2016

EDPS 870: Introduction to Educational and Psychological Measurement—A Peer Review of Teaching Project Benchmark Portfolio

Anthony D. Albano

University of Nebraska-Lincoln, albano@unl.edu

Albano, Anthony D., "EDPS 870: Introduction to Educational and Psychological Measurement—A Peer Review of Teaching Project Benchmark Portfolio" (2016). *UNL Faculty Course Portfolios*. 29.
http://digitalcommons.unl.edu/prtunl/29

EDPS 870: Introduction to Educational and Psychological Measurement

A Peer Review of Teaching Project Course Benchmark Portfolio

Spring 2016

Anthony D. Albano, PhD
Assistant Professor
Educational Psychology
University of Nebraska-Lincoln
albano@unl.edu

Abstract

This portfolio was compiled at the completion of the Peer Review of Teaching Project, professional development available to faculty at the University of Nebraska-Lincoln. The purpose of the project is to support faculty in evaluating and documenting the effectiveness of their teaching. The main objective of this portfolio is to summarize the results of my participation in the project during the 2015/2016 academic year. The portfolio summarizes key features of the course and my teaching methods, along with examples of student work.

Keywords: Measurement, Test development, Psychometrics, Reliability, Validity

Table of Contents

Portfolio Objectives

This portfolio was compiled at the completion of the Peer Review of Teaching Project (PRTP), professional development available to faculty at the University of Nebraska-Lincoln (UNL). The purpose of the project is to support faculty in evaluating and documenting the effectiveness of their teaching. A series of meetings, workshops, and self and peer evaluations provide opportunities to reflect on methods of course design, instruction, and assessment, and procedures for analyzing results so as to highlight student learning, and, as a result, effective teaching.

The main objective of this portfolio is to summarize the results of my participation in PRTP during the 2015/2016 academic year. First, I describe the course that I targeted in this project, including the content, the context in which it is taught, and background on enrollment and student demographics and affiliations. Next, I summarize my teaching methods, materials, and activities, with descriptions of how they have evolved over the past four years. I then situate the course within the broader curriculum of the Educational Psychology department at UNL, and the field of educational and psychological measurement in general. I describe examples of student work that highlight the growth that students experience in the course. Finally, I outline planned changes for my next offering of the course, and then conclude the portfolio with a description of what I personally gained from my participation in the project. Appendices contain my most recent syllabus for the course and examples of student work.

Feedback from peers on this portfolio would be appreciated. Below are some considerations that are of particular interest to me.

1. One of the challenges in a service course such as this is adapting the content to the needs of the learners, especially those without a background in statistics, psychometrics, or quantitative methods. Does the content seem appropriately broad and applied without sacrificing too much of the underlying theory and technicality of measurement theory?
2. What other resources could I use to supplement or replace articles currently in the list of assigned readings?
3. Roughly half of the course covers applied psychometric analysis for test development, scoring, and item analysis. These include item analysis, reliability analysis, generalizability theory, and a small amount of item response theory. Is this amount of coverage appropriate for this audience?
4. Psychometric analyses are taught using the statistical software R. I have found R to be the most flexible and accessible software available for the analyses we conduct. However, some students find it challenging. How can the introduction to R be improved? What additional resources could be provided. How could it be better integrated into instruction?

Course Overview

Description of the course

Introduction to Educational and Psychological Measurement, EDPS 870, is a course in the theory and applications of educational and psychological measurement. Main topics for the course include measurement applications in research and practice, reliability, validity, item writing, test design, and statistical analysis of test data. These topics are reviewed in lecture, assigned readings, and class and group discussions; they are explored within group assignments and a course project; and they are assessed via performance on assignments, the course project, and quizzes given in class. Additional details can be found in the course syllabus, included in **Appendix A**.

Students

The course is designed for graduate students in the social sciences. Students come primarily from the Educational Psychology department, within the College of Education and Human Sciences, and have backgrounds and affiliations in school and counseling psychology; cognition, learning, and development; and quantitative, qualitative, and psychometric methods. Students also often come from other departments within the college, such as Teaching, Learning, and Teacher Education; Child, Youth, and Family Studies; Educational Administration; and Nutrition. Finally, there are usually a handful of students from outside the college as well. Roughly half of the students in a given semester are taking the class specifically to inform their thesis or dissertation research.

Broader curricula

The course is required by programs within the Educational Psychology department, and is a prerequisite to an advanced measurement course offered in the department, EDPS 970. It fulfills a requirement for the Mixed Methods Certificate and contributes to the minor in Quantitative, Qualitative, and Psychometric Methods.

It is less clear how the course integrates into requirements or course sequences for other departments. However, based on student feedback, the course has been designed to meet the expressed needs of students to have exposure to the fundamentals of measurement and good practices for developing and evaluating educational and psychological tests. Because this is in part a methods course, it provides students with opportunities to develop core understanding and skills that will support them both in future courses that require understanding of these methods, and in their research where these methods must be applied.

Course goals

My main goals for the course are to help students 1) understand the theory and principles that are the foundation of sound educational and psychological measurement, and 2) successfully apply this understanding to practical applications in measurement development and evaluation. After taking this course, students should be able to think critically about the different stages of the test development process, and how each stage supports the primary objective of valid measurement. Students should be able to articulate the purpose of measurement within a given context, and then effectively develop an instrument or select an existing instrument that meets that purpose.

Achieving these goals will prepare students to be better teachers and researchers, as they will have the ability to get the most out of the measurement process. Their teaching, research, and counseling will thus improve as a result of taking the course.

The full list of learning objectives for the course is included in **Appendix B**. The objectives detail what students completing the course are expected to know and be able to do. The objectives are organized into the following ten topic areas.

1. Measurement, scales, and scoring
2. Testing applications
3. Cognitive test construction
4. Affective test construction
5. Reliability
6. Interrater reliability
7. Item analysis
8. Item response theory
9. Validity
10. Test evaluation

Teaching Methods

Overview

The teaching methods used in this course have evolved over the past four years since I started teaching it in the spring of 2013. In that time, I have taught six sections of the course, with the instruction, assignments, and assessments being modified to different degrees in each section. These modifications have been in response to student feedback about the teaching methods they find more and less effective. They have also been in response to my own perceptions of the teaching methods and course content that is most relevant and useful to my students.

In the spring 2016 section of the course, my instructional methods involved lecture and explanations of course content for roughly half of our time together in class, with class and group activities and discussion for the remaining time. My assessment methods involved take-home assignments, reflective writing assignments, and in-class quizzes. Instruction and assessment methods are each discuss in more detail below.

