Boston College
Lynch School of Education

Psychometric Theory - ED669
(Spring 2004)

Psychometric Theory is the discipline that addresses the measurement and quantification of psychological phenomena (often refereed to as "latent traits"). Strictly speaking, psychological phenomena are not directly observable. Typically, they must be inferred from observations taken on some behavior that may be observed and is assumed to operationally define the unobservable characteristic (or "variable") that is of interest. An operational definition is most useful when it delineates the kinds of items or tasks that represent the high/hard and low/easy boundaries of the variable (whether it addresses behaviors, skills, attitudes, etc) of interest and differential points between those boundaries. After this theory driven specification of the variable has occurred, a "scale" comprised of independent items is developed to measure the hypothesized unidimensional variable. Data are then gathered, and various statistical models are then employed, to determine the extent to which the scale (or "measurement instrument") functioned as intended.

Instructor: Prof. Larry H. Ludlow
Campion Hall 336C 617-552-4221 Ludlow@bc.edu

Theme Quotes:

1. "The Reader may here observe the Force of Numbers, which can be successfully applied even to those things, which one would imagine are subject to no Rules. There are very few things which we know, which are not capable of being reduc'd to a Mathematical Reasoning; and when they cannot, it's a sign our Knowledge of them is very small and confus'd; and where a mathematical reasoning can be had, it's a great folly to make use of any other, as to grope for a thing in the dark, when you have a Candle standing by you." John Arbuthnot, 1692. In I. Todhunter, *A History of Mathematical Theory of Probability*. (Macmillan,p48-51, 1865).

2. "Psychometry, it is hardly necessary to say, means the art of imposing measurement and number upon operations of the mind...". F. Galton, Psychometric Experiments. *Brain, II*, 149-162, 1879.

3. "...that until the phenomena of any branch of knowledge have been subjected to measurement and number, it cannot assume the status and dignity of a science." Galton.

4. "I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of *science*, whatever the matter may be." Sir William Thomson, Lord Kelvin. Electrical Units of Measurement. Popular Lectures and Addresses, Vol 1 of 3. (London: Macmillan, 1889, p. 73-74)

5. "The grand, and indeed only, character of truth is its capability of enduring the test of universal experience, and coming unchanged out of every possible form of fair discussion". Sir John Herschel.

6. "Whatever exists, exists in some amount." E. L. Thorndike.

Ludlow's Challenge:

If it exists, it can be measured;
If it can't be measured, it doesn't exist.
Ludlow, L.H. Psychometrics Lectures, Boston College, February 1996

Course Objectives

A) Introduce you to Classical Test Theory (or "True Score Theory") and Item Response Theory (the Rasch model, in particular); and

B) Provide an opportunity for you to develop competent, practical data analysis and statistical/psychometric consulting skills.

You will spend considerable time in the library and on the computer. [It is assumed that you will exert individual initiative in solving computing/analysis problems as they arise.]

COURSE ASSESSMENTS

You will be evaluated on the following components:
a) data analyses (CTT, IRT)
b) literature reactions
c) measurement essay
d) Rasch presentation (either in class or at the NEERO conference)
e) class participation

Literature Reactions

The literature reactions (theory critiques, reviews, reaction papers) will take the form of at least 1-2 pages, (greater length is acceptable but is not encouraged) typed and double-spaced. They will be handed in the first **six** class meetings after the initial lecture. Their purpose is to introduce the literature to you and, in turn, your interests to me.

1) Begin the main body of your discussion with a direct quote from the article and its page number. Following the quote, write an analysis of its meaning to you. Your analysis should not be a paraphrased rendition of the quote but illustrative of your independent thinking on an interesting idea. For example, identify what may be wrong with the author's thinking on a question and suggest how the approach could be improved. Or, when your quote captures the brilliance of someone's thinking suggest ways its application may be broadened. Or, how can what we typically accept as "standard procedure" be improved by an obviously better way? Or, when you have encountered a particularly interesting topic, discuss its research potential

for you or its potential for incorporation into your current employment or dissertation interests. Or, you may wish to challenge "Ludlow's Challenge."

2) Organize the reaction papers and reviews according to the format shown below. In this form, your name and date are in the upper right hand corner and the full literature citation is in the upper left hand corner of the document.

Pearson, K. The Grammar of Science. Your Name
London : Adam and Charles Black, 1900. Date

" The classification of facts and the formation of absolute judgments upon the basis of this classification-judgments independent of the idiosyncrasies of the individual mind-essentially sum up the *aim and method of modern science*."

Page 6

Now would follow your reaction to the quote.

3) Your **first** Reaction Paper is to answer the question "What is Measurement?" You may consult any of the materials in this syllabus. BUT, I want to know what you in your own words think constitutes measurement. Your remaining Reaction Papers will be of the form presented in steps (1) and (2) above.

4) No papers are due for the evening in which analyses are submitted.

Data Analyses

The data analyses will consist of your interpretation of the output from the psychometric and statistical software programs (including the critical output). You may supply your own data or you may solicit Lynch School faculty for data. A reasonable way to satisfy this course component is to analyze the same data set for each psychometric model. The report should describe the sample, the variable being measured, items of the instrument (including their number and scoring format), the psychometric model and its psychometric properties, the interpretation of whether or not the data fit the model, and what modifications (if any) would improve the instrument.

Measurement Essay

The measurement essay will integrate your literature reactions and your understanding of class discussions. This is an opportunity for you to formally summarize your understanding of the essentials of measurement. One reasonable way in which to satisfy this component is to take a single topic and focus each reaction paper on some aspect of that topic. The measurement essay would then trace the development of your research. This essay should be 5-10 pages in length (potentially longer), typed, double-spaced, and fully referenced. In your essay you may include a discussion of topics that remain confusing, or appear as potentially researchable. Potential topics might include: authentic assessment, item banking, tailored testing, computer adaptive testing, Rasch applications, standard setting, one-parameter versus three-parameter models, differential item functioning (DIF), comparisons of estimation algorithms, goodness of fit tests, etc. You might

even address how, if any, your interpretation of the first reaction paper "What is measurement?" has shifted/clarified/been re-defined over the course of the semester.

Rasch Presentation

Your last data analysis will close with the Rasch model. You will provide a brief (15-20 minute) class presentation of your results.

Required Texts

Andrich, D. (1988). *Rasch Models for Measurement.* Newbury Park: Sage.

Bond. T. (2000). *Applying the Rasch model: Fundamental measurement*. LEA

Crocker, L. & Algina, J. (1986). *Introduction to Classical & Modern Test Theory*. NY: Holt, Rinehart & Winston.

Hambleton, R.K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage.

Wright, B.D. & Stone, M. (1979). *Best Test Design.* Chicago: MESA Press.

Prerequisite courses
ED462: Assessment and Test Construction
ED469: Intermediate Statistics
ED667: General Linear Models

PROPOSED TOPICS

1. History: Psychophysics to Psychometrics.

Principals, principles, and theoretical development.

