EDUC821: Advanced Validity Theory and Test Validation

Stephen G. Sireci, Ph.D.
156 Hills House South
(413)545-0564 (phone)
(413)545-1523 (fax)
sireci@acad.umass.edu
http://people.umass.edu/~sireci/

Office Hours for Spring 2014:
Tuesdays: 12:30-2:30
Friday: 1:00-2:00
Other times by appointment

Course Objectives

Validity has been described as the most important aspect of test quality, yet many psychometricians have trouble articulating what validity is, and even more trouble validating tests. The purpose of this course is to introduce you to different perspectives and theories of test validity and to the process of accumulating validity evidence for educational tests. My goal is to familiarize you with many of the seminal articles in the validity literature and to illustrate how the process of test validation changes according to the specifics of the testing situation. We will address validation issues in educational testing, employment testing, certification testing, and other areas. Upon successful completion of this course, students will have a firm grasp of the technical and philosophical aspects of test validity and will have the skills to initiate and carry out a validity agenda for an educational testing program.

Topics to be covered in this course include:

- Origins and evolution of validity theory
- The AERA, APE, & NCME *Standards for Educational and Psychological Testing*
- Sources of validity evidence
- Responsibilities of test developers and test users
- Unitary conceptualization of validity
- Social considerations and equity issues in testing
- Statistical methods for validating test scores and evaluating tests
- Accumulating evidence in support of a validity argument
- Validating passing scores and other standards
- Assessing special populations—access and equity issues
- Legal versus psychometric criteria for evaluating tests
- Validating a theory of action

Course requirements

It is expected students will attend and actively participate in all classes. The reading load for this course is relatively heavy. I expect you to come to class prepared to discuss the extremely interesting reading assignments for that day. In addition, there will be weekly assignments, a midterm assignment, and a final assignment.

Grading: The requirements above are given a weight to determine your final grade as follows:

Activity	% of Grade
Attendance/Participation	20%
Weekly Assignments	30%
Midterm project	20%
Final project	30%

Attendance/participation and all assignments are graded on a 0-100 scale. Each missed class reduces the attendance/participation grade by 10 points. Medical illness and other acceptable emergencies will be exceptions to this policy. Final grades of 94-100 receive an A, 90-93 receive an A-, 87-89 receive a B+, 81-86 receive a B, 78-80 receive a C+, 70-77 receive a C, and below 70 receive an F.

Late assignments: 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 assignment submitted two days late).Unforeseen emergencies, as determined by the professor, will be exceptions to this policy.

Textbook

The only "textbook" required for this class is the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), which is available from AERA at www.AERA.net (see http://www.aera.net/AboutAERA/KeyPrograms/SocialJustice/StandardsforEducationalandPsychologicalTestin/tabid/10938/Default.aspx. We will use this resource throughout the course. Even though the next version is due out within the next year, this will be an important resource for you now and it will be interesting to see how it compares with the next version. In addition to this text, I will distribute intellectually stimulating articles each week. I suggest you assemble these articles into some type of binder for the course.

If you feel the need to purchase a supplementary text, the book *The Concept of Validity* edited by Lissitz (2009) would be a good choice. Another option would be *Test Validity* edited by Wainer and Braun (1988), which is very good, but a bit dated. The citation for these books and several others are included in the extensive list of references appended to this syllabus.

Resources for learning course material

You have at least four resources for helping understand the material presented in this course.

1) Me: I will do my best to present material clearly in class. Your class notes should be useful for completing assignments and examinations. In addition, I am available outside of class during my office hours and by appointment. You can also ask me questions using e-mail. See the top of this syllabus for office hours and e-mail address.

2) The reading assignments: I selected these assignments because I think they are exceptional for understanding the material taught in the course and represent significant contributions to the validity literature. The only exceptions are the articles I authored. I stuck those in there just to impress you and because it helps my ego to force others to read them.