Instruction

Since I started teaching the course, I have steadily decreased the amount of time I spend lecturing to allow for more time to interact with students and have students interact with one another. In my first section of the course in spring 2013, at least 80% of class time consisted of lecture, often more. Although students were free to ask questions during lecture, and I regularly solicited their input and sought their active participation, group and class discussions and other interactive activities were limited. I currently spend 50% or less on lecture and then the majority of class time on discussions and activities. In my lectures, I review key concepts from the assigned readings. Discussions are then used to get students involved in interpreting and applying what they have been assigned to read. I have found that this use of class time is both more engaging and more likely to encourage students to come to class prepared to contribute. The result is more effective teaching and learning.

Here is an example from one of the last class meetings of the semester, where we cover the topic of test evaluation. Readings for this meeting include the chapter from the course notes corresponding to this topic, along with an article presenting a meta-analysis of predictive validity evidence for the GRE. In class, I would lecture for about 20 minutes on the main ideas from the course notes. Next, we would discuss as a class the assigned reading on the GRE. For about 30 minutes, students would summarize what they learned from the reading, while I asked questions about the variety of methods used to document predictive validity evidence for the GRE. Finally, for the last 30 minutes or so, students would get into groups to read through technical reviews of the psychometric properties of the GRE. Each group would briefly present their evaluation of the results.

Assessment

My assessment methods have also evolved since the first time I taught this course in 2013. These methods include larger take-home assignments, smaller discussion and reflective writing assignments, and quizzes. The number of assignments and quizzes has remained somewhat consistent across sections of the course. There have typically been four to five of the larger homework assignments, sometimes called labs, where students work in groups to answer a series of questions, sometimes while also analyzing data. These take-home assignments are spread across the fifteen weeks of the semester, with an assignment roughly every two to three weeks. These assignments account for roughly 50% of the final grade in the class.

The number of discussion and reflective writing assignments has also been consistent over sections, hovering around 10. In the spring 2016 semester these were worth a total of 20% of the final grade. Finally, the number of exams, sometimes called quizzes, has also remained relatively stable around 2 or 3 per semester. For spring 2016, we had three sit-down, in-class exams which were worth a total of 30%.

The content and structure of the assignments, reflective writings, and exams and quizzes has changed substantially over the years. Formerly, assignments focused mainly on statistical and psychometric analyses of test data using SPSS. The final assignment required students to evaluate two tests based on their technical properties, reliability, and validity evidence. Currently, three of the assignments address pieces of the final project wherein students describe a hypothetical testing application and evaluating the appropriateness of two tests for their chosen application. The assignments require students to articulate their construct, describe appropriate forms of reliability and validity evidence for their application, and then select and evaluate two published tests. The remaining two assignments require students to analyze sample data sets and interpret the results of item analyses, reliability analyses, and item response theory analyses. Thus, statistical analysis is somewhat deemphasized in favor of interpreting the information reported for published tests.

Reflective writings were originally written as group discussions that took place online. However, these discussions proved to be tedious and students felt that they were not beneficial to their learning. Most of the 10 reflective writings now require students to write open-ended responses to questions on the supplemental readings. These questions help ensure students are completing the readings before they come to class. They also give students the opportunity to think about main ideas from the readings, and tie the concepts covered into issues that interest them. For example, a reflective writing on reliability included the two questions shown below.

1. *Suppose the internal consistency reliability of a test you're considering using is 0.82. Describe how this value would be expected to change based on changes in test length. When would a decrease test length be expected to decrease internal consistency? And when would a decrease be expected to increase internal consistency?*
2. *Explain what a corrected item-total correlation represents in the context of your own measurement application. What does the index capture for a given item, and what would lead it to increase or decrease for that item?*

These questions help the student apply what they are learning in the reliability readings to their own measurement applications, which they had articulated in a previous assignment.

Finally, exams and quizzes have also been modified considerably over the years. Exams were initially longer and more difficult, consisting partly of multiple-choice question and partly of constructed-response questions, with average scores sometimes as low as 75%. Previously, I also had very small quizzes given throughout the semester, sometimes as often as once per week. These were intended to provide quick snapshots of student progress over the course of the semester. In the spring 2016 semester, I eliminated the smaller quizzes. I also modified the exams to be predominately constructed-response, with minimal multiple-choice questions. At the request of students, more time was also given for testing. As a result, mean percentages on the three exams were 88%, 92%, and 100%.

For future sections of the course, I am considering eliminating the exams altogether. The main reason for doing so is to be more consistent with the summer online sections of the course, where proctored exams are difficult to arrange.

Rationale for teaching methods

I believe that active engagement of students is critical to their learning. This belief has led me to rely less on lecture and more on discussions and activities during class. I feel that students have a better understanding of the material as a result. They have also reported feeling more prepared for the various forms of assessment, as the in-class activities include questions and exercises that resemble those in the assignments, reflective writings, and exams.

I also believe that a variety of assessment methods should be used to get a comprehensive view of what a student knows and can do, both during the semester and at the conclusion of the semester. Assignments require that students take an in-depth look at the application of course topics to real-world problems. Reflective writings require that students think critically about key issues before they are discussed in class. I use the results from assignments and reflective writings to inform my instruction as we proceed through the course topics. When results indicate that students are struggling, we spend more time reviewing a topic to ensure that everyone is on the same page before moving on.

Student Learning

In this portfolio I focus on my most recent section of EDPS 870, from the spring 2016 semester. This happened to be my smallest section of the course, with only 13 students. Students seem to be opting for my more accessible online version of the course, taught five weeks during the summer. Ten of my 13 students in the spring of 2016 consented to having their grades analyzed and reported within this portfolio. In this section I will present a summary of my findings from this analysis.

Overall, the students in this section did very well, with the majority scoring above 95% for the semester. While such high grades may not be typical for an undergraduate course, they are not uncommon in a graduate course like this one. The consistently high scores over the course of the semester indicate that students were able to do well on all of the assignments and exams. However, growth in their understanding is still evident in their work. A few examples are highlighted here.

In the first week of the semester, students were required to write cognitive test questions on the learning objectives that had been covered during that week. The requirements for this item writing activity are outlined in the box below. Note that students posted their items to a website that I created for activities like this one.

