International Outcome Measurement Conference

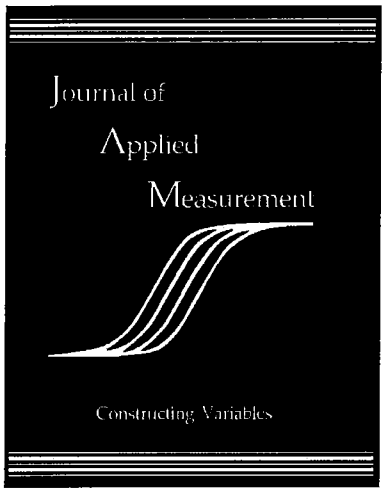
IOMC 2015

April 21 and 22, 2015

Chicago, IL

MetaMetrics®

LINKING ASSESSMENT WITH INSTRUCTION



International Outcome Measurement Conference

IOMC 2015 Welcome

Thank you for joining us for IOMC 2015. The International Outcome Measurement Conference began in May/June 1996, at the International House, University of Chicago, sponsored by Rehabilitation Foundation, Inc. and MESA Press. Presentations from that conference formed the primary content for *Physical Medicine and Rehabilitation: State of the Art Reviews* (V11, N2, 1997). There are five presenters and several participants from that conference in attendance today. The second IOMC was held in May 1998, at the International House, University of Chicago, sponsored by the *Journal of Outcome Measurement* and MESA Psychometric Laboratory. There are six presenters and several participants from that conference in attendance today. Presentations from that conference formed the primary content for later issues of the *Journal of Outcome Measurement* in Volumes 4 and 5. All of the issues of JOM are available as pdfs on the JAM/JAM Press website (www.jampress.org). The third IOMC was held at the University of Illinois at Chicago in June 2000, sponsored by the National Institutes of Health. Presentations from that conference formed the primary content for *Rasch Measurement in Health Sciences*, edited by Nikolaus Bezruczko and published by JAM Press in 2005. Ben Wright was the primary force behind this series of conferences. When his health was severely compromised in 2001, the conferences never continued.

In early 2014 Nick Bezruczko and I began an ongoing conversation about the current state of health outcome measurement and, in particular, the methods adopted by the NIH-funded PROMIS measures. The NIH PROMIS website describes these measures as follows:

PROMIS® tools measure what patients are able to do and how they feel. These tools are in the form of questionnaires that are asked in interviews, given to patients in written form, by computer, and by PDAs. The answers patients give are accumulated into reliable reports that can be used by patients and physicians to improve communication and manage diseases by finding the most appropriate treatment plan. PROMIS® tools also allow physicians to better understand how various treatments might impact what patients are able to do and the symptoms they experience.

Those conversations lead us to consider a return to our roots and a desire to reconvene the IOMC series these 15 years later. With the generous support of Jack Stenner and MetaMetrics, and *Journal of Applied Measurement* and JAM Press we welcome you to the latest edition of IOMC and welcome you to re-explore our roots in the quest for numbers, not numerals, that describe health outcomes.