WebCT Link Name **Required Readings:**

1. Chapters 1 and 3 of Crocker & Algina.
2. Preface of Bond
3. "Forward" of Wright & Masters.

Ludlow (1998) 4. Ludlow, L.H. (1998). Galton: The first psychometrician?. *Popular Measurement*, 1, 13-14.

Thurstone (1959) 5. Thurstone, L.L., Psychology as a quantitative rational science. In Thurstone, L.L. *The Measurement of Values*. University of Chicago Press, 1959.

Boring (1961) 6. Boring, E.G. The beginning and growth of measurement in psychology. In Woolf, H. (Ed.) *Quantification*. Bobbs-Merrill, 196l.

Ludlow & Alvarez-Salvat (2000) 7. Ludlow,L.H. & Alvarez-Salvat,R. (2000). Fechner: The man in the mask. *Popular Measurement, 3*, 5-6.

Suggested Readings:

1. Thurstone, L.L., Attitudes can be measured. In Thurstone, L.L. *The Measurement of Values*. University of Chicago Press, 1959.

2. Boring, E.G. Gustav Theodor Fechner. In Boring, E.G. *A History of Experimental Psychology* (2nd ed.). Prentice-Hall, 1950.

3. Kuhn, T.S. The function of measurement in modern physical science. In Woolf, H. (Ed.) *Quantification*. Bobbs-Merrill, 1961.

4. Thurstone, L.L. Psychophysical analysis. In Thurstone, L.L. *The Measurement of Values*. University of Chicago Press, 1959.

5. Jones, L.V. The nature of measurement. In *Educational Measurement* (2nd ed). Thorndike, R.L. (Ed) (2nd Ed). American Council on Education, 1971.

6. Stevens, S.S. Mathematics, measurement, and psychophysics. In Stevens, S.S. (Ed). *Handbook of Experimental Psychology*. Wiley, 1951.

7. Galton, F. (1879). Psychometric experiments. *Brain, II*, 149-162.

2. Classical True Score Theory:

Theory, assumptions, applications. SPSS reliability and factor analysis computer output interpretation of TASC data.

WebCT Link Name **Required Readings:**

1. Chapters 5-7, 13-14 of Crocker & Algina.

Spearman (1904) 2. Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Allen & Yen (1979) 3. Allen, M.J. & Yen, W.M. Classical True-Score Theory (Ch. 3) in *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole, 1979.

Lord (1980) Chapter 1 4. Lord, FM. *Applications of Item Response Theory to Practical Testing Problems*. Ch 1, LEA, 1980.

CTT TASC output 5. Ludlow, L.H. CTT TASC output.

Suggested Readings:

1. Traub, R.E. & Rowley, G.L. (1991). Understanding reliability. *Educational Measurement: Issues and Practice*, *10*, 37-45.

2. Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review*, *72*, 143-155.

3. Loevenger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, 5, 493-504.

4. Thurstone, L.L. Psychological Implications of Factor Analysis. Psychometric Laboratory Paper #44. The University of Chicago, Sept., 1947.

5. Thurstone, L.L. Psychological Assumptions of Factor Analysis. Psychometric Laboratory Paper #51. The University of Chicago, Feb.,1949.

6. Gould, J. (1981). Chapter 6 in *The Mismeasure of Man*. NY: Norton.

7. Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *American Journal of Psychology*, *15*, 201-293.

8. Hattie, J., Jaeger, R.M. & Bond, L. (1999). Persistent methodological questions in educational testing. *Review of Research in Education*, 24, Chapter 11. Washington, DC: AERA.

9. Traub. R.E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 8-14.

10. SPSS Chapter on Reliability calculations: http://www.spss.com/tech/stat/algorithms/11.0/reliability.pdf

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

12. Wainer, H. (1986). Can a test be too reliable? *Journal of Educational Measurement*, 23, 171-173.

13. Masters, G. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, 25, 15-29.

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

15. Cronbach, L.J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 3,* 391-418.

16. Mislevy, R.J. (2004). Can there be reliability without "reliability". *Journal of Educational and Behavioral Statistics, 29,* 241-244.

17. Moss, P.A. (2004). The meaning and consequences of "reliability". *Journal of Educational and Behavioral Statistics, 29,* 245-249.

3. Guttman's Scale Theory:

Theory, assumptions, applications. Interpretation of Hillock's Taxonomy of Reading Skills Hierarchy.

WebCT Link Name	**Required Readings:**
Stouffer (1950)	1. Stouffer, S.A. An Overview of the Contributions to Scaling and Scale Theory. In *Measurement and Prediction*, Stouffer, S.A. et al., Princeton University Press,1950.
Guttman	2. Guttman, L.L. The Basis for Scalogram Analysis. In op cit.
Ludlow & Hillocks (1985)	3. Ludlow, L.H. & Hillocks, Jr., G. (1985). Psychometric Considerations in the Analysis of Reading Skill Hierarchies. *Journal of Experimental Education, 54,* 15-21.

4. Item Response Theory:

Basics - item and test characteristic curves, the information function, one-parameter dichotomous /rating scale/partial credit models.

WebCT Link Name	**Measurement Theory**
Jaeger (1987)	**Uncon. v. Con. Max. Like.** **PROX to UCON via N-R**
Ludlow & Haley (1999)	
Lord (1980) Chapter 2	
Reeve: Intro to Modern	

Required Readings:

1. Chapter 15 of Crocker & Algina.
2. Jaeger, R.M. (1987). Two decades of revolution in educational measurement!? *Educational Measurement: Issues and Practice*, 6-14.
3. Ludlow, L.H. & Haley, K.C. (1999). Newton: The pinball wizard?. *Popular Measurement*, 2, 5-7.
4. Hambleton, et al Ch 1-4.
5. Bond Ch 1-2
6. Lord, F.M. *Applications of Item Response Theory to Practical Testing Problems*. Ch 2, LEA, 1980
7. Reeve, B.B. An Introduction to modern measurement theory. Division of cancer control and population sciences, National Cancer Institute.
8. Ludlow, L.H. Unconditional versus conditional maximum likelihood. Lecture notes.
9. Ludlow, L.H. PROX to UCON via Newton-Raphson. Lecture notes.

Suggested Overview Readings-Past/Present/Future:

1. Bock, R.D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, 21-32.

2. Fischer, G.H. & Molenaar, I.W. (Eds.) (1995). *Rasch Models: Foundations, recent developments, and applications*. NY: Springer. (**see Ch 1**).

3. Hambleton, R. (1989). Principles and selected applications of item response theory. In Linn, R.L. (Ed.). *Educational Measurement*. (3rd ed). NCME, AERA: McMillan.

4. Mislevy, R.J. (1987). Recent developments in item response theory with implications for teacher certification. In *Review of Research in Education*, Rothkopf, E.F. (Ed.) Vol. 14, Washington: AERA.