> *For this assignment you're going to write and submit for review two questions in Proola. The grade level for each should be college, and the subject should be Assessment Literacy.*
>
> *The first item should be selected-response with at least 3 response options, addressing learning objective 1.2. Compare and contrast measurement scales, including nominal, ordinal, interval, and ratio. See this item for a simple example: http://proola.org/items/74. Your item should be unique, but you're welcome to follow a similar approach.*
>
> *The second question should be constructed-response, i.e., short essay, addressing any one of these learning objectives:*
>
> *1.9. Define criterion referencing and identify contexts in which it is appropriate.*
> *1.10. Define norm referencing and identify contexts in which it is appropriate.*
> *1.11. Compare and contrast norm and criterion score referencing.*
>
> *Include an example response to the question within Proola that is thorough and deserving of full credit. See http://proola.org/items/43 for a selected-response item that could easily be converted into a constructed-response one.*

Results from this assignment revealed students limited understanding of principles of effective item writing. In subsequent assignments, students were required to comment on one another’s items, and then revise their items based on the feedback they received. The quality of their items consistently improved. Students’ work is available at the website www.proola.org. **Appendix C** contains two examples.

Growth was also evident in students' ability to articulate how course topics applied to real-world problems. One example from a reflective writing is provided in **Appendix C**. This example demonstrates the student's depth of knowledge regarding threats to validity evidence. They explain briefly what these threats are, and then discuss examples of them within the context of their own measurement application.

As noted above, the take-home assignments in this course require students to describe a hypothetical measurement application, and then describe how reliability and validity evidence could be collected for such an application. In their final assignment they identify two tests that could be used to support decision-making within their measurement applications and then they evaluate these tests. In this final assignment they are demonstrating their mastery of learning objectives from multiple topic areas. **Appendix C** contains an example of a complete final assignment for one group of two students. Their responses demonstrate a strong understanding of the material and how it can be applied to a real-world problem.

Summary

Lessons learned

Participating in this project and completing this portfolio have helped me be more thoughtful about what I expect students to learn and how I articulate those expectations and then teach and assess their learning in relation to them. I have utilized learning objectives since I started teaching this course in 2013. However, in this project I was more critical of my objectives and how they were taught and assessed. I was more careful about designing assessments around the learning objectives. As a result, the examples of student work discussed above can be linked directly to the objectives.

Participation in this project also encouraged me to be more thoughtful about collecting evidence of student learning. Until spring 2016, I had never required students to write and revise test items multiple times over the course of the semester. Thus, it was difficult for them and for me to understand how they had growth in this topic area. Measuring growth is often a challenge in a course such as this, where the content changes significantly over the course of the semester. However, I used my reflective writing assignments to revisit previous topics, integrating them with current ones, so as to document growth. Students seemed to appreciate this opportunity.

Future plans

In future sections of this course, I would like to improve my supplemental assigned readings to be more relevant to students. I will explore activities where they conduct literature reviews and bring to class examples of measurement in contexts that interest them. Related to this issue, in future sections of the course I plan to also address the issues presented at the start of the portfolio, that is, adapting the course content to the needs of the students, identifying an appropriate emphasis on psychometric analyses, and providing adequate resources for learning the statistical software. These are issues that students have identified as being important and I believe that in addressing them the course can be improved.

Appendix A: Syllabus

EDPS 870: Introduction to Measurement

Spring 2016 Syllabus

Course Info

Course: EDPS 470/870, Section 001
Term: Spring, 2016
Location: TEAC 112

Instructor: Tony Albano, PhD
Office: TEAC 235
Phone: 402-472-8911
Email: albano@unl.edu

Course Summary

This is a course in the theory and practice of educational and psychological measurement. Main topics include measurement applications in research and practice, reliability and validity, item writing and test design, and statistical analysis of test data. It is assumed that you have taken a course in introductory statistics.

Materials

There is no required textbook for the course. Electronic copies of assigned readings and course notes will be provided. Regular access to the internet is required. Assignments will be submitted online. Note that we will be using the Canvas learning management system (canvas.unl.edu) rather than Blackboard for all online activities. You can access Canvas in the same way that you access Blackboard.

The statistical software *R* will be used to complete some of the assignments in this course. *R* is free, open-source, and available online at cran.r-project.org. Note that *R* is run using syntax and a command line interface, rather than a point-and-click interface. We will use *RStudio*, free at www.rstudio.com, for running *R*.

We will also use a free web application called Proola (www.proola.org) to complete some assignments. You need to register for an account. Note that any information you post to this website will be publicly available to a community of educators and other students learning about assessment.

Grading

Readings will be assigned on a given topic before the topic is discussed in class. To get the most out of this class, keep up with readings as they are assigned and contribute to discussions.

Electronic copies of all readings will be posted online. A course outline, with topics, readings, quizzes, discussion activities, and assignments by date will be provided.
Participation in online activities accounts for 20 points or 20% of your final grade. These activities will all take place online and will involve developing and responding to discussion questions.

There will be four take-home assignments. Correct answers for these assignments will be discussed on the day they are due; as a result, any assignments submitted late will receive a score of zero. Each assignment is worth 10 points, totaling 40 points or 40% of your final grade. An in-class presentation that summarizes the four assignments will be worth 10 points or 10%.

Quizzes account for the remaining 30 points or 30% of your final grade. Each quiz will be worth 10 points. You must notify me at least 24 hours before a scheduled quiz if you cannot complete it; in extenuating circumstances, you will be allowed to schedule a make-up within the following two days; otherwise, you will receive a score of 0.

Final grades will be based on 20% participation, 40% assignments, 10% final presentation, and 30% quizzes. There will be no extra-credit assignments. The scale below will be used to determine final grades. A percentage of points below 60 will be assigned a failing grade. I will round grades to the nearest integer.

% Points	Grade	% Points	Grade
97-100	A+	77-79	C+
93-96	A	73-76	C
90-92	A-	70-72	C-
87-89	B+	67-69	D+
83-86	B	63-66	D
80-82	B-	60-62	D-

Some Standard Policies

Statement of academic dishonesty: Academic honesty is essential to the existence and integrity of an academic institution. The responsibility for maintaining that integrity is shared by all members of the academic community. To further serve this end, the University supports a Student Code of Conduct that addresses the issue of academic dishonesty.

Diversity statement: The University of Nebraska-Lincoln is committed to a pluralistic campus community through Affirmative Action and Equal Opportunity. We assure reasonable accommodation under the Americans with Disabilities Act. Students with disabilities are encouraged to contact me for a confidential discussion of their individual needs for academic accommodation. It is the policy of the University of Nebraska-Lincoln to provide flexible and individualized accommodation to students with documented disabilities that may affect their ability to fully participate in course activities or to meet course requirements. To receive accommodation services, students must be registered with the Services for Students with Disabilities (SSD) office, 132 Canfield Administration, 472-3787 voice or TTY.

