EDUC632A: FUNDAMENTALS OF TEST CONSTRUCTION

Stephen G. Sireci, Ph.D.
156 Hills South
sireci@acad.umass.edu
http://www-unix.oit.umass.edu/~sireci
(413)545-0564 (voice)

Office Hours:
Monday: 1:30—3:00, Friday: 1:30—3:00
Other times by appointment

Course Syllabus and Schedule for Fall 2014

Course Objectives: This course will provide information on how to build and evaluate educational tests and how to effectively and appropriately interpret test results. Students will learn about the advantages and disadvantages of different assessment formats such as selected response items, performance assessments, and computer-based testing. Specifically, students will learn how to:

- describe fundamental aspects of test quality such as reliability and validity
- operationally define testing purposes
- develop a variety of item formats including multiple-choice and constructed-response items
- develop answer keys and scoring rubrics for different item formats
- evaluate tests and items using statistical and qualitative methods
- incorporate meaning into test score scales using both norm-referenced and criterion-referenced procedures
- use standard setting techniques to set "passing scores" and other performance standards on tests
- develop appropriate documentation to properly communicate the quality of an assessment
- understand the utility of educational assessments within the broader context of educational policy and decision making

The common theme unifying these knowledge and skill areas is the promotion of equity and fairness in educational testing. In addition, the course stresses the role of educational testing in improving student learning. Students will learn how to build quality tests aimed towards promoting valid score interpretation, and will learn how to evaluate the use of a specific test for a specific purpose. Measuring psychological phenomena such as what a student "knows and is able to do" is a complex endeavor. Test construction is both art and science; both aspects will be stressed in this course. Upon successful completion of this course, students will know how to (a) develop tests, (b) choose among already existing tests for a specific purpose, (c) use the results of standardized tests to help make decisions about students and educational systems, and (d) identify flaws in educational assessments.

Some specific topics covered in the course are:

- Purposes of Educational Tests
- Standards for Teacher Competence in Educational Assessment
- Standards for Educational and Psychological Testing
- Fundamental Elements of Test Quality (e.g., reliability, validity)
- Developing Multiple-Choice Items
- Developing and Scoring Constructed-Response Items
- Developing Portfolio Assessments
- Item Analysis

Specific topics covered (continued):

- Evaluating the Validity of Score-Based Inferences
- Standard Setting
- Innovative Item Formats and Computer-Based Testing
- Test Accommodations for individuals with disabilities and for English learners
- Sensitivity Review
- Ethical Issues in Test Construction, Selection, Administration, and Interpretation

Course Requirements

A. Attendance and Participation: Students expecting to receive course credit will need to attend all (or nearly all) classes, work their way through the suggested readings, and complete several assignments. In addition, students are expected to actively participate in class.
B. Assignments: In addition to weekly homework assignments, there are two major assignments for the course:

1) **Test Development**: Each student is required to develop an educational achievement test. This semester, we will be developing math and reading tests for adult basic education students in Massachusetts. Throughout the semester, students will progressively work on the design and development of their test. This process will involve writing items that may eventually be field-tested on adult education students and end up on the Massachusetts Adult Proficiency Tests.

2) **Technical Manual Development**: Each student is required to develop a technical manual that describes the development process for their test and provides important psychometric data for their test. Instructions for developing this manual will be provided in class. Chapter 7 of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) should be used to guide development of the manual.

Grading: Students' final grades are determined by their attendance and participation in class, and by their performance on their weekly assignments, draft test, final test, and technical manual. Late assignments will be reduced by one-letter grade for each day late (e.g., a maximum grade of "C" will be given to an exceptional draft test submitted two days late). Unforeseen emergencies, as determined by the professor, are exceptions to this policy. The table below illustrates the weighting used in calculating grades.