5. Mislevy, R.L. (1996). Test theory reconceived. *Journal of Educational Measurement*, 379-416.

6. Reckase, M.D. The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25-36.

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

8. van der Linden, W. & Hambleton, R. (Eds.). (1997). *Handbook of Modern Item Response Theory*. NY: Springer. (**see Ch 1**).

9. Yamamoto, K. & Kulick, E. (1999).Scaling Methodology and Procedures for the TIMSS Mathematics and Science Scales. In *TIMSS Technical Report*, Ch. 14.

10. Item Response Theory web site:

General Measurement Articles:

1. Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath & S. H. Lovibond (Eds.), *Mathematical and Theoretical Systems: Proceedings of the 24th International Congress of Psychology of the International Union of Psychological Science, Vol. 4* (pp. 7-16). North-Holland: Elsevier Science Publishers.

2. Andrich, D. (1996). Measurement criteria for choosing among models with graded responses. In *Categorical variables in developmental research: Methods of analysis* (pp. 3-35). Academic Press, Inc.

3. Fisher, W. P., Jr. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice. Vol. II* (pp. 36-72). Norwood, New Jersey: Ablex Publishing Corporation.

4. Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*, 398-407.

5. Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355-383.

6. Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3(1),* 281-288.

7. Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know*. Hillsdale, NJ: LEA.

5. The Rasch Model:

Purpose, assumptions, estimation procedures, item and person fit, residual analysis, applications. Computer output interpretation of TASC and TAMP
data sets.

WebCT Link Name **Ludlow & Haley (1995)**

Wright (1980) Forward **Ludlow & O'Leary (1999)**

Wright (1980) Afterward **Details of Rasch Model Estimation**

IRT TASC output **Required Readings:**

1. Chapters 1-5 of Wright & Stone.

2. Wright, B.D. (1980). "Foreward", and "Afterward". In Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press.

3. Ludlow, L.H. & Haley, S.M. (1995). Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement, 55,* 967-975.

4. Ludlow, L.H. & O'Leary, M. (1999). Omitted and not reached items: Practical data analysis implications. *Educational and Psychological Measurement, 59, 615-630.*

5. Ludlow, L.H. Details of Rasch Model Estimation. Lecture notes.

6. Ludlow, L.H. IRT TASC output.

Suggested Readings:

1. Wright, B.D. (1967). Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton: Educational Testing Service, 85-101.

2. Whitely, S.E. & Davis, R.V. (1974). The nature of objectivity with the Rasch model. *Journal of Educational Measurement,* 163-178.

3. Andrich, D. (1978). Relationships between the Thurstone and Rasch Approaches to Item Scaling. *Applied Psychological Measurement*, 449-460.

4. Englehard, G. (1984). Thorndike, Thurstone, and Rasch: A Comparison of their methods of scaling psychological and educational tests. *Applied Psychological Measurement*, 21-38.

5. Brink. N. (1972). Rasch's logistic model vs. The Guttman model. *Educational and Psychological Measurement*, 32, 921-927.

6. Hambleton, R. & Jones, R. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice*, 38-47.

7. Gable, R., Ludlow, L., Wolf, M. (1990). The use of classical and Rasch latent trait models to enhance the validity of affective measures. *Educational and Psychological Measurement*, 50, 869-878.

8. McNamara, T. (1996). Raters and ratings: Introduction to multi-faceted measurement. Concepts and procedures in Rasch measurement. Ch 5 & 6 in *Measuring Second Language Performance*. London: Longman.

9. Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, 42, 7-16.

10. Drasgow, F. & Lissak, R.I. (1983). Modified parallel analysis: A procedure for examining latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, 3, 363-373.

Two related articles:

1. Leonard, M. (1980). Rasch promises: A layman's guide to the Rasch method of item analysis. *Educational Researcher, 22,* 188-192.

2. Willmont, A. (1980). What does Rasch promise? A reply to Rasch promises by Martin Leonard. *Educational Researcher, 22,* 193-197.

Five related articles:

1. Divgi, D.R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23,* 283-298.

2. Henning, G.(1989). Does the Rasch model work for multiple-choice items? Take another look: A response to Divgi. *Journal of Educational Measurement, 26,* 91-97.

3. Andrich, D. (1989). Statistical reasoning in psychometric models and educational measurement. *Journal of Educational Measurement, 26,* 81-90.

4. Goldstein, H. (1979). Consequences of using the Rasch model for educational assessment. *British Educational Research Journal, 5,* 211-220.

5. Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33,* 234-246.

General other-discipline articles:

1. Alphen A., Halfens, R., Hasman, A., & Imbos, T. (1994). Likert or Rasch? Nothing is more applicable than good theory. *Journal of Advanced Nursing, 20,* 196-201.

2. Spray, J. (1987). Recent developments in measurement and possible applications to the measurement of psychomotor behavior. *Research Quarterly for Exercise and Sport, 58,* 203-209.

4. Massof, R.W. (2002). The measurement of vision disability. *Optometry and Vision Science,* 79(8), 516-552.

Related Books:

1. Fischer, G.H. & Molenaar, I.W. *Rasch Models: Foundations, Recent Developments, and Applications*. NY: Springer, 1995.

2.Wilson, M. (ed). *Objective Measurement: Theory Into Practice*. Volume 1-5. Norwood, NJ: Ablex, 1992-2000.

3. Wright, B.D. & Stone, M.H. *Best Test Design*. Chicago: MESA Press, 1979.

Other:

Any issue of Rasch Measurement: Transactions of the Rasch Measurement Special Interest Group. (see me for their location)

Variable Development and Application Examples:

1. Hillocks, Jr. G. & Ludlow, L.H. (1984). A taxonomy of skills in reading and interpreting fiction. *American Educational Research Journal,* 7-24.

2. Ludlow, L.H. (1985). A strategy for the graphical representation of Rasch model residuals. *Educational and Psychological Measurement, 45*, 851-860.

3. Ludlow, L.H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement, 10*, 217-229.

4. Ludlow, L.H. & Hwang, R. (1990). Evaluating district-level performance relative to the system. *Educational Research Quarterly, 14*, 29-37.

5. Ludlow, L.H. & Guida, F.V. (1992). The Test Anxiety Scale for Children as a Measure of academic anxiety. *Educational and Psychological Measurement, 51*, 1013-1021.

6. Ludlow, L.H. & Lunz, M. (1998). The Job Responsibilities Scale: Invariance in a longitudinal prospective study. *Journal of Outcome Measurement*, 2, 326-337.

7. Ludlow, L.H. (1998). Scale invariance from a three-dimensional graphical perspective: Visualizing an eigenvector. *Educational and Psychological Measurement*, 58, 166-178.

8. Ludlow, L.H. (1999). The structure of the Job Responsibilities Scale: A multi-method analysis. *Educational and Psychological Measurement, 59, 962-975.*

9. Coster, W.J., Mancini, M.C. & Ludlow, L.H. **(**1999)**.** Factor structure of the School Function Assessment. *Educational and Psychological Measurement, 59, 665-677.*

10. Coster, W., Ludlow, L.H. & Mancini,M. (1999). Using IRT variable maps to enrich understanding of rehabilitation data. *Journal of Outcome Measurement*, 3, 123-133.

