

EPSY 546 - Educational Measurement

Fall 2003 EPASW 2217 Tuesday 2:00 - 4:50

Professor: Dr. George Karabatsos Office Hours: Tue 10am-12pm (EPASW 1034)
Information: Phone: 413-1816 E-mail: georgek@uic.edu

Prerequisites: ED 501 - Data and Interpretation in Educational Inquiry
 EPSY 503/ED 503 - Essentials of Quantitative Inquiry in Education
 or equivalents, or consent.

Required text:

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to Nonparametric Item Response Theory*. Thousand Oaks, CA: Sage.

In addition, certain journal articles serve as additional reading assignments.

Introduction:

This course teaches Psychometrics, the practice that aims to establish measurement scales for psychological traits (e.g., ability, attitudes), as they manifest through data sets where persons responds to a set of test items, or where judges rate persons on their ability on various tasks.

It is straightforward to establish scales to measure directly observable, physical objects. However, it is not straightforward to establish scales for psychological traits (e.g., ability, attitudes), because such traits are not directly observable (latent). In fact, the establishment of a measurement scale for latent traits requires tests of the hypothesis that a given set of test data is consistent with the scale. This course covers various unidimensional psychometric models designed to measure such latent traits. Examples will be drawn primarily from the fields of education, psychology, and health care.

This course serves as a natural prerequisite to the course EPSY 535 – Item Response Theory (IRT). The IRT course covers more parametric Item Response Theory models (generalized partial credit models, graded response model, unfolding models, multidimensional models, cognitive models), and goes into greater depth with respect to the classical and Bayesian statistical inference frameworks of IRT, including parameter estimation, model fit evaluation, and model selection.

Course Objectives:

- 1) Introduce students to classical and contemporary statistical models for constructing scales for latent trait measurement. Models include those arising from nonparametric Item Response Theory, Parametric Item Response Theory (especially Rasch models), Classical Test Theory, Generalizability Theory, (exploratory and confirmatory) Factor Analysis, and Cultural Consensus Theory.
- 2) Introduce non-parametric Item Response Theory as a general framework for measurement. In fact, special cases of this framework include parametric Item Response Theory (including Rasch models), Classical Test Theory, and (exploratory and confirmatory) Factor Analysis.
- 3) Introduce students to classical statistical methods for testing the unidimensionality of a test, for evaluating the reliability and validity of tests, and for estimating the parameters of item response theory models.
- 4) To provide opportunities for students to develop data analytic skills using software such as SPSS, Winsteps, FACETS, EQS, and various psychometric programs written in S-PLUS that I have written.
- 5) Have students become acquainted with special applications of psychometrics, including test equating, computer adaptive testing, item banking, and detecting item bias.
- 6) Introduce students to concepts and methods of Bayesian inference, especially with respect to estimating, testing, and selecting models of item response theory. This is important because Bayesian inference is the most fundamentally-sound framework of statistical inference, and recent advances in computational methods have led to an explosion of research involving the applications and methods of Bayesian inference (especially in the Educational Testing Service, for example).
- 7) To provide opportunities for students to present their results of psychometric data analyses.
- 8) Have students write 2-page summaries (and critiques) of theoretical and applied measurement articles.
- 9) Have students write and present a 15 page manuscript suitable for publication in a psychometric journal.

Students will spend substantial amounts of time in the library and on the computer.

It is assumed that students will exert individual initiative in solving computing/analysis problems as they arise.