Activity	Weight
Attendance/Participation	.15
Weekly assignments	.30
First Test Draft	.05
Final Test Form	.25
Technical Manual	.25

Attendance/participation and all assignments are graded on a 0-100 scale. Final grades of 94-100 receive an A, 90-93 receive an A-, 87-89 receive a B+, 81-86 receive a B, 79-80 receive a B-, 77-78 receive a C+, 70-76 receive a C, and below 70 receive an F.

Suggested Textbook

Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching (9th edition)*. Upper Saddle River, NJ: Prentice-Hall.

This is a terrific book. If this is your first course in educational measurement, you should get it. If you have other books that cover test construction, you may not need it because I will provide numerous handouts on all topics covered in the course. There are 10th and 11th editions of this book, but I am using the 9th edition because it contains essentially the same material and is about 25% of the cost of the 11th edition. Feel free to use the 9th or a later edition. You can also use Gallagher's (1997) text *Classroom Assessment for Teachers*, but it is a bit dated.

Recommended Text

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association. Available for purchase at: http://www.aera.net/Publications/OnlineStore/BooksPublications/tabid/11736/BKctl/ViewDetails/SKU/AERWSTDEPTNEW/Default.aspx.

Required Readings

Each week you will have one or more reading assignments. Most of these readings are listed in the bibliography that appears next. I will distribute or post copies of all reading assignments. Articles that are likely to appear as reading assignments are denoted with an asterisk (*). The bibliography is stratified by topic area to facilitate finding publications in specific areas of interest.

Fundamentals of Test Construction Bibliography

Item Development/Formats

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education*, 4, 305-318.

Bennet, R., & Ward, W. (1993). *Construction versus choice in cognitive measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement, 53*, 241-247.

Cronbach, L. J. (1946). Response sets in objective tests. *Educational and psychological measurement*, 6, 475-494.

Downing, S. M., & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10, 61-82.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of Testing* (pp. 329-347). Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M. (1992). The effectiveness of several multiple-choice item formats. *Applied Measurement in Education*, 5, 73-88.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale: Lawrence Erlbaum.

*Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 2, 37-50.

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.

Haladyna, T. M., & Shindoll, R. R. (1989). Item shells: A method for writing effective multiple-choice items. *Evaluation & The Health Professions, 12,* 97-106.

Kreiter, C. D., & Frisbie, D. A. (1989). Effectiveness of multiple true-false items. *Applied Measurement in Education*, 2, 207-216.

Lukhele, R. Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, 31, 234-250.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*, 207-218.

Martinez, R. J., Moreno, R., Martin, I, & Trigo, M. E. (2009). Evaluation of five guidelines for option development in multiple-choice item writing. *Psicothema, 21*, 326-330.

Mentzer, T. L. (1982). Response biases in multiple-choice test item files. *Educational and Psychological Measurement*, 42, 437-448.

Osterlind, S. J. (1989). *Constructing test items*. Hingham, MA: Kluwer.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24(2)*, 3-13.

Sireci, S.G, Wiley, A., & Keller, L.A. (2002). An empirical evaluation of selected multiple-choice item writing guidelines. *CLEAR Exam Review, 13(2)*, 20-26.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, 31, 113-123.

Validity

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40, 955-959.

*American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A pre- liminary theory of action for summative and formative assessment. *Measurement, 8(2–3)*, 70–91.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20(4)*, 6-18.

Burger, S. E., & Burger, D. L. (1994). Determining the validity of performance-based assessment. *Educational Measurement: Issues and Practice,* 13(1), 9-15.

*Crocker, L. M., Miller, D., and Franks E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education*, 2, 179-194.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, New Jersey: Lawrence Erlbaum.

D'Agostino, J. V., et al. (2008). The rating and matching item-objective alignment methods. *Applied Measurement in Education, 21*, 1-21.

Hambleton, R. K., (1984). Validating the test score In R.A. Berk (Ed.), *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press, pp. 199-230.

