Research and Evaluation Methods Doctoral Comprehensive Exam

Sample Exam Questions

Draft: April 22, 2011

Introduction

The faculty members in the Research and Evaluation Methods Doctoral Program (REMP) at the University of Massachusetts Amherst are pleased to present this compilation of comprehensive exam questions to help REMP students prepare for the written portion of their comprehensive exams. The purpose of the comprehensive exam is to ensure students have sufficient knowledge of the fields of research design, educational statistics, and psychometric methods, to proceed toward candidacy for the Ed.D. degree. Successful completion of the Comprehensive Exam is an important milestone for REMP students because it indicates mastery of the important knowledge and skills taught in our Program and that are important for success in the fields of educational research and psychometrics.

The comprehensive exam consists of a written exam and an oral exam. In this "practice test," we provide sample questions for the written portion of the exam. For further information regarding the entire comprehensive exam process, please refer to the document "Research and Evaluation Methods Program Comprehensive Exam," which describes all policies and procedures related to the exam.

Content and Format of the Written Portion of the Comprehensive Exam

The written examination is designed to assess both the student's knowledge and understanding of research design, measurement theory, and statistical methods, as well as the student's ability to apply this knowledge to solve real-world problems. Students will receive open-response (i.e. essay) examination questions, developed by REMP faculty members, that address three broad topics:

Topic 1: Research Designs and Statistics

Topic 2: Measurement Theory (including Structural Equation Modeling, Scaling, Validity Theory, Item Response Theory, and Classical Test Theory)

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

Topics 1 & 2 will be administered as two four-hour sessions over the course of two consecutive days in a secure room supervised by a committee member. Topic 3 will be administered on the third day and students will have 24-hours to complete the question(s). The number of items per topic may vary depending on the breadth, or depth, of the individual items. Students will receive a common set of questions, but may (at the discretion of the comprehensive exam committee & based on their specific research interests) receive one or more individualized questions.

What follows are examples of the types of questions/items students will encounter on the exam. The examples are stratified by the three Topics.

Day 1: Research Design and Statistics (35 points total) 4 hours

1. A school administrator is interested in whether character-education video games, in conjunction with the traditional weekly school assemblies, improves the behavior of her students. She wants to compare the inappropriate acts of one group of students who are given the opportunity to play the video games to the number of inappropriate acts of students who simply attend the weekly school assemblies, at the end of two months. Design a research study for the administrator, improving on her original ideas for the study, while considering possible threats to internal and external validity. Specifically discuss how you would collect data, the strengths and weaknesses of the research design, and the most appropriate analysis for the collected data.

(Research Design) (Estimated points: 7. Estimated Time: 45 minutes.)

2. Consider the process of conducting a hypothesis test about a population parameter. Using the following table, which shows the cross-classification of your test decision and the true state of the null hypothesis, answer the following questions:

(Statistics I) (Estimated Points: 6. Estimated Time: 30 minutes.)

Truth			
Test Decision		H_o True	H_o False
	Reject H_o		
	Not Reject H_o		

(a) Explain in words the concepts of type-I error rate, type-II error rate, and power in statistically precise terms as they relate to the four cells of the table.
Suppose that a directional z-test is used with alpha = 0.05 to test
Ho: μ equal to 500
Ha: μ greater than 500
Where σ=50 and N = 100.

(b) Sketch the sampling distributions for the sample mean for when μ=500 and for when μ=490 on a common axis.

(c) Assuming that the truth is μ=490, shade in and label the appropriate regions on your figure in part (b) that correspond to the four cells of the table.

(d) Explain (or demonstrate with another graph) how the sketch would change if the sample size was decreased.

3. Some researchers are investigating the effect of socioeconomic status (SES) and home environment on the academic success of students. They design a survey to collect the necessary data, and match it to the statewide testing results of the students. The SES measures include: mother's education, father's education, mother's income, father's income, the number of hours worked by the mother, the number of hours worked by the father, and receipt of free/reduced lunch. The home environment measures include number of books in the home, amount of time spent watching TV during the week by the student, amount of time spent on the internet during the week by the student on tasks not related to homework, the number of TVs in the home, and amount of time the student spends alone, or with a non-parental childcare provider.

(Statistics III) (Estimated Points: 7. Estimated Time: 45 minutes.)