3) The handouts: I will give you numerous handouts throughout the semester. These handouts are designed to summarize and supplement the lectures. I strongly recommend you review them in completing assignments and exams.

4) Each other: I encourage you to discuss class content and reading assignments with your classmates. Illuminating class discussion is a critical feature of this course.

Plagiarism policy:

It is expected that you will speak with others about course content and even work collaboratively on some class assignments. However, 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. In the University's Academic Regulations, plagiarism is defined as "knowingly representing the words or ideas of another as one's own work in any academic exercise. This includes submitting without citation, in whole or in part, prewritten term papers of another or the research of another, including but not limited to commercial vendors who sell or distribute such materials." See http://www.umass.edu/registrar/sites/default/files/academicregs.pdf. 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 reasonable academic accommodations to help you succeed in this course. If you have a documented disability that requires an accommodation, please notify me within the first two weeks of the semester so that I can make appropriate arrangements to provide any needed accommodations.

TENTATIVE Class Schedule

Spring 2014

Date	Topic	Readings*
1/27	Overview: What we all know about validity The 5 Sources of Validity Evidence	Lehman (1999); McGinn (1999); Jenkins (1946); Guilford (1946); Ebel (1961); AERA et al. (1999)
2/3	Validity Past, Present, and Future	Sireci (2009b), Rulon (1946), Cronbach & Meehl (1955)
2/10	Early Conceptualizations of Validity The Construct of Construct Validity	Campbell & Fiske (1959) Pitoniak, Sireci, & Luecht (2002)
2/18	*NOTE: THIS IS A TUESDAY* Multitrait-Multimethod Matrix	Messick (1989b, 13-34)
2/24	Validity as a Unitary Concept	Messick (1989b, 34-63)
3/3	Construct Validation	Messick (1989b, 63-92) Shepard (1993)
3/10	Test Interpretation and Test Use	Sireci (1998) Bhola, Impara, & Buckendahl (2003) Martone & Sireci (2009)
3/17	(**Midterms due**) SPRING BREAK—GO SOMEWHERE WARM!	
3/24	Is Content Validity Validity? Content Validity and Alignment	Kane (1992, 2006, 2013) AERA et al. (1999), pp. 9-24, 67-70
3/31	Argument-Based Approach to Validity	Linn (1984); Sireci & Talento-Miller (2006); Zwick & Schlemer (2004)
4/7	Gathering and Analyzing Criterion-related Evidence of Validity	AERA et al. (1999), pp. 163-169; Linn (2009); Sireci (in press); U.S. Dept. of Education (2009)
4/14	Validating a Theory of Action Validity for Accountability Testing	AERA et. al. (1999), pp. 91-108; Geisinger (2005); Sireci, Han, & Wells (2008); Phillips (2000); Sireci (2005a)
4/23	*NOTE: THIS IS A WEDNESDAY* Assessing Students with Disabilities and Linguistic Minorities	Lane (2014); Sireci & Geisinger (1998); Sireci & Parker (2006)
4/28	Social Consequences of Testing Employment Testing Legal Criteria for Evaluating Tests	
5/7	**Finals Due (No Class)**	

*Readings will be distributed on the date they are listed and will be discussed the next class.

EDUC821: Advanced Validity Theory and Test Validation

Bibliography
(Required reading assignments are indicated by *)

ACT (2000). *Content validity evidence in support of ACT's educational achievement tests: ACT's 1998-1999 national curriculum study*. Iowa City, IA: Author.

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

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, Language, and Assessment 1(5)*. Available from http://www.jtla.org.

American Educational Research Association (2000, July). AERA Position Statement: High-stakes testing in preK-12 education. Downloaded January 31, 2005 from http://www.aera.net/policyandprograms/?id=378.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

American Psychological Association, Committee on Test Standards. (1952). Technical recommendations for psychological tests and diagnostic techniques: A preliminary proposal. *American Psychologist*, 7, 461-465.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51*, (2, supplement).