Richard Smith and Nick Bezruczko
Conference Organizers

International Outcome Measurement Conference

IOMC 2015 Conference Schedule

April 21 and 22, 2015

Hampton Inn by Hilton, 160 E. Huron St., Chicago, IL

Day 1

7:30 am – 8:00 am	**Registration - Continental Breakfast** – 7th Floor
8:00 am – 8:15 am	**Welcome** (Pennsylvania)
8:15 am – 9:45 am	**Symposium 1**– Plenary session (Pennsylvania) Health care outcome metrology: *"Build it and they will come"* (Part I) William P. Fisher, Jr. - Building the field of dreams: The unrealized scientific and economic power of health care outcome metrology Jeremy Hobart - Alternate paths to successful clinical outcome measurement Stefan Cano - Individual vs group level measurement: Implications for health care outcome economics Laurie Burke - On systematically measuring the right things well enough vs locally measuring the wrong things really well William P. Fisher, Jr. - Moderator
9:45 am – 10:00 am	**Break**
10:00 am – 11:00 am	**Paper session 2A** (Pennsylvania) Nikolaus Bezruczko, Teresa Stanley, Maureen Battle, Cynthia Latty, and Shu-Pi Chen - Pathological consequences of evaluating simulation caregiver training with nonlinear, ordinal ratings: Measuring caregiver response to tracheostomy emergencies Stefanie A. Wind - Evaluating the quality of multiple-choice assessments with Mokken scale analysis Skye Barbic - Moderator **Paper session 2B** (Independence) Alexandra Rouquette, Jean-Benoit Hardouin, and Joël Coste - Differential Item Functioning (DIF) and subsequent bias in group comparisons using a multi-item scale: A simulation study James J. Thompson - What are you measuring? Dimensionality and reliability analysis of ability and speed in medical school didactic examinations Richard Smith - Moderator
11:00 am – 11:15 am	**Break**
11:15 am – 12:15 pm	**Paper session 3A** (Pennsylvania) Chih-Ying Li and Craig A. Velozo - Using Rasch analysis to generate Medicare G-Code Modifiers and develop a treatment framework in the attention domain Craig Velozo , Leigh Lehman, Ickpyo Hong, and Chih-Ying Li - Use of Rasch analysis to generate G-Code Modifiers for CMS outpatient reimbursement Nikolaus Bezruczko - Moderator

Paper session 3B (Independence)
Matthew W. Grady, Haiqin Chen, Chien-Lin Yang, and David Waldschmidt - Comparing the Rasch and Rasch testlet model for a health care licensure examination
Haiqin Chen, Matthew W. Grady, Chien-Lin Yang, and David Waldschmidt - Application of the multilevel Rasch testlet model for dual local dependence to empirical data in the health care field
Michael R. Peabody - Moderator

12:15 pm – 1:30 pm **Lunch Break**

1:30 pm – 3:00 pm **Paper session 4A** (Pennsylvania)
Ngadiman Djaja, Monika Janda, Catherine Olsen, and David Whiteman - Diagnostic discrimination of the Skin Cancer Risk (SCR) scale: Application of item response theory
Sherri L. LaVela, Sara Locatelli, Carol Kostovich, and Megan Gosch - Developing the Respirator Comfort, Wearing Experience, and Function Instrument using Rasch partial credit model analysis
Benjamin Fox - The application of Rasch measurement theory to dementia research
Stefan Cano - Moderator

Paper session 4B (Independence)
Nick Marosszeky - Not a fan of Fan (1998)! Item response theory and classical test theory: An empirical comparison of their item / person statistics
Peter Hagell and Albert Westergren - Sample size and statistical conclusions from tests of fit to the Rasch measurement model according to the RUMM2030 program
Robert Furter - Test speededness: Collecting evidence to support form length decisions
Richard Smith - Moderator

3:00 pm – 3:30 pm **Break**

3:30 pm – 5:00 pm **Paper session 5A1** (Pennsylvania)
Melissa Hofmann - Assessment of acute trauma exposure response for FIRE-EMS personnel
Jane Summer and William P. Fisher, Jr. - Ontological midwifery of caring in nursing: Practical measures for management
Jonathan Comins - Forward-backward translation vs. dual-panel translation methods for adaption of patient-reported outcome measures - how can we determine which is best?
Peter Hagell - Moderator

Paper session 5A2 (Independence)
Brett Berg, Karen Atler, and Anne G. Fisher - Constructing a health outcome measure of occupational experience: An application of Rasch measurement methods
Chris Wera - Development of a brief screening measure for depression and problem drinking
Ickpyo Hong, Annie N. Simpson, Chih-Ying Li, and Craig A. Velozo - Development of an upper extremity function measurement model
Craig Velozo - Moderator