(a) What size sample should the researchers aim to get to conduct this study?
(b) There are many variables in this model. Discuss some of the problems that might arise using this many variables in the model. Be sure to include a discussion of multicollinearity, how it might be measured, and how it can be addressed.
(c) Suppose the researchers are concerned that there are too many variables in the model and would like to reduce the number. How should they select the most important variables?
(d) Suppose the researchers choose to eliminate several of the independent variables. How can they assess the quality of the resulting model relative to the full model?

4. An educational researcher at the University of Massachusetts Amherst was interested in studying the effects of educational television programs on toddlers= mathematical development. She randomly assigned 20 toddlers, between eighteen and twenty months of age, to one of four video conditions: Sesame Street, Bill Nye the Science Guy, Mr. Rogers Neighborhood, or Rug Rats. The first three video conditions represented different forms of educational television programming; the fourth represented entertainment television programming. In each group, the toddlers viewed a one-hour video every day for a week. At the end of the week, a graduate student, who did not know to which group each toddler belonged, measured each toddler=s ability to count to ten. Table 1 presents the number of consecutive digits from one to ten that each child recited.

(Statistics II) (Estimated Points: 5. Estimated Time: 30 minutes.)

Table 1
Data for UMASS Educational Television Study (N=20)

Video Condition	**Number of Correctly Recited Digits for Each Child**
Sesame Street	9, 9, 8, 7, 7
Bill Nye the Science Guy	5, 7, 6, 3, 9
Mr. Roger’s Neighborhood	8, 6, 9, 5, 7
Rug Rats	1, 3, 4, 5, 1

(a) What are the independent and dependent variables for this study?

(b) Are there differences among the four video conditions with respect to the number of correctly recited digits? Provide a summary table of your results.

(c) Is there a difference between educational television and entertainment television?

(d) Form all pairwise contrasts and interpret them. Defend your selection of the specific multiple comparison procedure used. Are there differences among the treatments?

5. Express the linear regression model in matrix notation. Explain what each part of this expression would look like using a concrete example, with at least two independent variables.

(Multivariate Statistics I) (Estimated Points: 5. Estimated Time: 45 minutes.)

(a) How would you estimate the linear regression coefficients using matrix algebra?
(b) How can you express the assumptions of regression in matrix notation?
(c) Find $E(\underline{\beta})$ and $Var(\underline{\beta})$
(d) Using the results of part (c) find $E(\underline{y})$ and $Var(\underline{y})$
(e) Formulate an appropriate hypothesis to test in your concrete example. Write this hypothesis as $C\underline{\beta} = \underline{0}$

(Multivariate Statistics II) (Estimated Points: 5. Estimated Time: 45 minutes.)

6. Principal Components Analysis (PCA) and Exploratory Factor Analysis (EFA) are often used interchangeably, and are often confused for one another.

(a) Discuss the similarities and differences between PCA and EFA
(b) Provide a context when you would use PCA instead of EFA
(c) Provide a context when you would use EFA instead of PCA
(d) In both procedures, different methods can be used to rotate the solution. These methods can produce either orthogonal or oblique rotations. What is the difference between the two? How do you decide which to do?
(e) In either procedure, there is a need to make a judgment about how many components or factors to retain. How does that judgment get made?

Day 2: Measurement Theory (40 Points) 4 hours

1. A common question asked of young psychometricians when they are looking for a job goes something like this, "You must have seen in our job ad that 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 DIF, and so on.

(IRT) (Estimated Time: 60 minutes. Points: 10.)

(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 your reasoning. We are not really sure ourselves as a state department about what we should be setting up in our agency for the next 10 years. A lot of the other states have already switched to using IRT in their technical work.

(b) For which of the many applications do you think IRT is best (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 and please explain your reasoning? (e.g., software, details of specific applications, new knowledge)

(d) Who, if anyone, is working on these problems—who are some of the key persons, and what are they doing?"

(e)And then the conversation goes on, "Oh, and one more question. Tomorrow I need to go to our board and in 10 minutes explain IRT to some policy-folks. My own status will be helped if I can actually help them understand IRT. They hear and read about when discussions about NAEP, MCAS, SAT, etc. come up. In bullet form, can you give me at least 10 points that I should emphasize in my remarks? If you want to sketch out some graphics to use with your 10 points (e.g., what an ICC is or some other basic concepts look like), they would be helpful too. We can find someone tomorrow to put some power point slides together."