*Joint Committee on Testing Practices (2004). *Code of Fair Testing Practices in Education*. Washington, DC: American Psychological Association. Available for download at http://www.apa.org/science/fairtestcode.html.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*,527-535.

Kane, M. (2006). Validation. In R. L. Brennan (Ed). *Educational measurement* (4th edition, pp. 17-64). Washington, DC: American Council on Education/Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73.

Kuehn, P. A., Stallings, W. M., Holland, C. L. (1990). Court-defined job analysis requirements for validation of teacher certification tests. *Educational Measurement: Issues and Practice*, 9 (4), 21-24.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment among curriculum, assessments, and instruction. *Review of Educational Research 4*, 1332-1361.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed.) (pp. 13-103). Washington, D.C.: American Council on Education.

Nelson, D. S. (1994). Job analysis for licensure and certification exams: Science or politics? *Educational Measurement: Issues and Practice*, 13(3), 29-35.

O'Neil, T., Sireci, S. G., & Huff, K. F. (2004). Evaluating the consistency of test content across two successive administrations of a state-mandated science assessment. *Educational Assessment, 9*, 129-151.

Porter, A. C., Polikoff, M. S., Zeidner, T., & Smithson, J. (2008). The quality of content analyses on state student achievement tests and content standards. *Educational Measurement: Issues and Practice, 27(4)*, 2-14.

*Sireci, S.G. (1998). Gathering and analyzing content validity data. *Educational Assessment*,5, 299-321.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research.*

Sireci, S. G. (2007). On test validity theory and test validation. *Educational Researcher, 36(8),* 477-481.

Sireci, S. G. (2009). Packing and upacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions and Applications* (pp. 19-37). Charlotte, NC: Information Age Publishing Inc.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*, 99-104.

Sireci, S.G., & Geisinger, K.F. (1998). Equity issues in employment testing. In J.H. Sandoval, C. Frisby, K.F. Geisinger, J. Scheuneman, & J. Ramos-Grenier (Eds.). *Test interpretation and diversity* (pp. 105-140). American Psychological Association: Washington, D.C.

Sireci, S.G., & Green, P.C. (2000). Legal and psychometric criteria for evaluating teacher certification tests. *Educational Measurement: Issues and Practice*, 19(1), 22-31, 34.

Smith, I. L., & Hambleton, R. K. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues and Practice*, 9(4), 7-10.

Wainer, H., & Braun, H. (1988). *Test validity*. Lawrenceville, NJ: Erlbaum.

Wainer, H., & Sireci, S. G. (2005). Item and test bias. *Encyclopedia of social measurement volume 2*, 365-371. San Diego: Elsevier.

Ying, L., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice, 26(4)*, 29-37.

Standard Setting

*Cizek, G. J. (1996). Setting passing scores. [An NCME instructional module]. *Educational Measurement: Issues and Practice*, 15 (2), 20-31.

Cizek, G. J. (2001). *Standard setting: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31-50.

Davis-Becker, S. L., Buckendahl, C. W., & Gerrow, J. (2011). Evaluating the bookmark standard setting method: The impact of random item ordering. *International Journal of Testing, 11*, 24-37.

Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsythe, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W. J., & Zwick, R. (2000). A response to "setting reasonable and useful performance standards" in the National Academy of Sciences "Grading the Nation's Report Card." *Educational Measurement: Issues and Practice, 19(2)*, 5-14.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice, 25 (1)*, 4-12.

Linn, R. L. (2003, September 1). Performance standards: Utility for different uses of assessments. *Educational Policy and Analysis Archives, 11(31)*. Retrieved September 1, 2003 from http://epaa.asu.edu/epaa/v11n31.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Meara, K. P., Hambleton, R. K., & Sireci, S. G. (2001). Setting and validating standards on professional licensure and certification exams: A survey of current practices. *CLEAR Exam Review, 12 (2)*, 17-23.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice, 27(4),* 15-29.