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, D.C.: Author.

American Psychological Association (2000). *Report of the task force on test user qualifications*. Washington, DC: Author.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for educational and psychological tests*. Washington, D.C.: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement 10*, 67-78.

Anastasi, A. (1954). *Psychological testing*. New York: MacMillan.

Anastasi, A. (1980). Abilities and the measurement of achievement. *New Directions for Testing and Measurement, 5,* 1-10.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37,* 1-15.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, New Jersey: Lawrence Erlbaum.

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

Bennett, R. E., & Ward, W. C. (Eds.) (1993). *Construction versus choice in cognitive measurement*. Hillsdale, NJ: Erlbaum.

*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.

Binet, A. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'Année Psychologique, 12,* 191-244.

Binet, A., & Henri, B. (1899). La psychologic individuelle. *Amiee Psychol., 2,* 411-465.

Bingham, W.V. (1937). *Aptitudes and aptitude testing*. New York: Harper.

Braun, H. I., Jackson, D. N., & Wiley, D. E. (Eds.) (2002). *The role of constructs in psychological and educational measurement*. Mahwah, NJ: Erlbaum.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement*: *Issues and Practice, 17*(1), 5-9,30

Briggs, D. C. (2012, April). *Making inferences about growth and value-added: Design issues for the PARCC consortium.* Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.

Byrne, B. M., & van der Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10,* 107-132.

California Community Colleges (2001). *Standards, policies, and procedures for the evaluation of assessment instruments used in the California community colleges (4th edition).*

Camara, W. J. (2013). Defining and measuring college and career readiness: A validation framework. *Educational Measurement: Issues and Practice, 32*(4), 16-27.

*Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cascio, W.F., Outtz, J., Zedeck, S., & Goldstein, I.L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4*, 233-264.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issues and Practice, 22(3)*, 5-11.

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. (1971). Test Validation. In R.L. Thorndike (Ed.) *Educational measurement* (2nd ed., pp. 443-507). Washington, D.C.: American Council on Education.

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.

*Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Cureton, E.E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 621-694).

D'Agostino, J. V., & Bonner, S. M. (2009). High school exit exam scores and university performance. *Educational Assessment, 14*, 25-47.

D'Agostino, J. V., Welsh, M., & Corson, N. M. (2007). Instructional sensitivity of a state's standards-based assessment. *Educational Assessment, 12*, 1-22.

Dorans, N.J. & Lawrence, I. M. (1987). The internal construct validity of the SAT. (Research Report). Princeton, NJ: Educational Testing Service.

Ebel, R. L. (1956). Obtaining and reporting evidence for content validity. *Educational and Psychological Measurement, 16*, 269-282.

*Ebel, R.L. (1961). Must all tests be valid? *American Psychologist*, 16, 640-647.

Ebel, R.L. (1977). Comments on some problems of employment testing. *Personnel Psychology, 30*, 55-63.

Embretson (Whitley), S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bulletin, 93*, 179-197.

Fitzpatrick, A.R. (1983). The meaning of content validity. *Applied Psychological Measurement, 7*, 3-13.

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (Eds.) (1993). *Test theory for a new generation of tests.* Hillsdale, NJ: Erlbaum.

Geisinger, K. F. (1992). The metamorphosis in test validity. *Educational Psychologist, 27*, 197-222.

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

Geisinger, K. F. (2000). Psychological testing at the end of the millennium: A brief historical review. *Professional Psychology: Research and Practice, 31*, 117-118.

*Geisinger, K. F. (2005). The testing industry, ethnic minorities, and those with disabilities. (2005). In R. Phelps (Ed.), *Defending standardized testing* (pp. 187-203). Mahwah, NJ: Erlbaum.

Gierl, M.J., Leighton, J.P., & Hunka, S.M. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19(3)*, 34-44.

Goodenough, F. L. (1949). *Mental testing*. New York: Rinehart.