*11. Ludlow,L.H. & Mahalik,J.R. (2001). Congruence between a Theoretical Continuum of Masculinity and the Rasch Model: Examining the Conformity to Masculine Norms Inventory. *Journal of Applied Measurement, 2,* 205-221.

*12. Haley SM, Coster WJ, Andres PL, Ludlow LH, Ni P, Bond TLY, Sinclair SJ & Jette AM. (2004). Activity Outcome Measurement for Post-acute Care. *Medical Care*, 42, 49-61.

*13. Coster WJ, Haley SM, Andres PL, Ludlow LH, Bond TLY & Ni P. (2004). Refining the conceptual basis for rehabilitation outcome measurement: Personal Care and Instrumental items. *Medical Care*, 42. 62-72.

TAMP/PEDI Projects:

1. Gans & Haley, et al. (1988). Description and interobserver reliability of the Tufts Assessment of Motor Performance. *American Journal of Physical Medicine and Rehabilitation, 2,* 202-210.

2. Haley & Ludlow, et al. (1991). Tufts Assessment of Motor Performance: An empirical approach to identifying motor performance categories, *Archives of Physical Medicine and Rehabilitation, 72*, 359-366.

3. Ludlow & Haley. (1991). Polytomous Rasch models for behavioral assessment: The Tufts Assessment of Motor Performance. In *Objective Measurement,* Vol. 1, Wilson, M. (Ed.) Ablex.

4. Ludlow, Haley & Gans. (1992). A hierarchical model of functional performance in rehabilitation medicine: The Tufts Assessment of Motor Performance. *Evaluation and the Health Professions, 15*, 59-74.

5. Haley & Ludlow. (1992). Applicability of the hierarchical scales of the Tufts Assessment of Motor Performance for school-aged children and adults with disabilities. *Physical Therapy, 72*, 191-206.

6. Fisher, A.G., Bryze, K.A., Granger, C.V., Haley, S.M., Hamilton, B.B., Heineman, 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 assessment. *International Journal of Educational Research, 21*, 579-593.

7. Haley, S.M., Ludlow, L.H. & Coster, W.J. (1993). Pediatric Evaluation of Disability Inventory: Clinical Interpretation of summary scores using Rasch rating scale methodology. *Physical Medicine and Rehabilitation Clinics of North America: New Developments in Functional Assessment, 4*, 529-540.

8. Ludlow, L.H. & Haley, S.H. (1996). Effect of context in rating of mobility activities in children with disabilities. *Educational and Psychological Measurement, 56*, 122-129.

6. Operation of Psychometric computer programs:

SCALE, WINSTEPS, RUMM, PARSCALE, BILOG, QUEST/CONQUEST.

7. The Two-and Three-parameter IRT Models:

Purpose, assumptions, estimation, model fit, applications.

Baker, F.B. (1992). *Item Response Theory: Parameter Estimation Techniques.* NY: Marcel Dekker.

Hambleton, R.K. (Ed) (1983). *Applications of Item Response Theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*, Boston: Nijhoff.

Hambleton, R.K., Swaminathan, H. & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Sage.

Harris, D. (1989). Comparison of 1-,2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*. NCME Instructional Module, Spring, 35-41.

Hulin, C.L., Drasgow, F. & Parsons, C.K. (1983). *Item Response Theory: Application to Psychological Measurement*. Homewood, IL: Dow Jones- Irwin.

Lord, F.M. (1983). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

Van der Linden, W. & Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory.* NY: Springer.

Wainer, H. & Messick, S. (1983*). Principals of Modern Psychological Measurement*. Hillsdale, NJ: Erlbaum.

8. Technical Applications of IRT:

Item banking, adaptive testing, item and test bias, equating, test construction, differential item functioning (DIF), scale anchoring, cut-scores, plausible values.

<u>Differential Item Functioning</u>:

Berk, R.A. (Ed.) (1982). *Handbook of Methods for Detecting Test Bias*. Baltimore, MD: Johns Hopkins University Press.

Camilli, G. & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, NJ: Sage.

Clauser, B.E. & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17, 31-44.

Crane, P.K., van Belle, G. & Larson, E.B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine*, 23, 241-256.

Engelhard, G. Jr., Hansche, L. & Rutledge, K.E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education*, 4, 347-360.

French, A.W. & Miller, T.R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, 3, 315-332.

Hambleton, R.K. & Jones, R.W. (1994). Comparison of empirical and judgmental procedures for

detecting differential item functioning. *Educational Research Quarterly*, 1, 21-36.
Holland, P.W. & Wainer. H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.
Jodoin, M.G. & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 4, 329-349.
Miller, T.R. & Spray, J.A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, 2, 107-122.
Millsap, R.E. & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 4, 297-334.
Rogers, H.J. & Swaminathan, H. (1993). A comparison of the logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, 2, 105-116.
Swaminathan, H. & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 4, 361-370.
Teresi, J.A., Kleinman, M. & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine*, 19(11-12), 1651-1683.
Welch, C. & Hoover, H.D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education*, 1, 1-19.
Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.
Zwick, R. & Thayer, D.T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics*, 3, 187-201.

Computerized Adaptive Testing:

Sands, W.A., Waters, B.K. & McBride, J.R. (Eds). (1997). *Computerized Adaptive Testing: From Inquiry to Operation*. Washington, DC: APA.
Wainer, H., Dorans, N.J., Flauger, R., Green, B.F., Mislevy, R., Steinberg, L. & Thissen, D. (1990). *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Erlbaum.
Weiss, D. (Ed.). *New Horizons in Testing*.

Equating:

Angoff, W.H. (1984). *Scales, Norms, and Equivalent Scores*. Princeton: ETS.
Holland, P.W. & Rubin, D.B. (1982). *Test Equating*. NY: Academic Press.
Kolen, M.J. & Brennan, R.L. (1995). *Test Equating: Methods and Practices*. NY: Springer.
Linn, R.L. & Kiplinger, V.L. (1995). Linking statewide tests to the NAEP: Stability of results. *Applied Measurement in Education, 8*, 135-155.
Mislevy, R.J., Sheehan, K.M. & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30*, 55-78.

Cut-scores:

Berk, R.A. (1986). A consumer's guide to setting performance standards

on criterion-referenced tests. *Review of Educational Research*, 56,137-172.
Glass, G.V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237-261.
Jaeger, R.M. (1989). Certification of Student Competence. In R.L.Linn (Ed.), *Educational measurement* (3rd ed., pp 485-514). New York: American Council on Education and Macmillan.
Kane, M. (1994). Validating the performance standards associated with cutscores. *Review of Educational Research*, 64, 425-461.

Thurstone Bibliography

Thurstone, L.L. A method for scoring tests. *Psychological Bulletin*. 16, 1919, 235-240.