COURSE SCHEDULE FOR FALL SEMESTER 2003

DATE	TOPICS	READING
Aug	26 Assignments and tasks Scales of measurement The item response function (IRF) and the theory of conjoint measurement. Basic (and minimum) properties of measurement and unidimensionality.	
Sept	2 <u>Nonparametric Item Response Theory (NIRT): A general model for measurement</u> Methods for testing unidimensionality (monotonicity of IRFs) Covariance method, rest-score regression, H statistics Methods for testing invariant item ordering (non-intersection of IRFs): Covariance method, Rosenbaum's methods, H^T statistics, CMH test for item bias.	1
	9 <u>Finish/Review Nonparametric Item Response Theory</u> <i>Presentations on NIRT</i>	2
	16 <u>Classical test theory:</u> True scores, test reliability, test validity, norming/standardizing scores. <i>Presentations on NIRT, classical test theory</i>	3
	23 <u>Exploratory Factor Analysis:</u> Factor extraction (Principal components, principal axis, maximum likelihood) Factor rotation (Varimax/orthogonal, and oblique methods). <i>Presentations on classical test theory, exploratory factor analysis</i>	4
	30 <u>Confirmatory Factor Analysis:</u> Testing for hypothesized factor structures. <i>Presentations on exploratory factor analysis, confirmatory factor analysis</i>	5
Oct	7 <u>Rasch models for dichotomous items:</u> Model properties, Interpreting parameter estimates, testing person fit and item fit, testing local independence, parameter estimation, detecting item bias. <i>Presentations on confirmatory factor analysis, Rasch models for dichotomous items</i>	6
	<u>Rasch models for polytomous items:</u> 14 Model properties, Interpreting parameter estimates, testing person fit and item fit, testing local independence, parameter estimation, rating scale optimization, detecting item/rating scale bias. <i>Presentations on Rasch models for dichotomous items, Rasch models for polytomous items</i>	7
	21 <u>Special Topics:</u> Test Equating, computer adaptive testing (CAT), Item banking. <i>Presentations on Rasch models for polytomous items, and on equating/CAT/item banking</i>	8
	28 <u>Multi-Faceted Psychometrics:</u> Generalizability theory, Rasch FACETS model <i>Presentations on equating/CAT/item banking, and on G-theory/FACETS</i>	9
Nov	4 Lab-based Exam (2 hours). <i>Presentations on G-theory/FACETS</i>	
	11 <u>Bayesian inference and Item Response Theory:</u> Model Formulation and Estimation.	10
	18 <u>Bayesian inference and Item Response Theory:</u> Model Testing and Selection.	10
	25 Present Final Papers	
Dec	2 Present Final Papers	
	9 Exam Week: Final Papers due	

READINGS (required readings in bold)

Reading 1: Non-parametric item response theory.

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to Nonparametric Item Response Theory*. Fort Worth: Harcourt Brace Jovanovich College Publishers. pp. 1-64.

Reading 2: Non-parametric item response theory.

Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to Nonparametric Item Response Theory*. Fort Worth: Harcourt Brace Jovanovich College Publishers. pp. 91-148.