5:00 pm **End of Day 1**

6:30 pm – End **Optional Conference Dinner**

Dao Thai Restaurant
230 E. Ohio St., Chicago, IL
Menu located at: http://www.daothai.com/pdf/Dao_menu_dinner140721.pdf

Day 2

7:30 am – 8:00 am **Registration - Continental Breakfast** – 7th Floor

8:00 am – 8:05 am **Welcome** (Pennsylvania)

8:05 am – 9:45 am **Symposium 1 (continued)** – Plenary session (Pennsylvania)
Health care outcome metrology: *"Build it and they will come"* (Part II)

Jack Stenner - Reading measurement in education as a model metrology network for health care
Rob Cavanagh - An unmodern perspective on the role of educational measurement in globalization
Robert Massof - Lions Low Vision Rehabilitation Network (LOVRNET): A system that uses outcome measures for quality improvement through continuous professional
Maureen K. Powers, William P. Fisher, Jr., Robert Massof, and Mark Wilson - Integrating visual symptoms and visual skills to model and measure functional binocular vision
William P. Fisher, Jr. - Moderator

9:45 am – 10:00 am **Break**

10:00 am – 11:00 am **Paper session 7A** (Pennsylvania)
Myriam Blanchin, Elodie De Bock, Gildas Kubis, Tanguy Le Néel, Véronique Sébille, and Jean-Benoit Hardouin - Rasch and CTT-based approaches for joint analysis of group and time effects of longitudinal Patient Reported Outcomes: impact of informative and non-informative missing data
Véronique Sébille, Myriam Blanchin, Alice Guilleux, Mohand-Larbi Feddag, and Jean-Benoit Hardouin - Methods of power and sample size determination of clinical studies based on Rasch measurement
Jeremy Hobart - Moderator

Paper session 7B (Independence)
Michael R. Peabody, Kelly D. Bradley, and Melba Custer - Assessing the validity of a continuum-of-care survey: A Rasch measurement approach
Sarah Thomas, Karen M. Schmidt, Monica Erbacher, and Cindy Bergeman - Sliding scales and changing rulers: Anchoring the longitudinal measurement of positive affect
Nikolaus Bezruczko - Moderator

11:00 am – 11:15 am **Break**

11:15 am – 12:15 pm **Paper session 8A** (Pennsylvania)
Carol T. Kostovich, Beyza Aksu Dünya, Lee A. Schmidt, and Eileen G. Collins - A Rasch rating scale analysis of the Presence of Nursing Scale-RN
Christophe Chénier, Gilles Raîche, Céline Gélinas, Nadine Talbot, and Bianca Carignan - Rasch analysis of the Critical-Care Pain Observation Tool (CPOT)
Skye Barbic - Moderator

Paper session 8B (Independence)
Trudy Mallinson - Addressing response dependence in repeated measures of rehabilitation outcomes in the unidimensional Rasch model

Robert Massof and Judith Goldstein - The role of item filtering in measures of low vision rehabilitation outcomes
Richard Smith - Moderator

12:15 pm – 1:30 pm **Lunch Break**

1:30 pm – 3:00 pm **Paper session 9A** (Pennsylvania)
Deborah M. Rooney, Bruce L. Tai, Oren Sagher, Albert J. Shih, and Luis Savastano - Validation of performance measures from a Novel Ventriculostomy Simulator using *Standards* framework
Thomas Salzberger and Stefan Cano - Investigating a lack of discrimination between two adjacent response categories in the Rasch model for ordered categories in health measurement
Curt Hagquist - Do 7 items provide as good measurement as 13 items? A comparison of a short and long version of Kidscreen
Peter Hagell - Moderator

Paper session 9B (Independence)
Paula Petry and William Fisher - Applying published instrument development research in a workshop evaluation: Practical use of Rasch-calibrated instruments with small samples
Ying Du - Investigating knowledge growth during pediatric residency training using Rasch and linear mixed models
Trudy Mallinson - Moderator