Green, P., & Sireci, S.G. (1999). Legal and psychometric issues in testing students with disabilities. *Journal of Special Education Leadership*.

*Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-439.

Guion, R. M. (1977). Content validity: the source of my discontent. *Applied Psychological Measurement, 1*, 1-10.

Guion, R. M. (1978). Scoring of content domain samples: the problem of fairness. *Journal of Applied Psychology, 63*, 499-506.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology, 11*, 385-398.

*Gulliksen, H. (1950a). Intrinsic validity. *American Psychologist, 5*, 511-517.

Gulliksen, H. (1950b). *Theory of mental tests*. New York: Wiley.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23(1)*, 17-27.

Hambleton, R. K. (1980). Test score validity and standard setting methods. In R.A. Berk (ed.), *Criterion-referenced measurement: the state of the art*. Baltimore: Johns Hopkins University Press.

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.

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.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18(4)*, 5-9.

Haertel, E. H. (2013). Getting the help we need. *Journal of Educational Measurement, 50(1)*, 84-90.

Holland, P.W., & Wainer, H. (Eds.). (1993*). Differential item functioning*. Hillsdale, New Jersey: Lawrence Erlbaum.

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.

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

International Test Commission (2010). *International Test Commission guidelines for translating and adapting tests*. Author. Available for download at http://www.intestcom.org.

Jaeger, R. M. (1998). Evaluating the psychometric qualities of the National Board for Professional Teaching Standards' Assessments: A methodological accounting. *Journal of Personnel Evaluation in Education, 12*, 189-210.

Jarjoura, D. & Brennan, R.L. (1982). A variance components model for measurement procedures associated with a table of specifications. *Applied Psychological Measurement, 6*, 161-171.

*Jenkins J. G., (1946). Validity for what? *Journal of Consulting Psychology, 10*, 93-98.

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

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

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. (2009). Validating the interpretations and uses of test scores. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions and Applications* (pp. 39-64). Charlotte, NC: Information Age Publishing Inc.

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

Kelley, T.L. (1927). *Interpretation of educational measurement*. Yonkers-on-Hudson, NY: World Book Co.

Kobrin, J., L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008a). Differential validity and prediction of the SAT. *College Board research report no. 2008-4*. New York: The College Board.

Kobrin, J., L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008b). Validity of the SAT for predicting first-year college grade point average. *College Board research report no. 2008-5*. New York: The College Board.

Koenig, J.A., Sireci, S.G., & Wiley, A. (1998). Evaluating the predictive validity of MCAT scores across diverse applicant groups. *Academic Medicine, 73*, 65-76.

Kuncel, N., Campbell, J. P., Ones, D. (1998). Validity of the Graduate Record Examination: Estimated or tacitly known? *American Psychologist, 53(5)*, 567-568.

LaDuca, A. (1994). Validation of professional licensure examinations. *Evaluation & the Health Professions, 17*, 178-197.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema*. doi: 10.7334/psicothema2013.258.

Lane, S., Parke, C. S., & Stone, C. A. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17(2)*, 24-28.

Lane, S., & Stone, C. A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21(1)*, 23-41.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23(4)*, 6–15.

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

Leach, M.M. & Oakland, T. (2007). Ethics standards impacting test development and use: A Review of 31 ethics codes impacting practices in 35 countries, *International Journal of Testing, 7*(1). 71-88.

*Lehman, N. (1999, September 6). Behind the SAT. *Newsweek, 134(10)*, 52-57.

Lehman, N. (1999). *The big test*. New York: Farrar, Straus, & Giroux.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16*, 294-304.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. Washington, D.C.: American Council on Education.

Linn, R. L. (1982). Ability testing, individual differences, prediction, and differential prediction. In A. Wigdor & W. Garner (Eds.), *Ability testing: Uses, consequences, and controversies* (pp. 335-388). Washington, DC: National Academy Press.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21*, 33-47.