Tentative Course Outline

Date	Topic	Reading	Assignment
1/12	0. Overview, Stat Review	Syllabus, C0	
1/14	1. Measurement, Scales, Scoring	C1, Stevens	
1/19	1. Measurement, Scales, Scoring	Popham, ETS, Wiki Beck	D1
1/21	2. Testing Applications	C2, NAEP, Hombo, myIGDIs	
1/26	2. Testing Applications	Proola, Personality Project	D2
1/28	3. Cognitive Test Construction	C3, Haladyna, Webb	
2/2	3. Cognitive Test Construction	Pearson, UH Rubrics	D3
2/4	4. Affective Test Construction	C4, Clark	
2/9	Review		D4, A1
2/11	**Quiz 1**		
2/16	5. Reliability	C5, Traub	
2/18	5. Reliability (lab)	Torfs, R examples	D5
2/23	6. Interrater Reliability	C6, Goodwin	
2/25	6. Interrater Reliability (lab)	R examples	D6
3/1	7. Item Analysis	C7, Software Notes	
3/3	7. Item Analysis (lab)	Ferketich, Spector, Livingston	D7
3/8	8. IRT	C8, ACT Report	
3/10	8. IRT (lab)	Hambleton	D8, A2
3/15	Review		
3/17	**Quiz 2**		
3/22	**No Class – Spring Break**		
3/24	**No Class – Spring Break**		
3/29	9. Validity	C9, TBD	
3/31	9. Validity	Hubley, Haynes	D9
4/5	**Buros Fieldtrip**	C10, Midgley	A3
4/7	**Buros Fieldtrip**		
4/12	10. Test Review	GRE Report, Kuncel	
4/14	10. Test Review	TBD	D10
4/19	Review		
4/21	**Quiz 3**		
4/26	Presentations		A4
4/28	Presentations		

Reading List

Label	Reference
Stevens	Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103* (2684), 677-680.
Popham	Popham, W., J. & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Jounal of Educational Measurement, 6*, 1-9.
ETS	Educational Testing Service (2012) GRE Guide to the Use of Scores
Wiki Beck	Beck Depression Inventory. (n.d.). In Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Beck_Depression_Inventory
NAEP	U.S. Department of Education, National Center for Education Statistics. (2013). *An Overview of NAEP* (NCES 2013-455).
Hombo	Hombo, C. M. (2003). NAEP and No Child Left Behind: Technical challenges and pratical solutions. *Theory and Practice, 42*, 59-65.
myIGDIs	myIGDIs: RTI in early childhood
Proola	http://proola.org/learn (Links to an external site.)
Personality Project	http://personality-project.org/readings-taxonomies.html (Links to an external site.)
Haladyna	Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309-334.
Webb	Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. *Language Arts*.
Pearson	Item Writing Guide
Assessment Systems	Assessment Systems Corporation. (2011). *Item writing and review guide*.
UH Rubrics	https://manoa.hawaii.edu/assessment/howto/rubrics.htm (Links to an external site.)
Clark	Clark, L. A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319.
Traub	Traub, R. E. & Rowley, G. L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice, 10*, 37-45.
Torfs	Torfs, P. & Brauer, C. (2012). A (very) short introduction to R.
Goodwin	Goodwin, L. D. (2001). Interrater agreement and reliability. *Measurement in Physical Education and Exercise Science, 5*, 13-34.
Ferketich	Ferketich, S. (1991). Focus on psychometrics: Aspects of item analysis. *Research in Nursing and Health, 14*, 165-168.
Spector	Conducting the Item Analysis
Livingston	Livingston, S. A. (2006). Item analysis. Handbook of test development, 421-441.
ACT Report	Introduction to test development for credentialing: Item Response Theory

Hambleton	Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice, 12*, 38-47.
Hubley	Hubley, A. M. & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123,* 207-215.
Haynes	Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*, 238-247.
Midgley	Midgley, C., Kaplan, A., Middleton, M., Maehr, M. L., Urdan, T., Anderman, L. H., Anderman, E., & Roeser, R. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology*, *23,* 113-131.
GRE Report	TBD
Kuncel	Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162-181.

Appendix B: Learning Objectives

EDPS 870: Introduction to Measurement

Spring 2016 Learning Objectives

This document contains the learning objectives for the entire course, including the concepts you will need to know and things you will need to be able to do for quizzes and assignments. For the most part, you will not need to memorize equations or calculate statistics by hand, but you will need to understand what they represent. The objectives are organized into ten topics, plus a topic zero addressing prerequisite material for the course.

0. Stat Review

1. Identify and use common statistical terms. For example: X, n and N, μ, σ, σ^2, ρ and r.
2. Calculate and interpret frequencies, proportions, and percentages.
3. Create frequency distribution plots and histograms and use them to describe the central tendency, variability, and shape of a distribution.
4. Calculate and interpret measures of central tendency and describe what they represent, including the mean, median, and mode.
5. Calculate and interpret measures of variability and describe what they represent, including the standard deviation and variance.
6. Calculate and interpret the correlation between two variables and describe what it represents in terms of the shape, direction, and strength of the linear relationship between the variables.
7. Interpret a scatterplot, including what it indicates about the shape, direction, and strength of the linear relationship between two variables.

1. Measurement, Scales, and Scoring

1. Define the process of measurement, and describe how educational and psychological measurement differ from physical measurement.
2. Compare and contrast measurement scales, including nominal, ordinal, interval, and ratio.
3. Identify the measurement scales needed in provided applications.
4. Define the term *construct* and describe how constructs are used in measurement.
5. Explain why scales are preferable to individual items when measuring a construct.
6. Compare and contrast dichotomous and polytomous scoring.
7. Compare and contrast a composite score (e.g., a total test score) and a subscore (e.g., an individual item score) in terms of what they represent and how they are used.
8. Create a generic measurement model and define its components.
9. Define criterion referencing and identify contexts in which it is appropriate.
10. Define norm referencing and identify contexts in which it is appropriate.
11. Compare and contrast norm and criterion score referencing in terms of their strengths and limitations and how they are used in practice.
12. Compare three examples of norm referencing: grade, age, and percentile norms.

2. Reliability

1. Describe the assumptions of the classical test theory (CTT) model.
2. Identify and describe the components of the CTT model (X, T, and E) and how they relate to one another.
3. Describe the relationship between the reliability coefficient and standard error of measurement.
4. Calculate the standard error of measurement and describe it conceptually.

5. Describe the difference between systematic and random error, including examples of each.
6. Compare and contrast the three main ways of assessing reliability: test-retest, parallel-forms, and internal consistency.
7. Compare and contrast the four reliability study designs, based on 1 to 2 test forms and 1 to 2 testing occasions, in terms of the sources of error that each design accounts for.
8. Use the Spearman-Brown formula to predict change in reliability.
9. Describe the formula for coefficient alpha, the assumptions it is based on, and what factors impact it as an estimate of reliability (e.g., the number of items, the relationship between items, the dimensionality of the test).
10. Estimate different forms of reliability using statistical software.
11. Describe factors related to the test, the test administration, and the examinees, that affect reliability.