2. Think back to your study of classical test theory (CTT), and recall the many important results that have guided practice. You may not remember all the details of the equations and that's fine, but beginning with the classical test model, recall as many of the results from CTT that you can that impact on 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 all important standard error of measurement, and so. See if you can come up with at least 10 results and explain how they have impacted on educational measurement practices. [Spend about one hour on your answer.]

(Classical Test Theory) (Estimated Time: 60 minutes. Points: 10.)

3. One important area of analysis in educational and psychological testing is analysis of the dimensionality of an assessment. Describe three procedures that could be used to evaluate test dimensionality and describe the strengths and limitations of each procedure. If you were to perform a dimensionality analysis on an assessment, which methods would you use, and why?

Your response should list and describe three methods that have been used to evaluate the dimensionality of educational or psychological tests. Each method should be explained and the methods should be compared and contrasted to describe their strengths and limitations. Your response should also include a discussion of how you would evaluate dimensionality in a specific context. The reasons for your choice of method(s) should be clearly described.

(Scaling: Dimensionality) (Expected Time: 45 minutes. Points: 8.)

4. The Massachusetts Department of Education according to many experts has done an excellent job of defining state curricula and building tests (i.e., the MCAS) to match that curricula. Even the score reports now are receiving good reviews. At the same time, the state has been slow to conduct consequential validity studies to answer questions like “Are some of the best teachers leaving the state because we have too much testing in-state?” or “Are our high school graduates now doing better in college or in the work place?” or “Are the new reports understandable and meaningful to users? and there are many more. Suggest a couple of consequential validity studies. Pick one of them that to you is an especially interesting consequential validity question and then describe the problem, goals of the study, and the methodology of that study. In the methodology be sure to focus on the design, variables, data collection methods, and data analyses. Assume you have three years to do the study, and as much money as you need.

(Valdity) (Expected Time: 45 minutes. Points: 8)

Part B: Choose ONE of the following (3) problems. (Estimated Time: 30 minutes. Points: 4)

1. Describe the options for defining a latent variable's scale in SEM. What factors are important in deciding which method to implement? Is the same issue of scale indeterminacy also an issue in item response theory?

(SEM) (Estimated Time: 30 minutes. Points: 4)

2. The standard error of measurement (SEM) is a concept that is closely tied to reliability.

(a) What is the classical test theory formula for the SEM?

(b) What information do we get from the SEM? How should we use it?

(c) What factors influence the SEM? How can SEM be minimized?

(d)What are some of the cited limitations of the SEM in the classical test theory framework?

(CTT) (Estimated Time: 30 minutes. Points: 4.)

3.Define and describe item bias (also known as differential item functioning, or DIF). Describe two methods for detecting item bias and state the relative advantages and disadvantages of each.

(CTT/IRT) (Estimated Time: 30 minutes. Estimated Points: 4.)

Day 3: Applications of Measurement, Statistics, and Research Methods: (24 hours. Points: 25)

Our state department of education is building a new achievement test with lots of new item types. Probably the department's biggest concern is that the new test maybe multidimensional. Their second concern is, assuming the test is reasonably unidimensional, whether the graded response model will actually fit the data. So, here are the details

1. 20 items, each item is scored 0 to 4
2. 1000 persons
3. there is no missing data and all items were responded to by the candidates

The first task is to investigate the test dimensionality. It would be ideal if you could approach the check on dimensionality using two methods to check for convergence. Write up your method, findings, and conclusions.

The second task, regardless of what your findings are from the first task, is to apply the graded response model to the data, and carry out a model fit analysis. We would recommend that you use Parscale and Resid Plots 3 for your analyses but if you are more comfortable with other software, or prefer other software (for example, Multilog) it would be fine for you to use that software to answer the question. Write up your methods, findings, and conclusions.

If you had more time to do the work, can you think of other analyses you might run to help the state make decisions about test dimensionality and model fit.

Suppose the state were concerned about item parameter invariance over ethnic groups. Such a concern would always be justified, but maybe especially problematic with new item types. Describe how you would carry out this analysis and draw some graphics if these will you explain your answer.