- Bol, G.W. (1994). *Implicational scaling in child language acquisition: The order of production of dutch verbal constructions*. In M. Verrips & F. Wijnen (Eds.), *Papers from the Dutch-German colloquium on language acquisition* (Amsterdam Series in Child language Development). Amsterdam: University of Amsterdam, Institute for General Linguistics.
- Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17(1), 31-44.
- Daut, H., Van der Maesen, C.E., & Mokken, R.J. (1996). Political efficacy: A further exploration. *Acta Politica*, 31, 350-371.
- De Jong, A., & Molenaar, I.W. (1987). An application of Mokken's model for stochastic cumulative scaling in psychiatric research. *Journal of Psychiatric Research*, 21, 137-149.
- DeVries-Griever, A.H.G., & Meijman, T.F. (1987). The impact of abnormal hours of work on various models of information processing: A process model of human costs of performance. *Ergonomics*, 30, 1287-1299.
- Gillespie, M., TenVergert, E.M., & Kingma, J. (1987). Using Mokken scale analysis to develop unidimensional scales. *Quality and Quantity*, 21, 393-408.
- Gillespie, M., TenVergert, E.M., & Kingma, J. (1988). Secular trends in abortion attitudes: 1975-1980-1985. *The Journal of Psychology*, 122, 323-341.
- Hemker, B.T., Sijtsma, K., Molenaar, I.W. (1995). Selection of unidimensional scales from a multidimensional itembank in the polytomous Mokken IRT model. *Applied Psychological Measurement*, 19, 37-352.
- Junker, B.W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement*, 24, 65-81.
- Karabatsos, G. (2003). A comparison of the detection power of 36 person fit statistics. *Applied Measurement In Education*.
- Kempen, G.I.J.M., Miedema, I., Ormel, J., & Molenaar, W. (1996). The assessment of disability with the Groningen Activity restriction Scale. Conceptual framework and psychometric properties. *Social Science & Medicine*, 11, 1601-1610.
- Kingma, J. (1984). A comparison of four methods of scaling for the acquisition of early number concept. *The Journal of General Psychology*, 110, 23-45.
- Kingma, J., & Loth, F.L. (1985). The validation of a developmental scale of seriation. *Educational and Psychological Measurement*, 45, 321-328.
- Kingma, J., & Ten Vergert, E.M. (1985). A nonparametric scale analysis of the development of conservation. *Applied Psychological Measurement*, 9, 375-387.
- Meijer, R.R., Sijtsma, K., Smid, N.G. (1990). Theoretical and empirical comparison of the Mokken and Rasch approach to IRT. *Applied Psychological Measurement*, 14, 293-298.
- Meijman, T.F., Thunnissen, M.J., & DeVries-Griever, A.G.H. (1990). The after-effects of a prolonged period of day-sleep on subjective sleep quality. *Work & Stress*, 4, 65-70.
- Middel, B.P., & Van Schuur, W. (1981). Background characteristics, attitudes towards the European community and towards Dutch politics, of delegates from CDA, D'66, PvdA, and VVD. *Acta Politica*, 16, 241-263.
- Moorer, O., & Suurmeyer, T.P.B.M. (1993). Unidimensionality and cumulateness of the Lonliness scale using Mokken scale analysis for polychotomous items. *Psychological Reports*, 73, 1324-1326.
- Paas, L.J. (1998). Mokken scale characteristic sets and acquisition patterns of durable and financial products. *Journal of Economic Psychology*, 19, 353-376.
- Paas, L.J. (1999). Refining RFM-variables through Mokken scale analysis for the purpose of optimal prospect selection: Application to ownership patterns of financial products. *Journal of Market Focused Management*, 3, 275-294.
- Ringdal, G.I., & Ringdal, K. (1993). testing the EORTC quality of Life Questionnaire on cancer patients with heterogeneous diagnosis. *Quality of Life Research*, 2, 129-140.
- Roorda, L.D., Roebroek, M.E., Lankhorst, G.J., Van tilburg, T., & Bouter, L.M. (1996). Measuring functional limitations in rising and sitting down: Development of a questionnaire. *Archives of Physical Medicine and Rehabilitation*, 77, 663-669.
- Rosenbaum, P.R. (1984). Testing the monotonicity and conditional independence assumptions of item response theory. *Psychometrika*, 49, 425-535.

- Rosenbaum, P.R. (1987). Comparing item characteristic curves. *Psychometrika*, 52, 217-233.
- Scheiblechner, H. (2003). Nonparametric IRT: Testing the bi-isotonicity of isotonic probabilistic models (ISOP). *Psychometrika*.
- Sijtsma, K., & Meijer, R.R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement*, 16, 149-157.
- Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, 22, 3-32.
- Suarmeyer, T.P.B.M., Doeglas, D.M., Moum, T., Briancon, S., Krol, B., Sanderman, R., Guillemin, F., Bjelle, A., & Van den Heuvel, W.J.A. (1994). The Groningen Activity restriction Scale for Measuring Disability: Its utility in international comparisons. *American Journal of Public Health*, 84, 1270-1273.
- Van Schuur, W.H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response Theory. *Political Analysis*, 11, 139-163.
- Verweij, A.C., Sijtsma, K., Koops, W. (1996). A Mokken scale for transitive reasoning suited for longitudinal research. *International Journal of Behavioral Development*, 19, 219-238.
- Zinn, F.D., Henderson, D.A., Nystuen, J.D., & Drake, W.D. (1992). A stochastic cumulative scaling method applied to measuring wealth in Indonesian villages. *Environment and planning A*, 24, 1155-1166.