Thurstone, L.L. The anticipatory aspect of consciousness. *The Journal of Philosophy Psychology and Scientific and Scientific Methods*. 16, 1919, 561-568.

Thurstone, L.L. The learning curve equation. *Psychological Monographs*. 26, 1919, No. 114, p. 51.

Thurstone, L.L. The stimulus-response fallacy in psychology. *Psychological Review*. 30, 1923, 354-369.

Thurstone, L.L. A method of scaling psychological and educational tests. *Journal of Educational Psychology*. 16, 1925, 433-451.

Thurstone, L.L. The mental age concept. *Psychological Review*, 33, 1926, 268-278.

Thurstone, L.L. The scoring of individual performance. *Journal of Educational Psychology*. 18, 1927, 505-524.

Thurstone, L.L. The unit of measurement in educational scales. *Journal of Educational Psychology*. 18, 1927, 505-524.

Thurstone, L.L. A mental unit of measurement. *Psychological Review*. 34, 1927, 415-423.

Thurstone, L.L. The measurement of opinion. *Journal of Abnormal and Social Psychology*. 22, 1928, 415-430.

Thurstone, L.L. Attitudes can be measured. *American Journal of Sociology*, 33, 1928, 529-554.

Thurstone, L.L. The absolute zero in intelligence measurement. *Psychological Review*. 35, 1928, 175-197.

Thurstone, L.L. Fechner's Law and The Method of Equal Appearing Intervals. *Journal of Experimental Psychological*. 12, 1929, 214-224.

Thurstone, L.L. and Chave, E.J. *The Measurement of Attitude*. University of Chicago Press, 1929.

Thurstone, L.L. Theory of attitude measurement. *Psychological Review*. 36, 1929, 222-241.

Thurstone, L.L. *The Reliability and Validity of Tests*. Ann Arbor, Mich.: Edwards Brother. 1931.

Thurstone, L.L. The measurement of psychological value. In Smith, T.V. and W.K. Wright, (Eds). *Essays in Philosophy*. LaSalle, Illinois: The Open Court Publishing Co. 1929, 157-174.

Thurstone, L.L. The calibration of test items. *The American Psychologist*. 21, No. 3, 1947, 103-104.

Thurstone, L.L. The measurement of values. *Psychological Review*. 61, 1954, 47-58.

Thurstone, L.L. *The Measurement of Values*. University of Chicago Press, 1959, (Midway Reprint, 1974).

General References on Science and Measurement

Adams, E.W. On the nature and purpose of measurement. *Synthese*, 16, 1966, 125-169.

Allen, M.F. & Yen, W.M. *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole, 1979.

Berk, R.A. (Ed.) *Criteria-Referenced Measurement: The State of the Art*. Baltimore: Johns Hopkins Press, 1980.

Boring, E.G. *A History of Experimental Psychology*. NJ: Prentice-Hall, 1950.

Boring, E.G. *History, Psychology, and Science*. NY: Wiley, 1963.

Brannigan, A. *The Social Basis of Scientific Discoveries*. Cambridge, GB: Cambridge University Press, 1981.

Bridgman, P.W. L*ogic of Modern Physics*. NY: MacMillan Co. 1932.

Brown, W. & Thomson, G.H. *The Essentials of Mental Measuremen*t. Cambridge, GB: Cambridge University Press, 1921.

Butterfield, H. *The Origins of Modern Science*. NY: The Free Press, 1957.

Campbell, N.R. *Foundations of Science*. NY: Dover Publications, 1919.

Campbell, N.R. Symposium: Measurement and Its importance for Philosophy. *Aristolian Society*, 17, 1938, 121-142.

Carnap, R. *Philosophical Foundations of Physics*. Basic Books, 1966.

Cattell, J. Mental tests and measurements. *Mind*, 15, 373-381, 1890.

Churchman, C.W. & Ratoosh, P. *Measurement: Definitions and Theories*.

NY: Wiley, 1959.

Conant, J.B. *Science and Common Science*. Yale University Press, 1951.

Coombs, C.H. Thurstone's Measurement of Social Values Revisited 40 Years Later. *Journal of Abnormal and Social Psychology*, 1965, 2, 145-170.

Crocker, L. & Algina, J. *Introduction to Classical & Modern Test Theory*, NY: Holt, Rinehart & Winston, 1986.

Cronbach, L.J. Further evidence on response sets and test design. *Educational and Psychological Measurement*, 1950, 10, 3-31

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

DeGruijter, D.N.M. & von der Kamp, L.J. *Advances in Psychological and Educational Measurement*. NY: Wiley, 1976.

Edwards, A.L. *Techniques of Attitude Scale Construction*. NY: Appleton-Century-Crofts, Inc., 1957.

Edwards, A.L. & Thurstone, L.L. An internal consistency check for scale values determined by the method of successive integers. *Psychometrika*, 1952, 17, 169-180.

Ellis, Brian. *Basic Concepts of Measurement*. Cambridge, GB: Cambridge University Press, 1966.

Embretson, S. (Ed.). *Test Design: Developments in Psychology and Psychometrics*. NY: Academic Press, 1985.

Embretson, S.E. & Hershberger, S.L. (Eds.). *The New Rules of Measurement: What Every Psychologist and Educator Should Know*. LEA, 1999.

Galton, F. *Natural Inheritance*. NY:MacMillan, 1894.

Galilei, G. *Dialogue Concerning the Two Chief World Systems-Ptolemaic & Copernican*. Translated by Drake, S., Foreward by Einstein, A. (2nd ed.). London: University of California Press, 1967.

Ghiselli, E. *Theory of Psychological Measurement*. NY: McGraw-Hill, 1964.

Glaser, R. A Methodological Analysis of the Inconsistency of Responses to Test Items. *Educational and Psychological Measurement*, 9, 1949, 727-739.

Glaser, R. The Application of the Concepts of Multiple Operation Measurement to the Response Patterns on Psychological Tests. *Educational and Psychological Measurement*,

11, 1951, 372-382.

Guilford, J.P. (1954). *Psychometric Methods*. NY: McGraw-Hill.

Gullicksen, H. (1950). *Theory of Mental Tests*. NY: Wiley.

Gulliksen, H. & Messick S. *Psychological Scaling: Theory and Applications*. NY: Wiley, 1960.

Hanson, N.R. (1979). *Patterns of Discovery*. Cambridge, GB: Cambridge University Press.

Hawking, S. (1988). *A Brief History of Time*. NY: Bantam.

Jensen, A.R. (1980). *Bias in Mental Testing*. NY: Free Press.

Kaplan, A. *The Conduct of Inquiry: Methodology for Behavioral Science*. NY: Chandler, 1964.

Kirk, G.S. & Raven, J.E. *The Presocratic Philosophers*. Cambridge, GB: Cambridge University Press, 1962.

Kelley, T.L. (1928). *Crossroads in the Mind of Man: A Study of Differentiable Mental Abilities*. Stanford: Stanford University Press.

