have for you to answer:

1. Using the classical item analysis information available from PARSCALE, what is your opinion of the test items? How do the multiple-choice items compare in quality to the constructed response items? How would you recommend the MDESE use the classical item analysis evidence to address the challenge from the critic?

2. Include a print-out of the item parameter and standard error file in your submission. Do you see anything that is troubling in the file or the SEs?

3. For the GR items, show both the global b values and the differences provided by the Parscale printout, and then convert these values for the four items to the item threshold parameters in the GRM developed by Samejima.. We expect this form of the model will be less confusing for the critic.

4. Perhaps if the critic could see that model fit was good or excellent and the test was unidimensional, he/she may feel better. Via any of the acceptable methods, report on the level of model fit. At least run the data through the Resid Plot 2 software (Liang, Han & Hambleton, 2008[1]) and report your findings.

5. If the cut scores on the theta metric are -1.0, 0.50, and 1.5, what would you estimate the cut scores to be on the test score metric? Do the analysis if you can, but if you can't, at least explain how these raw score cut scores could be obtained.

6. Calculate the TIF for the test, and judge the suitability of the current test in relation to the cut scores. Provide the TIF and your opinion about the suitability of the test for the purpose of sorting candidates into proficiency categories. Any suggestions for how we might improve the test next year?

7. Probably the above analyses are sufficient, but can you think of other analyses that the state and the critic might value?

[1] Liang, T., Han, K. T, & Hambleton, R. K. (2008). ResidPlots-2: computer software for IRT graphical residual analyses, Version 2.0 [Computer Software]. Amherst, MA: University of Massachusetts, Center for Educational Assessment. Available at http://www.umass.edu/remp/software/residplots/.