Reading 3: Classical test theory / Reliability / Validity / Norming / Standardizing scores

- Kline, P. (1993). Reliability of tests: Practical issues. In Ch 1, *The Handbook of Psychological Testing*, pp. 5-15.**
- Kline, P. (1993). The validity of psychological tests. In Ch 2, *The Handbook of Psychological Testing*, pp. 15-28.**
- Kline, P. (1993). The classical model of test error. In Ch 3, *The Handbook of Psychological Testing*, pp. 29-41.**
- Kline, P. (1993). Standardising the test. In Ch 4, *The Handbook of Psychological Testing*, pp. 42-61.**
- Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist*, 50, 741-749.**
- Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, 74, 478-494.
- Campbell, D. T., & Fiske, D. W. (1969). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Crocker, L., Llabre, M., & Miller, M. D. (1988). The generalizability of content validity ratings. *Journal of Educational Measurement*, 25, 287-299.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Cortina, J.M. (1993). What is Coefficient Alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98-104.
- Cronbach, L. J. (1947). Test reliability: Its meaning and determination. *Psychometrika*, 12, 1-16.
- Cureton, E. E. (1950). Validity, reliability and baloney. *Educational and Psychological Measurement*, 10, 94-96.
- Fiske, D. W. (1973). Can a personality construct be validated empirically? *Psychological Bulletin*, 80, 89-92.
- Frederiksen, N. (1986). Construct validity and construct similarity: Methods for use in test development and test validation. *Multivariate Behavioral Research*, 21, 3-28.
- Green, S., Lissitz, R. W., & Mulaik, S. A (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, 37, 827-838.
- Guion, R. M. (1978). Content validity in moderation. *Personnel Psychology*, 31, 205-214.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563-575.
- Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, 1, 98-107.
- Messick, S. (1989). Validity (Chapter 2, pp. 13-103). In R. L. Linn, *Educational Measurement*, 3rd edition. New York: Macmillan.
- Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling*, 2(3), 255-273.
- Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17, 6-12.
- Novick, M.R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32, 1-13.
- Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods*, 5, 343-355.
- Reckase, M. (1998). Consequential validity from a test developer's perspective. *Educational Measurement: Issues and Practice*, 17, 13-16.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8, 81-84.

- Schmitt, N., & Ostroff, C. (1986). Operationalizing the behavioral consistency approach: Selection test development based on a content-oriented strategy. *Personnel Psychology*, 39, 91-108.
- Sijtsma, K., Molenaar, I.W. (1987). Reliability of test scores in nonparametric item response theory. *Psychometrika*, 52, 79-87.
- Zimmerman, D.W., Zumbo, B.D., & Lalonde, C. (1993). Coefficient alpha as an estimate of test reliability under violation of two assumptions. *Educational and Psychological Measurement*, 53, 33-49.

Reading 4: Exploratory Factor Analysis

- Kline, P. (1994). Principal Components Analysis. In Ch. 3, *An easy guide to factor analysis*, pp. 29-41. Routledge.**
- Kline, P. (1994). Other methods of factor analysis. In Ch. 4, *An easy guide to factor analysis*, pp. 42-55. Routledge.**
- Kline, P. (1994). Rotation of factors. In Ch. 5, *An easy guide to factor analysis*, pp. 56-79. Routledge.**
- Cella, D.F., Dineen, K., Arnason, B., Reder, A., Webster, K.A., Karabatsos, G., Chang, C., Lloyd, S., Mo, F., Stewart, J., & Stefoski, D. (1996). Validation of the Functional Assessment of Multiple Sclerosis Quality of life instrument. *Neurology*, 47, 129-139.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84-99.
- Phelps, L. (1995). Exploratory factor analysis of the WRAML with academically at-risk students. *Journal of Psychoeducational Assessment*, 13, 384-390.
- Smith, K. W. (1974). On estimating the reliability of composite indexes through factor analysis. *Sociological Methods and Research*, 2, 485-510.
- Thomson, B., & Daniel, L.G. (1996). Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement*, 56, 197-208.
- Wright, B.D., & Evitts, M.S. (1961). Direct factor analysis in sociometry. *Sociometry*, 24, 82.