Koestler, A. *The Sleepwalkers: A History of Man's Changing Vision of the Universe*. BY: MacMillan, 1959.

Kuhn, T.S. (1957). *The Copernican Revolution*. Cambridge, MA: Harvard University Press.

Kuhn, T.S. (1962). *Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Kuhn, T.S. (1977). *The Essential Tension*. Chicago, IL: University of Chicago Press.

Kyburg, Henry, K. (1970). *Probability and Inductive Logic*. NY: Macmillan.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, No. 140.

Lindquist, E.F. (Ed). *Educational Measurement* (1st Ed.). Washington, DC: American Council on Education, 1951.

Linn, R.L. (Ed.) *Educational Measurement* (3rd Ed.). NY: Macmillan, 1989.

Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, 3, 635-694.

Lord, F.M. *A Theory of Test Scores*. Psychometric Monograph Number 7, 1952.

Lord, F.M. & Novick, M.R. *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley, 1968.

Luce, Bush, & Galanter. *Handbook of Mathematical Psychology*. NY: Wiley, 1963.

Luce, R.D. & Tukey, J.W. Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement. *Journal of Mathematical Psychology*, 1964, 1, 1-27.

Lumsden, J. Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, 1978, 31, 19-26.

Lumsden, J. Variations on a theme by Thurstone. *Applied Psychological Measurement*, 1980, 4, 1-7.

Lumsden, J. Person reliability. *Applied Psychological Measurement*, 1977, 1, 477-482.

Lumsden, J. Test theory. *Annual Review of Psychology*, 1976, 27.

Maranell, G.M. (ED) *Scaling*. Chicago, IL: Aldine, 1974.

McDonald, R.M. *Test Theory*. LEA, 1999.

Messick, S. The standards problem. *American Psychologist*, October, 1975.

Messick, S. Validity. Chapter 1 in Linn, R. (Ed) *Educational Measurement* (3rd Ed.). NY: Macmillan, 1989.

Mosier, C. Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review*. 47, 1940, 355-366.

Mosier, C. Psychophysics and mental test theory: The constant process. *Psychological Review*, 48, 1941, 235-249.

Nunnally, J. *Psychometric Theory*. NY: McGraw-Hill, (1936, 1967).

Pearson, K. Regression, heredity and panmixia. *Philosophical Transactions*, 187A, 253-318, 1896.

Pearson, K. *The Grammar of Science*. London: A&C Black, 1890.

Pearson, K. (1914-1930). *The Life, Letters and Labours of Francis Galton*. Vol I, II, IIIA, IIIB. London: Cambridge University Press.

Pierce, C.S. *Values In A Universe of Chance*. NY: Dover.

Pierce, C.S. *Essays In The Philosophy of Science*. Vincent Tomas (ed.), NY: The Liberal Arts Press, 1959.

Popper, K.R. *The Logic of Scientific Discovery*. Harper Torchbooks, 1968.

Rozeboom, W.W. *Scaling Theory and Measurement*. Synthese, 16, 1966, 172-233.

Spearman, C. (1904b). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15*, 201-292.

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

Spearman, C. Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271-295, 1910.

Spearman, C. *The Abilities of Man: Their Nature and Measurement*. NY: Macmillan, 1927.

Spearman, C. & Jones, LL. W. (1951). *Human Ability: A Continuation of the Abilities of Man*. London: MacMillan.

Stevens, S.S. On the theory of scales of measurement, *Science*, 103, 1946, 677-680.

Thissen, D. & Wainer, H. (2001) (Ed.). *Test Scoring*. LEA.

Thomson, G.H. (1939). *The Factorial Analysis of Abilit*y. Boston: Houghton Mifflin.

Thorndike, E.L. (1907). Empirical studies in the theory of measurement. *Archives of Psychology*, Vol. XV, No. 3, 1-45.

Thorndike, E.L. *Handwriting*. Teacher's College Record, 11, 2, 1910.

Thorndike, E.L. *Theory of Mental and Social Measurement*. NY: Teachers College, Columbia University, 1916

Thorndike, E. et al. *The Measurement of Intelligence*. NY: Teachers College, Columbia University, 1925.

Thorndike, R.L. (Ed.) *Educational Measurement* (2nd Ed.). Washington, DC: American Council on Education, 1971.

Thorndike, R.L. *Applied Psychometrics*. Boston, MA: Houghton Mifflin, 1982.

Torgerson, W.S. *Theory and Methods of Scaling*. NY: Wiley, 1958.

Urban, F.M. *The Application of Statistical Methods to The Problems of Psychophysics*. Philadelphia, PA: The Psychological Clinic Press, 1908.

Urban, F.M. The Weber-Fechner Law and Mental Measurement. *Journal of Experimental Psychology*, 1933, 16, 221-238.

Van der Kamp, L.J., Langerak, W.F. & de Gruijter, D.N.M. *Psychometrics for Educational Debates*. NY: Wiley, 1980.

Van der Linden, W.J. & Hambleton, R.K. (Eds.). *Handbook of Modern Item Response Theory*. Springer, 1997.

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

Weiss, D. (Ed) *New Horizons in Testing*. NY: Academic Press, 1983.

Woolf, H. (Ed) *Quantification.* NY: Bobbs-Merrill Co., 1961.

Yoakum C.S. & Yorkers, R.M. *Army Mental Tests*. NY: Holt, 1920.

Rasch Bibliography (in my files)

Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danmarks Paedogogiske Institut, 1960 (Chicago: University of Chicago Press, 1980).

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, 19, 49-57.

Rasch, G. On Specific Objectivity: An Attempt At Formalizing the Request for Generality and Validity of Scientific Statements. *Danish Yearbook of Philosophy*, 1977, 14, 58-94.

Rasch, G. Objective Comparisons. Lectures given at the UNESCO Seminar, Oslo, 1964.

Rasch, G. An Individual-Centered Approach to Item Analysis with Two Categories of Answers (undated mimeograph).

Rasch, G. The Poisson Process as a Model for a Diversity of Behavioral Phenomena. Washington, DC: International Congress of Psychology, 1963.

Rasch, G. An Informal Report of the Present State of a Theory of Objectivity in Comparisons. In Proceedings of the NUFFIC International Summer Session in Science at "Het Oude Hof". L.J. von der Kamp and C.A.J. Viek, eds., Leiden, 1967.

Rasch, G. An Individualistic Approach to Item Analysis. In P.F. Lazarsfeld and N.W. Henry (Eds) *Readings in Mathematical Social Science*, p 84-108. Chicago: Science Research Associates, 1966.

Wright Bibliography

Wright, B.D. and Panchapakesan, N.A. A procedure for sample free item analysis. *Educational and Psychological Measurement*. 1969, 29, 23-48.

Wright, B.D. & Masters, G.N. *Rating Scale Analysis*, Chicago: MESA Press, 1982.

Wright, B.D. & Douglas, G.A. Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement*, 37, 1977, 573-586.