Reckase, M. D. (2006a). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice, 25 (2), 4-18.*

Reckase, M. D. (2006b). Rejoinder: Evaluating standard setting methods using error models proposed by Schulz. *Educational Measurement: Issues and Practice, 25 (3), 14-17.*

Sireci, S. G., Hambleton, R. K., & Pitoniak, M. J. (2004). Setting passing scores on licensure exams using direct consensus. *CLEAR Exam Review 15(1),* 21-25.

*Sireci, S. G., Randall, J., & Zenisky, A. (2012). Setting valid performance standards on educational tests. *CLEAR Exam Review, 23(2),* 18-27.

Sireci, S.G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education*, 12, 301-325.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Computer-Based Testing

Almond, R. G., Steinberg, L., S., & Mislevy, R. J. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *Journal of Technology, Learning, and Assessment, 1(5).* Available at http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml.

Drasgow, F., & Olson-Buchanan, J. B. (Eds.) (1999). Innovations in Computerized Assessment. Mahwah, NJ: Lawrence Erlbaum.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*, 20 (3), 16-25.

Luecht, R. M. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology*, available at http://www.testpublishers.org/Documents/JATT2005_rev_Criteria4CBT_RMLuecht_Apr2005.pdf

Luecht, R. L., & Sireci (2011). A review of models for computer-based testing. *Research report 2011-2012.* New York: The College Board.

Randall, J., Sireci, S. G., Li, X., & Kaira, L. (2013). Evaluating the comparability of paper- and computer-based science tests across sex and SES subgroups. *Educational Measurement: Issues and Practice, 31(4),* 2-12.

Sands, W. A., Waters, B. K. & McBride, J. R. (Eds.). (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.

*Sireci, S. G. (2004). Computerized-adaptive testing: An introduction. In J. Wall and G. Walz (Eds). *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 685-6947). Greensboro, NC: CAPS Press.

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S.M. Downing and T.M. Haladyna (Eds.), *Handbook of Testing* (pp. 329-347). Mahwah, NJ: Lawrence Erlbaum.

*Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, 12, 15-20.

Wainer, H. (Ed.). (2000). Computerized adaptive testing: A primer (2nd edition). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., & Kiley, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-201.

Zenisky, A.L., & Sireci, S.G. (2002). Technological innovations in large-scale assessment. Applied Measurement in Education, 15, 337-362.

Testing Special Populations

Abedi, J. & Ewers, N. (2013). *Accommodations for English learners and students with disabilities: A research based decision algorithm*. Smarter Balanced Assessment Consortium.

Almond, P., Winter, P., Cameto, R., Russell, M., Sato, E., Clarke, J., Torres, C., Haertel, G., Dolan, R., Beddow, P., & Lazarus, S. (2010). *Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities—A foundation for research*. Dover, NH: Measured Progress and Menlo Park, CA: SRI International.

Council of Chief State School Officers (1992). Recommendations for improving the assessment and monitoring of students with limited English proficiency. Washington, DC: Author.

Fisher, R. J. (1994). The Americans With Disabilities Act: Implications for measurement. *Educational Measurement: Issues and Practice*, 13(3), 17-26, 37.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, 6, 304-312.

Geisinger, K. F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education*, 7, 121-140.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*, 229-244.

Phillips, S. E. (1994). High-stakes testing accommodations: validity versus disabled rights. *Applied Measurement in Education*, 7, 93-120.

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P.W. Holland & H.Wainer (Eds.), *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Erlbaum.

Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher, 34(1)*, 3-12.

Sireci, S. G. (2009). No more excuses: New research on assessing students with disabilities. *Journal of Applied Testing Technology, 10 (2)*. Available at http://www.testpublishers.org/Documents/Special%20Issue%20article%201%20.pdf.

Sireci, S. G. (2011). Evaluating test and survey items for bias across languages and cultures. In D. Matsumoto and F. van de Vijver (Eds.) *Cross-cultural research methods in psychology* (pp. 216-240). Oxford, UK: Oxford University Press.