Reading 5: Confirmatory Factor Analysis

- Kline, P. (1994). Confirmatory factor analysis and path analysis. In Ch. 6, *An easy guide to factor analysis*, pp. 80-99. Routledge.**
- Kline, P. (1994). Interpreting confirmatory factor analysis and path analysis. In Ch. 10, *Confirmatory factor analysis and path analysis. An easy guide to factor analysis*, pp. 157-172. Routledge.**
- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103, 411-423.
- Coovert, M. D., Craiger, P. J., & Teachout, M. S. (1997). Effectiveness of the direct product versus confirmatory factor model for reflecting the structure of multimethod-multirater job performance data. *Journal of Applied Psychology*, 82, 271-280.
- Edwards, J. E., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, 5, 155-174.
- Hattrup, K., Schmitt, N., & Landis, R. S. (1992). Equivalence of constructs measured by job-specific and commercially-available aptitude tests. *Journal of Applied Psychology*, 77, 298-308.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424-453.
- Kieffer, K.M. (1998). Some comments on the analytic traditions in EFA as against CFA: An analysis of selected research reports. Manuscript presented at the American Educational Research Association's annual conference.
- Marsh, H. W., Balla, J. R., & MacDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Journal of Applied Psychology*, 103, 391-410.
- Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement*, 15, 47-70.
- Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S. & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, 105, 430-445.
- Reuterberg, S.E., & Gustafsson, J.E. (1992). Confirmatory factor analysis and reliability: Testing measurement model assumptions. *Educational and Psychological Measurement*, 52, 795-811.
- Stephens, G.K., Szajna, B., & Broome, K.M. (1998). The career success expectations scale: An exploratory and confirmatory factor analysis. *Educational and Psychological Measurement*, 58, 129-141.
- Thompson, B., Webber, L., & Berenson, G. S. (1987). Factor structure of a children's health locus of control measure: A confirmatory maximum-likelihood analysis. *Educational and Psychological Measurement*, 47, 1071-1080.

Reading 6: Rasch models for dichotomous items

Wright, B.D., & Masters, G.N. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press. Chapter 3 on Models for Measuring, pp. 38-59.

- Agresti A. (1997) A model for repeated measurements of a multivariate binary response. *Journal of the American Statistical Association* 92(437) 315-321
- Esdaille, M, Shaw, F., Smith, M., & Valgeirsottir, S., (1994). Educational applications of conjoint measurement models. *International Journal of Educational Research*, 21, 635-651.
- Fisher, A.G., et al. (1994). Applications of conjoint measurement to the development of functional assessments. *International Journal of Educational Research*, 21, 579-592.
- Maller S. J. (1997) Deafness and WISC-III item difficulty: invariance and fit. *Journal of School Psychology* 35(3) 299-314.
- Shen, L., Yen, J. (1997). Item dependency in medical licensing examinations. *Academic Medicine*, 72(Supplement), S19-S21.
- ***Many other articles are also available in issues of the *Journal of Applied Measurement*.
(see <http://www.jampress.org/>)

Reading 7: Rasch models for polytomous items

- Betemps, E.J., Smith, R.M., Baker, D.G., & rounds-Kugler, B.A. (2003). Measurement precision of the Clinician Administered PTSD scale (CAPS): A Rasch model analysis. *Journal of Applied Measurement*, 4, 59-69.
- Elder, C., McNamara, T., & Congdon, P. (2003). Understanding Rasch measurement: Rasch techniques for detecting bias in performance assessments: An example comparing the performance of native and non-native speakers on a test of academic English. *Journal of Applied Measurement*, 4, 181-197.
- Heinemann A. W., Harvey R. L., Mcguire J. R., Ingberman D., Lovell L., Semik P., Roth E. J. (1997) Measurement properties of the NIH Stroke Scale during acute rehabilitation. *Stroke* 28(6) 1174-1180.
- Ip, E.H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, 66, 109-132.
- King, J.A., & Bond, T.G. (2003). Measuring client satisfaction with Public Education I: Meeting competing demands in establishing state-wide benchmarks. *Journal of Applied Measurement*, 4, 111-123.
- King, J.A., & Bond, T.G. (2003). Measuring client satisfaction with Public Education I: Comparing schools with state benchmarks. *Journal of Applied Measurement*, 4, 258-268.
- Lee J. (1997) State activism in education reform: Applying the Rasch model to measure trends and examine policy coherence. *Educational Evaluation and Policy Analysis* 19(1) 29-43.
- Leplege A., Rude N., Ecosse E., Ceinos R., Dohin E., Pouchot J. (1997) Measuring quality of life from the point of view of HIV-positive subjects: the HIV-QL31. *Quality of Life Research* 6(6) 585-594.
- Myford, C.M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education*, 15, 187-215.
- Norquist, J.M., Fitzpatrick, R., Jenkinson, C. (2003). Rasch measurement in the assessment of Amyotrophic Lateral Schlerosis patients. *Journal of Applied Measurement*, 4, 249-257.
- Peterman A. H., Cella D., Mo F., McCain N. (1997) Psychometric validation of the revised Functional Assessment of Human Immuno deficiency Virus Infection (FAHI) quality of life instrument. *Quality of Life Research* 6(6) 572-584
- Prieto L., Alonso J., Ferrer M. (1997) Are the results of the SF-36 health survey and the Nottingham Health Profile similar? A comparison in COPD patients. Quality of Life in COPD Study Group. *Journal of Clinical Epidemiology* 50(4) 463-473
- Sijtsma, K., & Hemker, B.T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, 25, 391-415.
- Smith, Jr., E.V. (2001). Evidence for the Reliability of Measures and Validity of Measure Interpretation: A Rasch Measurement Perspective. *Journal of Applied Measurement*, 2, 281-311.
- Smith, Jr., E.V. (2000). Understanding Rasch measurement: Metric development and score reporting in Rasch measurement. *Journal of Applied Measurement*, 1, 303-326.
- Yen, W.M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational and Behavioral Statistics*, 30, 187-213.
- **Many other articles are also available in issues of the *Journal of Applied Measurement*
(see <http://www.jampress.org/>)