Perline, R., Wright, B.D. & Wainer, H. The Rasch model as additive conjoint measurement. *Research Memorandum Number 24*, The University of Chicago, Department of Education, Statistical Laboratory, 1977.

Wright, B.D. & Bell, S.R. Verifying the unconditional estimation procedure for Rasch item analysis with simulated data. *Research Memorandum Number 26*, The University of Chicago, Department of Education, Statistical Laboratory, 1977.

Wright, B.D. & Mead, R.J. The use of measurement models in the definition and application of social science variables. *Army Research Institute Report*, DAHC19-76-G-0011, 1977.

Wright, B.D. & Douglas, G.A. Best test design and self-tailored testing. *Research Memorandum Number 19*, The University of Chicago, Department of Education, Statistical Laboratory, 1975.

Wright, B.D. The Rasch Model for Test Construction and Person Measurement. Paper prepared for Fifth Annual Conference and Exhibition on Measurement and Evaluation, Office of the Los Angeles County Superintendent of Schools, 1978.

Wright, B.D. & Masters, G.N. The measurement of knowledge and attitude. *Research Memorandum Number 30*, The University of Chicago, Department of Education, Statistical Laboratory, 1980.

Masters, G.N. & Wright, B.D. A model for partial credit scoring. *Research Memorandum Number 31*, The University of Chicago, Department of Education, Statistical Laboratory, 1981.

Wright, B.D., Mead, R.J. & Bell, S.R. BICAL: Calibrating items with the Rasch model. *Research Memorandum Number 23*, The University of Chicago, Department of Education, Statistical Laboratory, 1980.

Wright, B.D. Solving Measurement problems with the Rasch model. *Journal of Educational Measurement*, 14, 1977, 97-116.

Wright, B.D. & Douglas, G.A. Rasch item analysis by hand. *Research Memorandum*

Number 21, The University of Chicago, Department of Education, Statistical Laboratory, 1976.

Wright, B.D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service, 1968.

Wright, B.D. Afterward in G. *Rasch Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press, 1980.

Wright, B.D. & Douglas, G.A. Best procedures for sample-free item analysis. *Applied Psychological Measurement*, 1, 1977, 281-294.

Guttman Bibliography (in my files)

Guttman, L.A. The Quantification of a Class of Attributes: A theory and Method for Scale Construction. In P. Hurst et al., *The Prediction of Personal Adjustment* (p 319-348). NY: Social Science Research Council, 1941.

Guttman, L.A. On Festinger's Evaluation of Scale Analysis. *Psychological Bulletin*, 44, 1947, 451-465.

Guttman, L.A. The Basis for Scalogram Analysis. In S.A. Stouffer et al., *Measurement and Prediction*. NY: Wiley, 1950.

Guttman, L.A. Questions and Answers about Scale Analysis. Research Branch, Information and Education Division, Army Service Forces, Report D-2, 1945.

Guttman, L. The Cornell Technique for Scale and Intensity Analysis. *Educational and Psychological Measurement*, 7, 1947, 247-279.

Guttman, L. A Basis for Scaling Qualitative Data. *American Sociological Review*, 9, 1944, 139-150.

Relevant Dissertations (Chapters with Rasch or IRT model descriptions)

Turner, J. (2003). *Examining an Art Portfolio Assessment Using a Many-Facet Rasch Measurement Model*. Boston College, Lynch School of Education, ERME.

Kennedy, A. (2003). *Interpreting the Progress in International Reading Literacy Study (PIRLS) Scale*. Boston College, Lynch School of Education, ERME.

Gregory, K. (2000). *The Two Parameter Multidimensional Random Coefficients Multinomial Logit Model*. Boston College, Lynch School of Education, ERME.

Yu, Yueh-Hsia. (2000). *The Development of a Functional Math Inventory in the Rasch*

Framework: An Integrated Approach for Constructing Instruments for Assessing Students With Severe Disabilities. Boston College, Lynch School of Education, ERME.

Haley, K. (1999). *Watkins-Farnum Revisited: Application of the Rasch Model to Measures of Musical Performance*. Boston College, Lynch School of Education, ERME.

Kelly, D. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) Achievement Scales Using Scale Anchoring*. Boston College, Lynch School of Education, ERME.

Gonzalez, E.J. (1994). *Estimating IRT Ability Parameters for extreme Scores When Using Maximum Likelihood Estimation Procedures*. Boston College, Lynch School of Education, ERME.

Williams, F. (1994). *Comparison of Classical Measurement Theory, Rasch and Linear Logistic Trait Models*. Boston College, Lynch School of Education, ERME.

Kalinowski, T. (1992). *The Latent Structure of Pain Intensity*. Clark University, Department Psychology.

Pierce, G. (1992). *A Comparison of the Identification of "Depth" as Used in Item Classification by the ICCP to the "Difficulty" of Those Items as Measured Using the Rasch Model Analysis*. Boston College, Lynch School of Education, ERME.

Leibowitz, S. (1990). *Measuring Change in Sensitivity to Ethical Issues in Computer Use*. Boston College, Lynch School of Education, ERME.

Hickey, B.L. (1987). *Comparison of Rasch and Three-Parameter Logistic Model Residuals*. Boston College, Lynch School of Education, ERME.

Ludlow, L.H. (1983). *The Analysis of Rasch Model Residuals*. University of Chicago, Department of Education, MESA.

Mead, R. (1976). *Assessment of Fit of Data to the Rasch Model Through Analysis of Residuals*. University of Chicago, Department of Education, MESA.

Special Edition Journals

Journal of Educational Measurement, Vol. 14, Part 2, 1977

Journal of Educational Measurement, Vol. 20, Part 2, 1983.
("Linking Achievement Testing to Cognitive Processes")

Journal of Educational Measurement, Vol. 21, Part 4, 1984.
("Issues in Item Banking")

Applied Psychological Measurement, Vol. 6, Part 4, 1982.
("Advances in Item Response Theory and Applications")

Applied Psychological Measurement, Dec. 1996.
("Multidimensional IRT")

Medical Care. Vol 42, Number 1 supplement, January 2004. ("International Conference on Objective Measurement: Applications of Rasch Analysis in Health Care").

New Directions for Testing and Measurement. (a series of quarterly publications on various current measurement models, testing issues, and applications of techniques-published by Jossey-Bass).

Proceedings: Invitational Conference on Testing Problems. (yearly publication of invited presentations at ETS-published by ETS).