3. Interrater Reliability

1. Describe the purpose of measuring interrater reliability, and how interrater reliability differs from traditional reliability.
2. Describe the difference between interrater agreement and interrater reliability.
3. Calculate and interpret indices of interrater agreement and reliability, including percentage agreement, Kappa, Pearson correlation, and intraclass correlation.
4. Identify appropriate uses of each interrater index, including the benefits and drawbacks of each.
5. Describe the three main considerations involved in using intraclass correlations.

4. Item Analysis

1. Identify items that may have been keyed or scored incorrectly.
2. Recode variables to reverse their scoring or keyed direction.
3. Calculate and interpret item difficulties and compare items in terms of difficulty.
4. Describe what item discrimination represents and how it is used in item analysis.
5. Calculate and interpret item discrimination indices.
6. Calculate and interpret item-total correlations and corrected item-total correlations.
7. Describe the relationship between item difficulty and item discrimination.
8. Describe how item bias can be estimated using scores on an external criterion variable.
9. Calculate and interpret alpha if item removed.
10. Distinguish between items that function well in a set and items that do not.
11. Remove items from an item set to achieve a target level of reliability.
12. Evaluate multiple-choice options using distractor analysis.

5. IRT

1. Describe the similarities and differences between item response theory (IRT) and CTT, as discussed in Hambleton, 1993.
2. Describe how IRT is used in item analysis and the test development process.
3. Identify the two main assumptions that are made when using a traditional IRT model (one addresses dimensionality the other addresses the number of item parameters).
4. Define the three item parameters and one ability parameter in the traditional, general IRT model, and describe the role of each in modeling performance.
5. Identify the main benefits of IRT over CTT.

6. Validity

1. Define validity in terms of test score interpretation and use.

2. Describe three main types of validity evidence (content, criterion, and construct) and identify examples of how each type is established, including the validation process involved with each.
3. Describe the structure and function of a test blueprint, and how it is used to provide evidence of content validity.
4. Identify appropriate sources of validity evidence for given testing applications and describe how certain sources of are more appropriate than others for certain applications.
5. Describe the unified view of validity and how it differs from and improves upon the traditional view of validity.
6. Identify threats to validity, including features of a test, testing process, or score interpretation or use that impact validity.

7. Testing Applications

1. Describe the general purpose of aptitude testing and some common applications.
2. Identify the distinctive features of aptitude tests.
3. Identify the main benefits and limitations in using aptitude tests in decision-making.
4. Describe the general purpose of standardized achievement testing and some common applications.
5. Identify the distinctive features of standardized achievement tests.
6. Identify the main benefits and limitations in using standardized achievement tests in decision-making.
7. Compare and contrast different forms of aptitude and achievement tests, and identify examples of each.

8. Cognitive Test Construction

1. Describe the purpose of a cognitive learning objective or learning outcome statement.
2. Describe how a test blueprint or test plan is used in cognitive test development to align the test to the content domain and learning objectives.
3. Compare items assessing different cognitive levels or depth of knowledge, e.g., higher-order thinking such as synthesizing and evaluating information versus lower-order thinking such as recall and definitional knowledge.
4. Identify and provide examples of selected-response item types (multiple-choice, true/false, matching) and constructed-response item types (short-answer, essay).
5. Compare and contrast selected-response and constructed-response item types, describing the benefits and limitations of each type.
6. Describe the overarching objective in item writing.
7. Write and critique cognitive items using the guidelines discussed in Haladyna, 1989 and Kline, 1986.

9. Affective Test Construction

1. Describe the main challenges associated with affective item writing discussed in Kline, 1986.
2. Describe how to reduce the effects of acquiescence and social desirability with affective items, as discussed in Kline, 1986.
3. Compare and contrast item types used in affective tests, describing the benefits and limitations of each type.
4. Describe how item analysis results are utilized in affective test development and piloting.
5. Write and critique affective items using the guidelines discussed in Haladyna, 1989 and Kline, 1986.

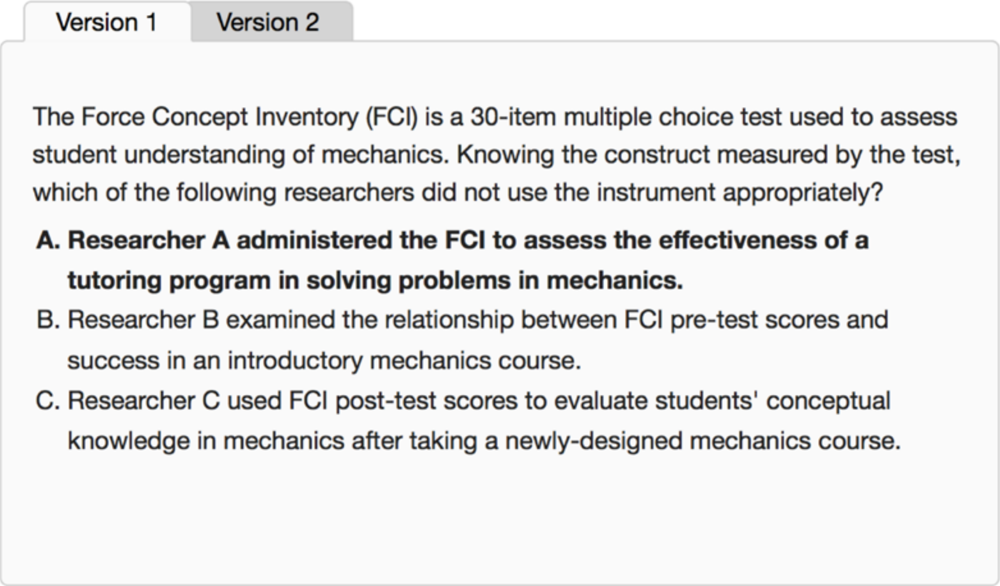
10. Test Review

1. Review and critique the documentation contained in a test review, test manual, or technical report, including:
   a. Data collection and test administration designs,
   b. Reliability analysis results,
   c. Validity analysis results,
   d. Scoring and reporting guidelines,
   e. Recommendations for test use.
2. Compare and contrast tests using reported information.
3. Use information reported for a test to determine the appropriateness of the test for a given application.

Appendix C: Examples of Student Work

Results from multiple item-writing activities demonstrated students' increasing ability to follow principles of effective item writing to improve item quality. Below are two example items from different students. The first version in each item is the original, and the second version is the revision demonstrating student growth.