Linn, R. L. (Ed.). (1989). *Educational measurement*, (3rd ed.). Washington, D.C.: American Council on Education.

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.

*Linn, R. L. (2009). The concept of validity in the context of NCLB. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions and Applications* (pp. 195-212). Charlotte, NC: Information Age Publishing Inc.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694 (Monograph Supplement 9).

Ludlow, L. H. (2001). Teacher test accountability: From Alabama to Massachusetts. *Education Policy Analysis Archives, 9(6).* Available at http://epaa.asu.edu/epaa/v9n6.html.

Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. *The Alberta Journal of Educational Research, 15,* 109-126.

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

*McGinn, D. (1999, September 6). The big score. *Newsweek, 134(10),* 46-51. Available at *http://www.newsweek.com/big-score-166300*.

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist, 30*, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, New Jersey: Lawrence Erlbaum.

Messick, S. (1989a). Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher, 18*, 5-11.

*Messick, S. (1989b). Validity. In R. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13-100). Washington, D.C.: American Council on Education.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128-165.

Mislevy, R. J. (2003). Rehabilitating psychometrics: Commentary on Pellegrino and Chudowsky's "the foundation of assessment." *Measurement: Interdisciplinary Research and perspectives, 1*, 162-165.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions and Applications* (pp. 83-108). Charlotte, NC: Information Age Publishing Inc.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*, 191-205.

National Council on Measurement in Education (2012). *Testing and data integrity in the administration of statewide student assessment programs*. Madison, WI: Author.

Nichols, P. D., Chipman, S. F. & Brennan, R. L. (Eds.) (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

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.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia, *Philosophical Transactions of the Royal Society A*, **187**, 253-318.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Penfield, R. D., & Miller, J. M. (2004). Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Applied Measurement In Education*, *17*(4), 359-370.

Penfield, R. D. (2010). Test-based grade retention: Does it stand up to professional standards for fair and appropriate test use? *Educational Researcher, 39*, 110-119.

Phelps, R. (Ed.), (2005). *Defending standardized testing*. Mahwah, NJ: Erlbaum.

Phelps, R. (Ed.), (2009). *Correcting fallacies about educational and psychological testing*. Washington, DC: American Psychological Association.

Phelps, R. (2012). The effects of testing on student achievement: 1910-2010. *International Journal of Testing, 12*, 21-43. **DOI:**10.1080/15305058.2011.602920

Phillips, S. E. (2000). GI Forum v. Texas Education Agency: Psychometric evidence. *Applied Measurement in Education, 13*, 343-385.

*Pitoniak, M. J., Sireci, S. G., & Luecht, R. M. (2002). A multitrait-multimethod validity investigation of scores from a professional licensure exam. *Educational and Psychological Measurement, 62*, 498-516.

Poggio, J. P., Glasnapp, D. R., Miller, M. D., Tollefson, N., & Burry, J.A. (1986, summer). Strategies for validating teacher certification tests. *Educational Measurement: Issues and Practice, 5(2),* 18-25.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29(4)*, 3-14.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5*, 285-301.

Popham, W. J. (1994). The instructional consequences of criterion-referenced clarity. *Educational Measurement: Issues and Practice, 13*, 15-20,39.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

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.aasa.org/issues_and_insights/assessment/Building_Tests.pdf.

Rabinowitz, S., & Brandt, S. (2001). Computer-based assessment: Can it deliver on its promise? *WestEd Knowledge Brief*. Downloaded January 30, 2005 from http://www.wested.org/cs/we/view/rs/568.

Raju, N. S., Lafitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517-529.

Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification exams. *Applied Measurement in Education, 14*, 369-415.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13-16.

*Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16,* 290-296.

Sanchez, E. I. (2013). Differential effects of using ACT college readiness assessment scores and high school GPA to predict first-year college GPA among racial/ethnic, gender, and income groups. *ACT research report series 2013-4*. Iowa City: ACT.