Relevant Journals

Applied Measurement in Education
Applied Psychological Measurement
Educational & Psychological Measurement
Educational Measurement: Issues and Practice
Educational Researcher
Journal of Abnormal & Social Psychology
Journal of Educational Measurement
Journal of Educational Psychology

Journal of Behavioral (formerly Educational) Statistics
Journal of Experimental Education
Journal of Mathematical Psychology
Journal of Outcome Measurement
Popular Measurement
Psychological Bulletin
Psychological Review
Psychometrika
Rasch Transactions (AERA SIG: Rasch Measurement)

Web Sites

Finally, check these interesting IRT-related web sites:
http://www.rasch.org/
http://work.psych.uiuc.edu/irt/tutorial.asp
http://www.acer.edu.au
(this site has excellent large-scale examples of Rasch applications)

And this list of relevant sources was compiled by former ERME student Steve Stemler

There are several bibliographies referenced from www.rasch.org/rmt/index.htm - See second panel.
For the Facets model, see www.winsteps.com/facetman/references.htm
There is a great list of Rasch references on the Journal of Applied Measurement web page (http://www.jampress.org/) under their "Guidelines for Manuscripts" link.
See the bibliography in "Applying the Rasch Model" by Trevor Bond and Christine Fox: www.erlbaum.com

Classical True-Score Theory Assignment
(Spring 2004-100 points)

Upon your data set, use SPSS procedures to perform a classical test theory (CTT) item analysis. Provide an answer to all of the following questions. An outline format is preferable. There is no need to try to write the assignment as a mini-publication at this point.

1. **Instrument and sample**:

Explain the purpose of your measurement instrument. What does the instrument purport to measure? Who developed it (wrote the items)? How many items are included? What is the scoring format? How many response options are provided? Is it a speeded test? How long does it take to answer? Is it a standardized or non-standardized instrument? Is it primarily for norm-referenced or criterion-referenced purposes?

Where did your sample come from? Who collected the data? How many subjects are there? Are they a subset of a larger study and, if so, briefly explain why they were specifically chosen. Are there any special characteristics about them? What is the population to whom they are generalizable?

2. Measurement model:

I. *Technical material*

a). Explain the statistical form of the CTT model and where it came from. Present the relevant expressions and explain them.

b). What are the primary assumptions of CTT? Present the expressions and explain them. What are potential problems that might violate them?

c). Why is the concept of parallel tests important for CTT—what is the theoretical problem that parallel tests solve?

d). Show how reliability can be expressed as the ratio of two variances.

e). What is the purpose for disattenuating a correlation—why are correlations attenuated? Show how to correct them and explain why they can lead to illogical results

f). Show why test reliability tends to increase with test length.

II. *From your data*

Report the following statistics and show how they were computed (what equations led to the statistics):

a). item difficulty (for dichotomous or rating scale data and what are generally accepted preferred difficulty ranges,

b) discrimination (use the corrected item-total correlation and explain why it is corrected and what are generally accepted preferred discrimination ranges),
c) reliability (split-half, KR-20, Cronbach alpha and when you might use one over the other), and
d) standard error of measurement (pick one form of reliability estimate and explain why you chose it and show how the SEM is used). Explain the various components of the equations.
e) explain the general purpose of a common factor analysis when it is applied to items of a test (how many factors do you hypothesize for your data?). Briefly explain what an eigenvalue and eigenvector are. Explain what the factor loadings are and what magnitudes we would like. What is a communality and what magnitudes would we like? What are some of the general ways to determine the number of factors in your data? What is the purpose of the scree plot? Why are solutions usually rotated and, in general, explain the difference between an orthogonal and an oblique rotation and why we might prefer one over the other. Explain the general principle of the varimax technique—are other techniques possible? Explain what the KMO, Bartlett's sphericity, and |R| are and why they are often general procedures computed to determine the appropriateness of factoring a correlation matrix.

3. **Analysis**:

a. Discuss the distributional characteristics of your item difficulties and person total scores, e.g., are they as intended, are they surprising? Discuss whether your discrimination estimates are reasonable or not. Are there any particular items with statistical problems (what are the statistical problems)? What might have caused the problems, if there are any? Should any items be removed or revised? Interpret the reliability coefficients you obtained.

b. Discuss the results of your initial factor analysis (was your correlation matrix "appropriate" for factoring?) and how you subsequently decided on the number of final factors to retain. What percent of variance was extracted by those factors and what is your opinion of the magnitude of the percent that was accounted for? Was the rotated and plotted final solution interpretable (just plot the first two factors)? What verbal labels did you apply to "name" the factors (and explain why you applied those names)? Was your solution expected or surprising (did you have any idea about what might result from the factor analysis)?

c. What is the reliability of each of the final factors in your solution? How many scores for each tested person would you recommend should be reported?

d. Summarize the quality of your data and what you would do next (e.g. leave it as is or modify the instrument or get a different sample, etc).

4. **Submit** your write-up and output. A useful way to write your analysis is to cut and paste into it the appropriate tables/graphs/figures that are output by SPSS rather than referring the reader to the pages of your output. (NOTE: pay attention to typo's and notation errors.)

Item Response Theory Assignment
(Spring 2004: 100 points)

Upon your data set, use SCALE/ WINSTEPS/ RUMM/ PARSCALE to perform an item response theory analysis.

1. **Instrument and sample**:

Explain the purpose of your measurement instrument. Who developed it? How many items are included? What is the scoring format? How many response options are provided? Is it a speeded test? How long does it take to answer? Where did your sample come from? How many subjects are there? Are there any special characteristics about them? Basically, I want you to remind me of the characteristics of the data used for the classical analysis.

For your data, what is the "variable" that is being measured? That is, what is the hypothesized structure that is to be tested by the Rasch model?

2. **Measurement model details**:

Explain the statistical components of the Rasch model. Why is it called a one-parameter model when clearly there is a parameter for both persons and items? What are the primary assumptions of the model?

Explain how the initial PROX person ability estimates and item difficulty estimates are computed. Why are persons and items with perfect correct or zero scores removed from analysis? What does the term "sufficient statistic" refer to? Explain (**in your own words**) what person and item "logits" are. How is the "expected" value for a person on any item computed?

Explain the general difference between a conditional and an unconditional estimation procedure. What does the term "maximum likelihood" refer to? Explain, in general basic terms, how the Newton-Raphson algorithm operates. What is its function?

How are person and item weighted fit statistics computed? Explain how person and item positive and negative fit statistics may be interpreted. What might be done if an item or person is considered to misfit the model?

3. **Analysis**:

Discuss the initial distributional characteristics of your item difficulties (does it appear to be a relatively easy or hard instrument) and person abilities (do they appear relatively capable or not). Are these findings as intended? For your data, what do "difficulty" and "ability" translate into?

Are there any particular persons with statistical problems? What might have caused them if there are? Are there any particular items with statistical problems? What might have caused them is there are? (How are you defining a "problem" and what have you done to try to locate their source?)

Explain what the "variable map" is and what it reveals about your data. Was your solution expected or surprising? What modifications, if any, would you suggest if the instrument were to be revised and re-administered?

Finally, compare and contrast the Rasch results to your previous classical analysis results. For example, is there any additional insight you have gained about your data? In addition, how does the standard error of measurement associated with a person's performance differ between the two models?

4. **Submit** your write-up and output. A useful way to write your analysis is to cut and paste into it the appropriate tables/graphs/figures that are output by the software rather than referring the reader to the pages of your output. (NOTE: pay attention to typo's and notation errors.)