Example item 1

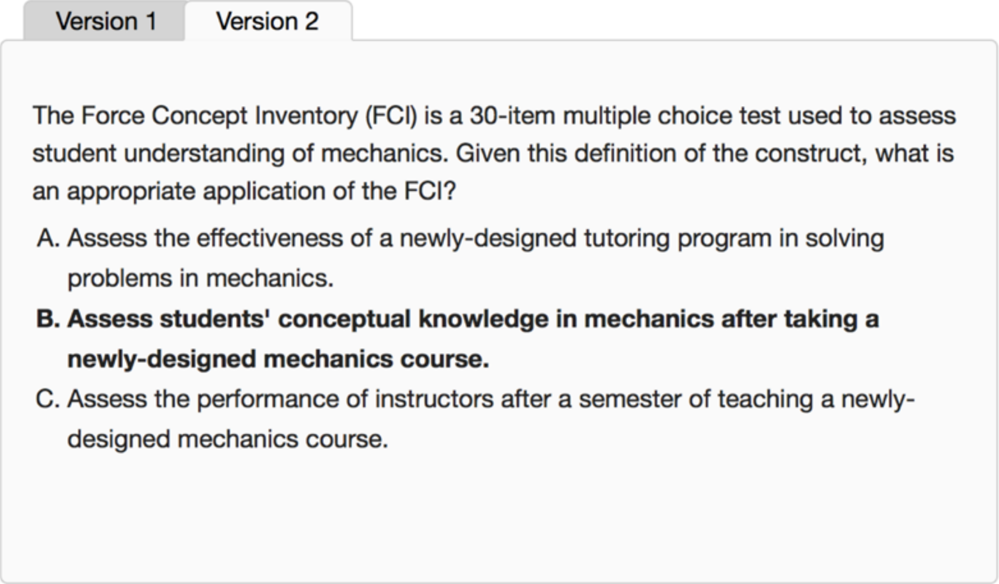


The Force Concept Inventory (FCI) is a 30-item multiple choice test used to assess student understanding of mechanics. Knowing the construct measured by the test, which of the following researchers did not use the instrument appropriately?

A. Researcher A administered the FCI to assess the effectiveness of a tutoring program in solving problems in mechanics.

B. Researcher B examined the relationship between FCI pre-test scores and success in an introductory mechanics course.

C. Researcher C used FCI post-test scores to evaluate students' conceptual knowledge in mechanics after taking a newly-designed mechanics course.



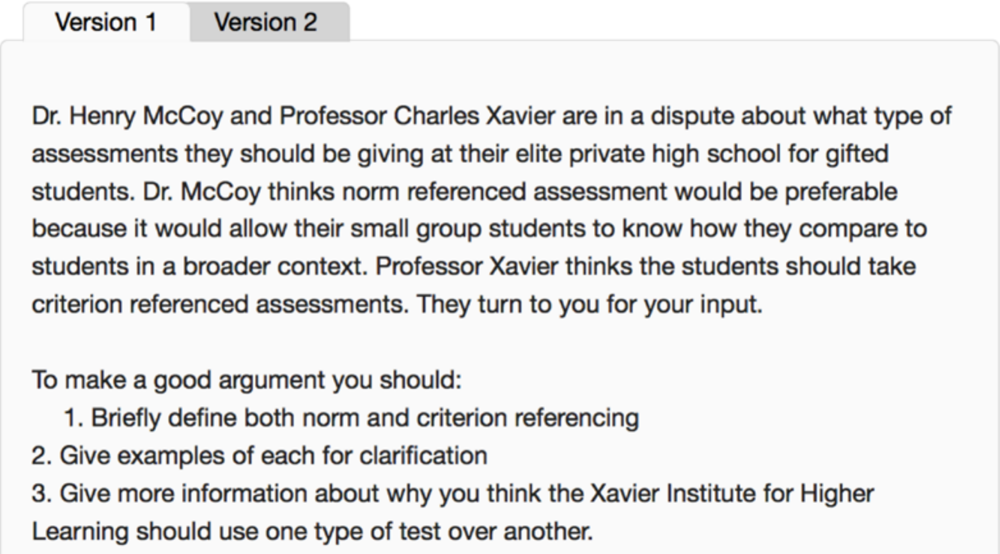
The Force Concept Inventory (FCI) is a 30-item multiple choice test used to assess student understanding of mechanics. Given this definition of the construct, what is an appropriate application of the FCI?

A. Assess the effectiveness of a newly-designed tutoring program in solving problems in mechanics.

B. Assess students' conceptual knowledge in mechanics after taking a newly-designed mechanics course.

C. Assess the performance of instructors after a semester of teaching a newly-designed mechanics course.

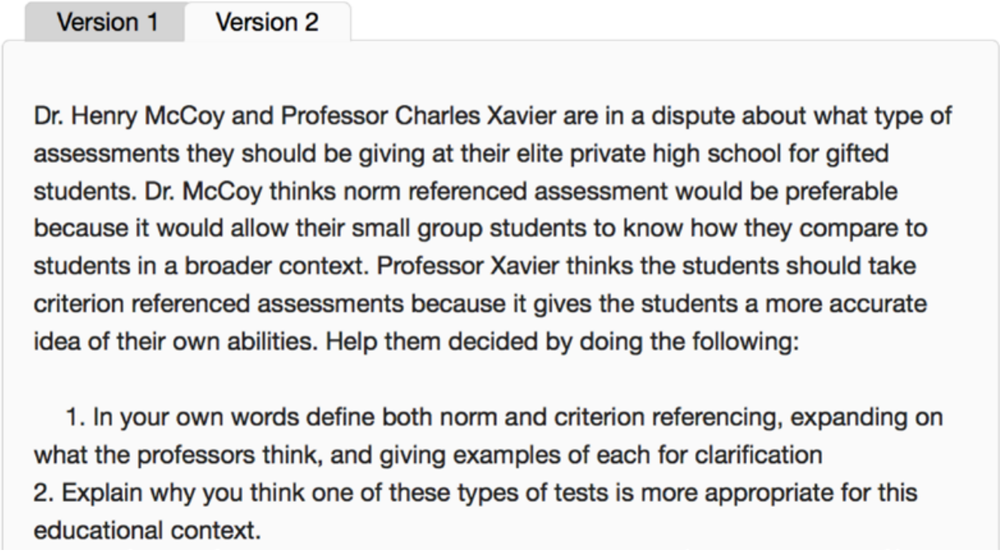
Example Item 2



Dr. Henry McCoy and Professor Charles Xavier are in a dispute about what type of assessments they should be giving at their elite private high school for gifted students. Dr. McCoy thinks norm referenced assessment would be preferable because it would allow their small group students to know how they compare to students in a broader context. Professor Xavier thinks the students should take criterion referenced assessments. They turn to you for your input.