Reading 8: Test equating, Computer Adaptive testing, Item banking.

- Choppin B. (1979) Testing the questions - the Rasch model and item banking. Chapter 5 in M. St.J. Raggett, C. Tutt, P. Raggett (Eds.) Assessment and Testing of Reading: Problems and Practices. London: Ward Lock Educational. In: <http://www.rasch.org/memo49.htm>**
- Linacre, J.M. Test Equating. In: <http://www.winsteps.com/winman/testequating.htm>**

- Meijer, R.R., & Nering, M.L. (1999). Computer adaptive testing: Overview and introduction. In: *Applied Psychological Measurement*, 23, 3, 187-194. Special Issue on Computerized adaptive testing.**
- Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. *Applied Psychological Measurement*, 11, 279-290.
- Chang, C.H., & Cella, D. (1997). Equating health-related quality of life instruments in applied oncology settings. In R.M. Smith (Ed.), *PM&R Secrets* (pp. 397-406). Philadelphia: Hanley & Belfus.
- Davis, L.L., Pastor, D.A., Dodd, B.G., Chiang, C., & Fitzpatrick, S.J. (2003). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. *Journal of Applied Measurement*, 4, 24-42.
- Lunz, M.E., Bergstrom, B.A., & Gershon, R.C. (1994). Computer adaptive testing. *International Journal of Educational Research*, 21, 623-633.
- Mills, C. N. (1999). Development and introduction of a computer adaptive graduate record examinations general test (Chapter 6, pp. 117-135). In F. Drasgow & J. B. Olson-Buchanan (eds.), *Innovations in Computerized Assessment*. Mahwah, NJ: Erlbaum.
- Schmitt, N., Gilliland, S. W., Landis, R. S., & Devine, D. (1993). Computer-based testing applied to selection of secretarial applicants. *Personnel Psychology*, 46, 149-166.
- Stahl, J. Bergstrom, B., Gershon, R. (2000). CAT administration of language placement examinations. *Journal of Applied Measurement*, 1, 292-302.
- Ward, A.W., & Murray-Ward, M. (1994). Guidelines for the development of item banks. An NCME instructional module. *Educational Measurement: Issues and Practice*, 13 (1), 34-39.
- Wolfe, E.W. (2003). Understanding Rasch measurement: Equating and item banking with the Rasch model. *Journal of Applied Measurement*, 1, 409-434.
- Wright, B.D., & Bell, S.R. (1984). *Item Banks: What, Why, How*. In: <http://www.rasch.org/memo43.htm>
- ***Any other article from the Special issue: Computer adaptive testing. (1999). *Applied Psychological Measurement*, 23, 187-261.
- ***Many other articles are also available in issues of the *Journal of Applied Measurement*. (see <http://www.jampress.org/>)