Sireci, S. G., DeLeon, B., & Washington, E. (2002, Spring). Improving teachers of minority students' attitudes towards and knowledge of standardized tests. *Academic Exchange Quarterly*, 162-167.

Sireci, S. G., Han, K. T., & Wells, C. S. (2008). Methods for evaluating the validity of test scores for English language learners. *Educational Assessment, 13*, 108-131.

*Sireci, S. G., & Mullane, L. A. (1994). Evaluating test fairness in licensure testing: the sensitivity review process. *CLEAR Exam Review*, 5 (2) 22-28.

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice, 25(3)*, 27-34.

Sireci, S. G., & Pitoniak, M. J. (2007). Assessment accommodations: What have we learned from research? *Large scale assessment and accommodations: What works?* In C. C. Laitusis & L. Cook (Eds.) (pp. 53-65).

Sireci, S. G., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research, 75*, 457-490.

*Thompson, S, & Thurlow, M. (2002, June). Universally designed assessments: Better tests for everyone! *Policy Directions, Number 14*. Minneapolis, MND: National Center on Educational Outcomes.

Performance Assessment

Clauser, B. E., Subhiyah, R. G, Nungester, R. J., Ripkey, D. R., Clyman, S. G., McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. *Journal of Educational Measurement*, 32, 397-415.

*Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4, 289-303.

Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment instrument. *Educational Measurement: Issues and Practice*, 12(2), 16-23.

*Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, 13(1), 5-8, 15.

Pearson, P. D. & Garavaglia, D. R. (1997). Improving the information value of performance items in large scale assessments. Paper commissioned by the NAEP Validity Studies Panel. Palo Alto, CA: American Institutes for Research.

Quellmalz, E. S. (1991). Developing criteria for performance assessments: The missing link. *Applied Measurement in Education*, 4, 319-331.

Shavelson, R. J., Baxter, G., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, 4, 347-362.

Cognitive/Principled (Evidence-Centered) Assessment Design

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25(4)*, 21-35.

Gorin, J. S. (2007). Test construction and diagnostic testing. In Leighton, J. & Gierl, M. (Eds.) *Cognitive diagnostic assessment for education* (pp. 173-201). Cambridge: Cambridge University Press.

Huff, K. & Goodman, Dean P. (2007). The demand for cognitive diagnostic assessment. In Leighton, J. & Gierl, M. (Eds.) *Cognitive diagnostic assessment for education* (pp. 19-60). Cambridge: Cambridge University Press.

Leighton, J.P. & Gierl, M.J. (Eds.). (2007). Cognitive diagnostic assessment for education: Theories and applications. Cambridge, MA: Cambridge University Press.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In *The Concept of Validity: Revisions, New Directions and Applications* (R. Lissitz, Ed.) (pp. xx-xx). Charlotte, NC: Information Age Publishing Inc.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the role of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 97-128). Mahwah, NJ: Lawrence Erlbaum.

Williamson, D. M., Mislevy, R. J., & Almond, R. G. (2004). Evidence-centered design for certification and licensure. *CLEAR Exam Review, 15(2)*, 14-18.

Portfolio Assessment

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, 13(3), 5-16.

Reckase, M. D. (1995). Portfolio assessment: A theoretical estimate of score reliability. *Educational Measurement: Issues and Practice*, 14(1),12-14, 31.

Miscellaneous

Anastasi, A. (1988). *Psychological testing* (6th edition). New York: Macmillan.

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service. (Reprint of chapter In R.L. Thorndike (Ed.) *Educational Measurement* (2nd Edition), Washington, DC: American Council on Education, 1971).

Berk. R. A. (Ed.), (1984). *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30(1)*, 3-12.