Sawyer, R. (1996). Decision theory models for validating course placement tests. *Journal of Educational Measurement, 33,* 271-290.

Schmidt, F. L. (1988). Validity generalization and the future of criterion-related validity. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 173-189). Hillsdale, New Jersey: Lawrence Erlbaum.

Schmidt, F. L., Viswesvaran, C., & Ones, D. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*, 901-912.

*Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 405-450.

Shepard, L. A. (1996). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16,* 5-24.

Shepard, L. A. (2003). Intermediate steps to knowing what students know. *Measurement: Interdisciplinary Research and perspectives, 1,* 171-177.

Sireci, S. G. (1997a). *Dimensionality issues related to the National Assessment of Educational Progress.* Commissioned paper by the National Academy of Sciences/National Research Council's Committee on the Evaluation of National and State Assessments of Educational Progress, [Document Number 619]. Washington, DC: National Research Council.

Sireci, S. G. (1997b). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16(1),* 12-19.

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

*Sireci, S. G. (1998b). The construct of content validity. *Social Indicators Research, 45*, 83-117.

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

Sireci, S.G. (2005b). Using bilinguals to evaluate the comparability of different language versions of a test. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.) *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Hillsdale, NJ: Lawrence Erlbaum.

Sireci, S. G. (2006). Content validity. In N. J. Salkind (Ed.) *Encyclopedia of measurement and statistics*. Thousand Oaks, CA: Sage.

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

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. (2009a). 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. (2009b). 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. (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. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*, 99-104.

*Sireci, S. G. (in press). A theory of action for validation. In R. Lissitz (Ed.). *The next generation of testing*. Charlotte: Information Age.

Sireci, S. G., Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., Shea, C. L., Han, K. T., Deng, N., Delton, J., & Hambleton, R. K. (2008). *Massachusetts Adult Proficiency Tests technical manual: Version 2*. Center for Educational Assessment Research Report No. 677. Amherst, MA: University of Massachusetts, Center for Educational Assessment. Available at http://www.umass.edu/remp/CEA_TechMan.html.

Sireci, S. G., & Faulkner-Bond (2014). Validity evidence based on test content. *Psicothema*. doi: 10.7334/psicothema2013.256.

Sireci, S. G. & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimensional scaling. *Applied Psychological Measurement, 16*, 17-31.

Sireci, S. G., & Geisinger K. F. (1995). Using subject matter experts to assess content representation: An MDS analysis. *Applied Psychological Measurement, 19*, 241-255.

*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.

*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., Hauger, J. B, Wells, C. S., Shea, C., & Zenisky, A. L. (2009). Evaluation of the standard setting on the 2005 grade 12 National Assessment of Educational Progress mathematics test. *Applied Measurement in Education, 22,* 339-358.

*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., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flawed items in the test adaptations process. In R.K. Hambleton, P. Merenda, & C. Spielberger (Eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). Hillsdale, NJ: Lawrence Erlbaum.

Sireci, S.G., Robin, F., Meara, K., Rogers, H.J., & Swaminathan, H. (2000). An external evaluation of the 1996 Grade 8 NAEP Science Framework. In N. Raju, J.W. Pellegrino, M.W. Bertenthal, K.J. Mitchell & L.R. Jones (Eds.) *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 74-100). Washington, D.C.: National Academy Press.

Sireci, S.G., Rogers, H.J., Swaminathan, H., Meara, K.,& Robin, F. (2000). Appraising the dimensionality of the 1996 Grade 8 NAEP Science Assessment Data. In N. Raju, J.W. Pellegrino, M.W. Bertenthal, K.J. Mitchell & L.R. Jones (Eds.) *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 101-122). Washington, D.C.: National Academy Press.

*Sireci, S. G., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admissions Test Scores. *Educational and Psychological Measurement*, *66*, 305-317.