Reading 9: Generalizability Theory and Many-FACET Rasch Measurement

- Linacre, J.M. (1996). Generalizability theory and Many-facet Rasch measurement. In G. Englehard, Jr., & M. Wilson (Eds.), *Objective Measurement: Theory into Practice Volume 3* (pp. 85-98). Norwood: Ablex Publishing Corporation.**
- Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability Theory. *American Psychologist*, 44, 922-932.**
- Allen J.M. and Schumacker R.E.(1998) Team Assessment Utilizing a Many-Facet Rasch Model. *Journal of Outcome Measurement* 2:2, 142-158.
- Bachman, L.F., Lynch, B.K., and Mason M. (1995) Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, 12, 238-257.
- Banerji, M. (2000) Construct Validity of Scores/Measures from a Developmental Assessment of Mathematics using Classical and Many-Facet Rasch Measurement. *Journal of Applied Measurement*, 1:2, 177-198.
- Barrett, S. (2001) The impact of training on rater variability. *International Education Journal*, 2 (1), 49-58
- Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4), 27-34.
- Campbell S.K., Kolobe T.H.A., Osten E.T., Lenke M., Girolami G.L. (1995). Construct Validity of the Test of Infant Motor Performance. *Physical Therapy* 75:7 p.585-596
- Chi, E. (2001) Comparing Holistic and Analytic Scoring for Performance Assessment with Many-facet Rasch Measurement. *Journal of Applied Measurement* 2:4, 379-388.
- Crowley, S.L., Thompson, B., & Worchel, F. (1994). The children's depression inventory: A comparison of generalizability and classical test theory analyses. *Educational and Psychological Measurement*, 54, 705-713.
- Engelhard, G., Jr. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, 5(3), 171-191.
- Engelhard, G., Jr. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93-112.
- Engelhard, G., Jr. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, 33(1), 56-70.
- Engelhard, G., Jr. (1997). Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, 1(1), 19-33.
- Engelhard, G.Jr & Stone, G.E. (1998) Evaluating the Quality of Ratings Obtained From Standard-Setting Judges, *Educational and Psychological Measurement*, 58(2), 179-196.
- Fisher A.G. (1993). The assessment of IADL motor skills: An application of many-faceted Rasch analysis.

American Journal of Occupational Therapy, 47(4), 319-329.

- Fisher A.G. (1997) Multifaceted measurement of daily life task performance: Conceptualizing a test of instrumental ADL and validating the addition of personal ADL tasks. *Physical medicine and rehabilitation: State of the Art Reviews*. 11(2) : 289-303.
- Fisher A.G., Bryze K.A., Granger C.V., Haley S.M., Hamilton B.B., Heinemann A.W., Puderbaugh J.K., Linacre J.M., Ludlow L.H., McCabe M.A. & Wright B.D. (1994) Applications of conjoint measurement to the development of functional assessments. *International Journal of Educational Research*, 21(6), 579-593.
- Linacre, J.M., Englehard, G., Tatum, D.S., & Myford, C.M. (1994). Measurement with judges: Many-faceted conjoint measurement. *International Journal of Educational Research*, 21, 569-577.
- Lunz, M.E., Wright, B.D., & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, 3(4), 331-345.
- Lynch B. & McNamara T.F. (1998) Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 15: 158-180.
- Myford, C.M., & Wolfe, E.W. (2002). Detecting and measuring rater effects using many-facet Rasch measurement. *Journal of Applied Measurement*, to appear.
- Myford, C.M., & Wolfe, E.W. (2002). When raters disagree, then what: Examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings, *Journal of Applied Measurement*, 3, 300-324.
- Webb, N.M, Rowley, G.L., & Shavelson, R.J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development*, 21, 81-90.
- Wolfe, E.W., Moulder, B.M., & Myford, C.M. (2001). Methods for detecting differential rater functioning over time (DRIFT). *Journal of Applied Measurement*, 2, 256-280.
- Zhu W., Cole, E. L. (1996). Many-faceted Rasch calibration of a gross motor instrument. *Research Quarterly for Exercise and Sport*, 67(1), 24-34.
- Zhu W., Ennis C.D., Chen A. (1998) Many-faceted Rasch modeling experts' judgement in test development. *Measurement in Physical Education and Exercise Science*, 221-39.
- ***Many other articles are also available in issues of the *Journal of Applied Measurement*.
(see <http://www.jampress.org/>)

Reading 10: Bayesian Inference Approaches to IRT

Karabatsos, G., & Batchelder, W.H. (2003). Markov Chain Monte Carlo For Test Theory Without An Answer Key. *Psychometrika*.

- Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, 2, 389-423.
- Patz, R.J., & Junker, B.W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24, 146-178.
- Karabatsos, G., & Sheu, C.-F. (2003). Order constrained Bayes inference for dichotomous models of unidimensional non-parametric item response theory. *Applied Psychological Measurement*, to appear.

ASSIGNMENTS**

A) Summaries and presentations of measurement articles

B) Computer-Based Exam

C) Data Analysis Presentation and Paper

A. Summaries of measurement articles (all summaries worth 20% of total grade):

Each student must make 2 to 4 summary presentations, depending on class size.

Each presentation should be approximately 25 minutes, and must include:

1. the background and purpose of the research/study, (3 points)
2. the methods (participants, procedures/analyses, and instruments), (3 points)
3. results, (3 points)
4. discussion/conclusion and implications, (3 points)
5. and your detailed critique. (5 points)

Please provide appropriate handouts and develop meaningful overheads.

You may use figures from the articles (with a citation) as overheads.

A 2-page summary of your presentation is due the week following your presentation, at the beginning of class.

B. Data Analyses Presentation and Paper (worth 40% of total grade)

The data analyses and paper will consist of the **relevant output** from the software programs and a complete report stating the results. You may supply your own data or you may solicit faculty (education or other) for data.

The paper must be 15 double spaced-pages, using 1-inch margins, and in APA format (computer generated output must be placed in the Appendix, and is not part of the 15-page limit).

The presentation and paper must include:

Introduction –

Describe in detail the rationale/theory underpinning the data you will analyze (5 points).

Methods – (not necessarily in the following order).

Describe sample characteristics (3 points).

Describe the items on your test(s) (including their number and scoring format) (3 points).

Describe the unidimensional variable(s) you intend to measure with the test(s) (3 points).

Describe the psychometric models that you intend to use for data analysis, and their properties in terms of the item response function. You must employ at least four distinct psychometric models (15 points).

Describe the methods you will use to evaluate the unidimensionality, reliability, and validity of each of your test(s) (15 points).

Results – (not necessarily in the following order).

Empirically justify your choice of the “best” model (5 points).

With respect to the best model, discuss the amount of evidence for unidimensionality (10 points), reliability (10 points) and validity of your test(s) (10 points) (or lack thereof).

With respect to the best model, justify each step in your test modifications (removing items, removing persons, etc...) (5 points).

Discussion –

What modifications (if any) would improve the instrument? (3 points)

What are the implications of your study, with respect to the measurement and applications in the field of interest? (3 points)

Everyone starts with 90 points. I will deduct points from each section if you incorrectly interpret your results, fail to report/describe or fail to fully report/describe any of the information we have covered in class that is relevant to your particular investigation.

Please provide appropriate handouts and develop meaningful overheads for your presentation.

C. Computer-Based Exam (40% total):

On a scheduled class day, you will be given files of various test data sets. You will be tested on your ability to use WINSTEPS, FACETS, SPSS, EQS, and S-PLUS to perform psychometric analyses of these data sets, and answer questions concerning the interpretation of these analyses.

FINAL GRADES

Final grades will be given out using the following scale:

90% - 100%	A
79% - 89%	B
68% - 78%	C
57% - 67%	D
56% - Lower	F

There are no exceptions to the above grading scale, and no extra credit work will be accepted.

Incompletes will be considered for students with extenuating circumstances.

Poor performance on assignments will not be considered in a request for an incomplete.

Plagiarism is not tolerated by UIC. Any student who plagiarizes will receive a failing grade (F) for the course.

Disability Services:

UIC strives to ensure the accessibility of programs, classes, and services to students with disabilities. Reasonable accommodations can be arranged for students with various types of disabilities, such as documented learning disabilities, vision, or hearing impairments, and emotional or physical disabilities. If you need accommodations for this class, please let your instructor know your needs and he/she will help you obtain the assistance you need in conjunction with the Office of Disability Services (1190 SSB, 413-2183).