Chakwera, E., Khembo, D., & Sireci, S. G. (2004). High-stakes testing in the warm heart of Africa: The challenges and successes of the Malawi National Examinations Board. *Education Policy Analysis Archives, 12(29)* (see http://epaa.asu.edu/epaa/v12n29/.

*Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20 (4), 19-27.

Council of Chief State School Officers (2012). *Distinguishing formative assessment from other educational assessment lables.* Washington, DC: Author.

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of testing* (pp. 329-347). Mahwah, NJ: Lawrence Erlbaum.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519-521.

Glaser, R. (1994). Criterion-referenced tests: Part I Origins. *Educational Measurement: Issues and Practice*, 13(4), 9-11.

Goldstein, H. (1994). Recontextualizing mental measurement. *Educational Measurement: Issues and Practice*, 12, 16-19, 43.

*Hambleton, R. K., & Zenisky, A. (2003). Advances in criterion-referenced testing methods and practices. In C. R. Reynolds & R. W. Kamphaus (Eds.). *Handbook of psychological and educational assessment of children* (2nd Ed., pp. 377-404).

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Washington, DC: Council of Chief State School Officers.

Linn, R.L. (1994). Criterion-referenced measurement: a valuable perspective clouded by surplus meaning. Educational Measurement: Issues and Practice, 13, 12-15.

Linn, R. L. (2000). Assessments and accountability. Educational Researcher, 29(2), 4-16.

Livingston, S.A. (1982). Estimation of the conditional standard error of measurement for stratified tests. *Journal of Educational Measurement*, 19, 135-138.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and abilities. In R. Linn (Ed.), *Educational measurement*, (3rd ed. (pp. 335-366). Washington, D.C.: American Council on Education.

Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice*, 11(2), 9-15.

Nunnally, J. C. (1978). *Psychometric theory*. New York: MacGraw-Hill.

Phye, G. D. (1997). *Handbook of classroom assessment*. San Diego, CA: Academic Press.

Popham, W. J. (1992). A tale of two test-specification strategies. *Educational Measurement: Issues and Practice*, 11(2), 16-17,22.

Popham, W. J., Baker, E. L., Berliner, D. C, Yeakey, C. C., Pelligrino, J. W., Quenemoen, R. F., Roderiquez-Brown, F. V., Sandifer, P. D., Sireci, S. G., & Thurlow, M. L. (2001, October). *Building tests to support instruction and accountability: A guide for policymakers*. Commission on Instructionally Supportive Assessment. Available at http://www.nea.org/accountability/buildingtests.html

.

*Sireci, S.G. (2005). The most frequently unasked questions about testing. In R. Phelps (Ed.), *Defending standardized testing* (pp. 111-121). Mahwah, NJ: Lawrence Erlbaum.

Sireci, S. G. (2008). Are educational tests inherently evil? In D. A. Henningfeld (Ed.). *At issue: Standardized testing* (pp. 10-16). Detroit: Thompson Gale.

Sireci, S. G., Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., Shea, C. L., Han, K., Deng, N., Delton, J., & Hambleton, R. K. (2008). *Massachusetts adult proficiency tests technical manual: Version 2*: Amherst, MA: Center for Educational Assessment. Available at http://www.umass.edu/remp/CEA_TechMan.html.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28, 237-247.

Stiggins, R. J. (1997). *Student-centered classroom assessment*. New York: Merrill.

Ungerleider, C. (2003). Large-scale assessment: Guidelines for policy makers. *International Journal of Testing, 3*, 119-128.

*Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement, 26,* 191-208.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15(1),* 22-29.

Wall, J. E., & Walz, G. R. (Eds.) (2004). *Measuring up: Assessment Issues for teachers, counselors, and administrators*. Greensboro, NC: CAPS Press.

Webb, N. L. (2006). Identifying content for student achievement tests. . In S.M. Downing and T.M. Haladyna (Eds.), *Handbook of Testing* (pp. 155-180). Mahwah, NJ: Lawrence Erlbaum.