To make a good argument you should:

1. Briefly define both norm and criterion referencing
2. Give examples of each for clarification
3. Give more information about why you think the Xavier Institute for Higher Learning should use one type of test over another.



Dr. Henry McCoy and Professor Charles Xavier are in a dispute about what type of assessments they should be giving at their elite private high school for gifted students. Dr. McCoy thinks norm referenced assessment would be preferable because it would allow their small group students to know how they compare to students in a broader context. Professor Xavier thinks the students should take criterion referenced assessments because it gives the students a more accurate idea of their own abilities. Help them decided by doing the following:

1. In your own words define both norm and criterion referencing, expanding on what the professors think, and giving examples of each for clarification
2. Explain why you think one of these types of tests is more appropriate for this educational context.

Results from a reflective writing assignment demonstrate students' ability to apply concepts to their own measurement applications.

> *The two major threats to validity discussed by Hubley and Zumbo (1996) are construct underrepresentation and construct-irrelevant variance. Construct underrepresentation refers to the failure of a measure to include important dimensions or facets of the construct. For instance, our testing application was designed to measure preservice teachers' knowledge of scientific models and modeling of the hydrologic cycle. To avoid construct underrepresentation, we lay out a blueprint identifying the important dimensions of what we intend to measure. Specifically, we want to make sure that the test includes questions about scientific models, scientific modeling, and the hydrologic cycle, to fully capture the construct that we are attempting to study. On the other hand, construct-irrelevant variance could arise from various sources - the measure is too broad and contains excess reliable variance associated with other distinct constructs, method variance, and error variance. As an example, our testing application could be too broad that some of the questions we have included about scientific modeling could be answered regardless of teachers' knowledge of the hydrologic cycle; an online version of the test (if we decide to create one) might have different results compared to a paper-based test; or the test may have a low coefficient alpha.*

The final group project for two students is reproduced here in its entirety. This project requires students to apply all of the concepts they have learned over the semester. Scores on this assignment were all above 95%.

Introduction

Our test purpose is to measure a preservice teacher candidate's knowledge on modeling in physics; to provide one type of quantitative information useful for designers of teacher preparation programs and of professional development programs. We want the measure to target areas in need of support. The construct we are measuring is preservice teachers' modeling knowledge in physics. This construct has two components – conceptual physics knowledge and scientific modeling of physics problems. The test is intended for preservice science teachers especially seeking for an endorsement in physical science or physics. Scores in the test would be used to assess conceptual physics knowledge and modeling of physics problems. Item-level information can be used to identify misconceptions in conceptual physics and modeling physics problems.

Literature Review

Developing and using models is one of the practices of science and engineering that the Next Generation Science Standards (NGSS) identify as essential for all students to learn (NGSS Lead States, 2013). While the term modeling is widely used in education, there is little research describing how teachers understand scientific models and how they enact modeling in a science classroom. Since secondary science teachers are expected to be able to design lessons that are aligned with the NGSS, they should have an understanding robust enough to be able to practice modeling and assess scientific models effectively. There is, therefore, a need for research on how subject matter knowledge and modeling knowledge can be developed among science teachers. A necessary precursor for this research endeavor is to develop an instrument that can be used to assess science teachers' subject matter knowledge and to determine how science teachers understand modeling and identify what conceptions or misconceptions they have.

There are several instruments that test subject matter knowledge in specific science subjects such as physics or physical science but they are often designed for students. Examples of these tests are the Misconceptions-Oriented Standards-Based Assessment Resources for Teachers (MOSART) (Sadler, Coyle, Miller, Cook-Smith, Dussault, & Gold, 2010), Force Concept Inventory (FCI) (Hestenes & Halloun, 1995), and the Brief Electricity and Magnetism Assessment (BEMA) (Ding, Chabay, Sherwood, & Beichner, 2006). The MOSART is a collection of science tests for middle school and high school students while the FCI and BEMA have been used to assess knowledge in mechanics, and electricity and magnetism, respectively. These tests are sometimes used as a proxy in research studies to measure teachers' subject matter knowledge since tests designed for teachers such as the Praxis subject assessments are often used as part of teacher licensing and certification.

Two potential unintended uses of our test would be as a high-stakes test to measure science teacher knowledge and modeling knowledge, and as a conceptual physics test. Drawing conclusions about preservice teachers' instructional effectiveness on scientific modeling would be a potential misuse of results. The test can be used as a diagnostic test or as a pre- and a *post-test component of professional development program. Determining experienced teachers mastery of scientific modeling would be inappropriate because the test was designed for preservice secondary science teachers.*

Test Descriptions of the NTE and CMEE

We did not find a test that focuses on assessing scientific modeling that was specifically designed for preservice science teachers. However, we found two tests that could potentially meet at least one of the components of the construct that we intend to measure. These tests are the National Teacher Examinations (NTE) and the Content Mastery Examination for Educators (CMEE) which are both applicable for preservice science teachers.

Both the NTE and the CMEE represent a suite of tests beginning teachers can use to show proficiencies in specific content or around specific concepts, used to inform teacher licensure. Because of the many different tests included in these suites, we decided to limit our focus on the physics or physical science components of the tests.

The NTE was developed by the Educational Testing Service (ETS). Information on this test was published in the Mental Measurements Yearbook in 1965. We found through our inspection of the NTE that it was a precursor to the Praxis exam, unfortunately, the Buros library does not carry the updated test, and thus we continued to use the older edition. The NTE was designed to measure the candidate's competence in the subject matter he/she hopes to teach, to prove useful in diagnosing strengths and weaknesses in the candidate's background preparation, to serve teacher education institutions in their efforts to improve teacher preparation and to provide one type of quantitative information useful in the selection for teachers. The test is meant to be given on scheduled days by qualified examiners in a large number of centers across the US. Colleges and universities may also schedule special testings for whole groups such as seniors majoring in education or new graduate students in education. The test is usually divided into two parts. The first part is called the Common Examinations which is administered in the morning on the day of the test. The Common Examinations take about 3 hours and include five components: Professional Information (90 items), English Expression (45 items), Social Studies, Literature, and Fine Arts (60 items), Science and Mathematics (50 items), and Nonverbal Reasoning (50 items). The second part of the NTE is the Optional Examinations which is administered in the afternoon on the day of the test. An examinee can elect to take one or two of the 80-minute Optional Examinations on 13 teaching fields based upon their area of teaching competence. Each of the optional tests has 85 to 120 items.