Smarter Balanced Assessment Consortium (2010, June 23). *Race to the top assessment program application for new grants: Comprehensive assessment systems, CFDA Number: 84.395B*. OMB Control Number 1810-0699.

Society for Industrial Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures.* Bowling Green, OH: Author. Available at http://www.siop.org/_Principles/principlesdefault.htm.

Spearman, C. (1904). General intelligence: objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18,* 161-169.

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50,* 912-927.

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement: Issues and Practice, 17*(2), 20-23

Tenopyr, M.L. (1977). Content-construct confusion. *Personnel Psychology, 30,* 47-54.

Terman, L. M., & Childs, H.G. (1912). A tentative revision and extension of the Binet-Simon measuring scale of intelligence. *Journal of Educational Psychology, 3*, 61-74.

Terman, L. M. (1924). The mental test as a psychological method. *Psychological Review*, *31*(2), 93-117. doi:10.1037/h0070938

Terman, L. M., Lyman, G. Ordahl, G., Ordahl, L., Galbreath, N., & Talbert, W. (1915). The Stanford revision of the Binet-Simon scale and some results from its application to 1000 non-selected children. *Journal of Educational Psychology, 6*, 551-562.

Thorndike, E. L. (1931). *Measurement of intelligence*. New York: Bureau of Publishers, Columbia University.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York: Wiley.

Thorndike, R. L. (Ed.). (1971). *Educational measurement* (2nd ed.). Washington. D.C.: American Council on Education.

Thorndike, R. L. (1982). *Applied Psychometrics*. Boston: Houghton-Mifflin.

Thurstone, L. L. (1932). *The reliability and validity of tests*. Ann Arbor, Michigan: Edwards Brothers.

Toops, H. A. (1944). The criterion. *Educational and Psychological Measurement, 4*, 271-297.

Tucker, L. R. (1961). Factor analysis of relevance judgments: an approach to content validity. Paper presented at the Invitational Conference on Testing Problems, Princeton, NJ [Reprinted in A. Anastasi (Ed.) *Testing problems in perspective* (1966), (pp. 577-586). Washington, D.C.: American Council on Education.

Turney, A. H. (1934). The concept of validity in mental and achievement testing. *Journal of Educational Psychology*, *25*(2), 81-95. doi:10.1037/h0072182

U.S. Department of Education (2009). *Standards and Assessments Peer Review Guidance: Information and Examples for Meeting Requirements of the No Child Left Behind Act of 2001 (revised with technical edits January 12, 2009)*. Washington, DC: Author.

Vogt, D. S., King, D. W., & King, L. A. (2004). Focus groups in psychological assessment: Enhancing content validity by consulting members of the target population. *Psychological Assessment, 16,* 231-243.

Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement, 30*, 1-21.

Wainer, H., & Braun, H. I. (1988). *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.

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

Webb, N. L. (1999, August). Alignment of science and mathematics standards and assessments in four states. *Research Mongraph No. 18*. Madison, WI: National Institute for Science Education, University of Wisconsin-Madison.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*(1), 7-25.

Wells, C. S., Baldwin, S., Hambleton, R. K., Sireci, S. G., Karantonis, A. & Jirka, S. (2009). Evaluating score equity assessment for state NAEP. *Applied Measurement in Education, 22,* 394-408.

Willingham, W. W. (1988). Testing handicapped people—the validity issue. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 89-103). Hillsdale, New Jersey: Lawrence Erlbaum.

Willingham, W. W. & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher, 12*, 10-14.

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

Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics*, (pp. 45-79). Elsevier Science B.V.: The Netherlands.

Zumbo, B.D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. History repeats itself again. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions and Applications* (pp. 65-81). Charlotte, NC: Information Age Publishing Inc.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible Modeling of Measurement Data For Appropriate Inferences: Important Advances in Reliability and Validity Theory. In David Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.

Zwick, R., & Schlemer, L. (2004). SAT validity for linguistic minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice, 23(1)*, 6-16.