Zenisky, A. L., Sireci, S. G., Martone, A., Baldwin, P., & Lam, W. (2009). *Massachusetts adult proficiency tests technical manual supplement: 2008-2009*. Amherst, MA: Center for Educational Assessment.

Academic Honesty Statement

The integrity of the academic enterprise of any institution of higher education requires honesty in scholarship and research; thus academic honesty is required of all students at the UMass. Academic dishonesty is prohibited in all programs of the University. Academic dishonesty includes but is not limited to: cheating, fabrication, plagiarism, and facilitating dishonesty. Appropriate sanctions may be imposed on any student who has committed an act of academic dishonesty. Any person who has reason to believe a student has committed academic dishonesty should bring the information to my attention. Instances of academic dishonesty not related to this course should be brought to the attention of the appropriate department Chair. Students are expected to be familiar with this policy and the commonly accepted standards of academic integrity, and so ignorance of such standards is not normally sufficient evidence of lack of intent.

Plagiarism policy: Direct copying of someone else's work is not allowed. Printing out someone else's computer output, and handing it in as your own work, is also not allowed. Passing off someone else's work as your own will result in failing this course. Please see me if you have questions about this policy, or if you have trouble completing any assignments.

Accommodation policy:

I strive to provide an equal educational opportunity for all students. If you have a physical, psychological, or learning disability, you may be eligible for academic accommodations to help you succeed in this course. If you have a documented disability that requires an accommodation, please notify me as soon as possible, but no later than the third class, so that we may make appropriate arrangements to provide any needed accommodations for class or assignments.

TENTATIVE CLASS SCHEDULE FOR FALL 2014

Listed below are the topics that will be covered as well as a list of suggested readings.
The dates listed for each topic are tentative.

Class	Topics	Readings
9/2	Purposes of Educational Tests Standards for Teacher Competence in Assessment Norm- and Criterion-referenced testing	Brookhart (2011); Cizek (2001) Text Chs. 1-2, & Appendix D
9/9	Reliability and Validity Fundamentals Planning the Development of a Test	Sireci (2005), Text Chs., 3-5 Hambleton & Zenisky (2003)
9/16	Defining Test Content and Developing Test Specifications	Text Ch. 6, Handouts
9/23	Assessment Format Options Writing Multiple-Choice (MC) Items	Text Ch. 8 Haladyna & Downing (1989)
9/30	Writing MC Items (continued) Writing Other Objectively Scored Items	Haladyna (1992); Text Ch. 7
10/7	Developing Performance Assessments	Text Chs. 10-11; Dunbar et al. (1991); Linn & Burton (1994)
10/14	**No Class—Monday Schedule @ UMASS**	
10/21	Scoring Performance Assessments	Ch. 11, Handouts
10/28	Evaluating Tests for Content Validity Sensitivity Review	Crocker, et al. (1989); Sireci (1998) Sireci & Mullane (1994)
11/4	Field Testing and Item Analysis	Text Ch. 14, Handouts; Wainer (1989)
11/12	**NOTE: This class is a WEDNESDAY** Incorporating Meaning Into the Test Score Scale Setting Standards on Educational Tests	Text Ch. 19 Cizek (1996)
11/18	Understanding Test Results Developing a Technical Manual	AERA, APA, & NCME (2014)
11/25	Computer-based testing	Sireci (2004); Sireci & Zeniksy (2005) Wainer (1993); Zenisky & Sireci (2002)
12/2	Test Accommodations, Universal Test Design, Evidence-Centered Test Design	Geisinger (1994); Phillips (1994) Thompson & Thurlow (2002); Text Ch. 12; Huff & Goodman (2007)
12/16	**Final Test Form and Technical Manual Due (no class)**	

Indicates Online Class

At this point, we anticipate 4 online classes, with students having an option to participate remotely in a 5th class via Fuze. Details to be discussed during the first class.