The CMEE was developed by the IOX Assessment Associates and was published in 1991 in Iowa. The test was designed for use in licensing teachers, counselors, administrators, and reading and media specialists. Upon initial conception, elements of this test mirrored ideas of the NTE , but eventually diverged focusing on ideas of diversity and bias as distinguishing factors. This test has a basic skills component that is also designed for use at entrance to colleges of education. The test is meant to be administered to groups of teachers, school administrators, other educational specialists, and entrants to colleges of education. Each test in the CMEE takes about 120 to 180 minutes to complete. The test is criterion-referenced. The

Physics test in the CMEE is scored in 5 different content areas: Mechanics, Heat, Wave Motion, Electricity and Magnetism, and Modern Physics.

Test Comparisons

NTE uses norm-referencing. Raw scores on each of the 5 Common Examinations and 13 Optional Examinations are converted to scaled scores with a mean of 60 and a standard deviation of 10 for the standardizing population. It should be noted the standardizing population is not the same for the common and optional test. On the other hand, CMEE uses criterion-referencing.

For our test purpose, criterion-referencing is more appropriate. Physics knowledge and scientific modeling knowledge are constructs defined in theory and this test would be formative, informing both instructors and pre-service teachers about high-need areas of additional support rather than trying to compare with another group. Because we will be measuring two different constructs both composite and component scores should be taken into consideration. Subscores would be given on each individual construct and a composite score will be derived focused on scientific modeling knowledge in physics for preservice teachers to better understand their strengths and weaknesses in these areas.

Multiple-choice items comprised the NTE and CMEE. However, the CMEE also requires a writing sample in its Basic Skills test. Due to the nature of the subject we are testing, in this case, physics knowledge and scientific modeling knowledge in physics, cognitive multiple choice items would be appropriate.

NTE is meant to be given on scheduled days by qualified examiners in a large number of centers across the US. Under some circumstances, colleges and universities can schedule special testing for groups such as seniors majoring in education or new graduate students in education. The test has two parts: Common Examinations and Optional Examinations. The Common Examinations, which usually take about 3 hours, are administered in the morning on the day of the test. The Common Examinations have 5 parts in a single booklet: Professional Information (90 items), English Expression (45 items), Social Studies, Literature, and Fine Arts (60 items), Science and Mathematics (50 items), and Nonverbal Reasoning (50 items). The Optional Examinations from 13 teaching fields are given in the afternoon. Each optional test takes 80 minutes. Each multiple choice item has an objective answer and does not require multiple raters. In the 1960s, the NTE test costs $9 to $13. The modern-day version of the NTE, which is the Praxis, costs $120.

The CMEE, similar to the NTE, is also meant to be administered in groups. Tests are administered through statewide programs on specified dates and at designated test centers. Each test has 100 multiple-choice items plus an essay in the Basic Skills Test and requires 2 to 3 hours. In 1994, one CMEE test costs $45. Today, the successor of the CMEE, which is the National Evaluation Series (NES) costs $95 per test. While the multiple-choice items do not require multiple raters, there was no information about the number of raters for the writing sample in the Basic Skills test.

For our test purpose, we intend to have the test given in groups of preservice secondary science teachers in their respective colleges or universities. Since we are using the test for *research, it would be given to the target population at no cost. The test would have 50 multiple-choice items and would take about 1 to 2 hours to complete. Each item would have a definite correct answer and would not require multiple raters.*

NTE reports that each of the common tests have K-R 20 values in the upper .80's and are thus reliable. The optional tests were reported to have K-R 20 values that ranged from .80 to .94 with the median value about .88. The reliability coefficients for the optional tests were reported for samples of about 300 cases. On the other hand, CMEE exams were field tested in two forms and reliability was established by examining equivalence, between alternate forms of each exam, of (a) content and (b) overall and within-section mean difficulty levels and overall within-section mean point-biserial coefficients. Overall mean difficulty (p-value) for the CMEE tests ranged from .56 to .77 and a median of .70. Between forms, the mean difficulty differed by .002 at the most. Overall mean discrimination (point-biserial r) for the CMEE tests ranged from .25 to .45 with a median of .31. Between forms, the mean discrimination differed by .008 at the most.

Results of the measures of reliability for both tests appear to be acceptable values. However, both test only had one source of reliability evidence. NTE reported K-R 20 values for internal consistency and did provide information about having multiple versions of their tests that might have needed evidence of parallel-forms reliability. CMEE reported parallel-forms reliability using point-biserial correlation from multiple field testing of alternate test forms but did not report internal consistency measures. CMEE appears to be more compelling because of the large number of field tests conducted although some were administered to small samples. Also, CMEE reported very small differences between levels of difficulty and consistency for parallel test forms. CMEE was more explicit about how reliability was conceptualized in relation to the purpose of presenting equivalent challenges to examinees across parallel forms.

The NTE establishes construct validity by performing job analyses to determine important knowledge and skills for beginning teachers as well as investigating state standards. The NTE also discusses how they piloted and field test some parts of their test. Content validity is established by having panels of experts work to develop test specifications (blueprints) and items. Items are then validated again by a second panel of experts. Criterion validity is not discussed and thus, no validity correlations were reported. A reviewer of the NTE explicitly called for more validation studies to be run and more clarification be given.
The CMEE provides a more robust description of its validity procedures. It achieves content validity through engagement with experts for each of their tests. They also go through three levels of content review defined as 'content identification', 'content determination', and 'national content review'. The content validation portion of their test is clearly outlined and detailed and result in the publishing of test blueprints for each of their content tests. Their construct validity is also well documented, identifying each subscale in their test blueprint as well as providing documentation justifying the inclusion and exclusion of content. The CMEE also extensively details pilot studies to strengthen their construct validity. However, no evidence could be found linking their construct to theory to discussing a nomological network of related ideas outside of the content itself. Finally, criterion validity is vaguely established by

loosely linking the CMEE to the NTE, no correlations were run, however, so this is a tenuous link at best. Overall, the CMEE provides more compelling validity evidence.

Conclusion

When considering the reliability, validity, and other pertinent information about the NTE and the CMEE compared to our testing applications, we found neither test truly met our needs. Ideally, we would have a test that is more explicitly calibrated towards modeling as well as the validity evidence to support its incorporation. If pressed, the CMEE would be more likely to meet our testing purpose. The reliability of both tests were reasonable in terms of their reported coefficients. The CMEE was also more explicit about how the misuse of the test would cause unreliable results. This lower reliability score is ameliorated by the richer description of the CMEE's validation process. Evidence can be found of content, criterion, and construct validity within their documentation. While some some semblance of each of those facets can be found in NTE's documents, CMEE's documentation is more organized and more thorough, clearly documenting the step by step processes, especially in terms of content validity.