3:00 pm – 3:15 pm **Break**

3:15 pm – 4:45 pm **Symposium 3** – Plenary session (Pennsylvania)
The relationship between measurement models and scale properties
Gunnar Grimby - On the treatment of ordinal scale data in rehabilitation medicine research
Nikolaus Bezruczko - Standards and practices to guide health outcomes measurement: A strategy to avoid measurement malpractice
Richard Smith, Lee McKenna, and Christie Plackner - A comparison of the information available in various measurement models
Richard Smith - Moderator

4:45 pm – 5:00 pm **Closing Remarks**

5:00 pm **End of Day 2**

All symposia will be plenary sessions. All paper sessions will be concurrent sessions, A and B in adjacent rooms.

All **A** paper sessions highlight clinical presentations and all **B** paper sessions highlight methodological issues.

Paper presentations will allow 20 minutes per presentation, with 20-30 minutes for discussion depending on the number of presentations.

Revised April 8, 2015

International Outcome Measurement Conference

IOMC 2015 Conference Abstracts

Day 1

8:15 am – 9:45 am **Symposium 1**– Plenary session (Pennsylvania)

Health Care Outcome Metrology (Part 1): "Build It and They Will Come"
William P. Fisher, Jr., Organizer

Health care research and practice tend to define measurement in terms of ordinal, instrument- and sample-dependent units. The contrast of these changeable units—usually of unknown magnitude—with the linear and invariant unit standards developed in the natural sciences is rationalized on the grounds that the concrete objects available in physics and chemistry provide epistemological and ontological advantages unavailable for more abstract psychological and social objects. Mathematical proofs, inspired teaching, readily available software, simplified methods, and empirically obtained and theoretically explained invariant units in thousands of research investigations have neither dispelled these preconceptions nor led to new metrological standards. This symposium proposes an alternative approach—the "Field of Dreams" approach, which asserts "Build It and They Will Come"—to improving the quality of outcome measurement in health care. The distinguishing characteristic of this approach is its focus on beginning from the idea of a standard unit that is translated into terms facilitating the realization of multiple allied social sectors' scientific and economic interests. The presentations begin from a statement of principles, work through several examples of standards in process, and conclude with reflections on FDA regulatory policy development regarding the use of patient-reported outcomes (PROs) in labeling and advertisements.

William P. Fisher, Jr.
Building the Field of Dreams: The Unrealized Scientific and Economic Power of Health Care Outcome Metrology

Following Einstein's insight that major problems cannot be solved from within the frame of reference that provoked them, a profoundly different way of thinking and acting is required to initiate a new paradigm of measurement in health care, psychology, and the social sciences. The thesis defended here is that the "fatal conceit" of the modern Cartesian presumption of an independent subject (Hayek, 1948) prevents formulation of the concepts, methods, and tools needed for broad scale improvements in the quality of psychological and social measurement. An alternative unmodern (Dewey, 2012) or amodern (Latour, 1990, 1993) frame of reference offers a fundamentally different basis for thinking about and doing measurement. The primary distinguishing characteristics of this alternative include its sense of knowledge as technology, the lack of a central authority over the use and development of language, its recognition of end users as having little or no understanding of how language and technology work, its acceptance of genuine method as a playful captivation in the flow of mutually implicated subjects and objects, and its focus on the wide distribution of standardized tools as providing the language unifying fields of research and practice. Thus, rather than continue waiting indefinitely for the modern project to arrive at its perpetually deferred fulfillment in an complete picture of the objective world, we should instead define the terrain, the equipment, and the rules, roles, and responsibilities of teams and players in the language game of measurement.

Jeremy Hobart
Alternate Paths to Successful Clinical Outcomes Measurement

In recent work, it has become evident that clinical outcome measurement still has much to achieve if it is to maximize its potential as a cornerstone of therapeutic trials, policy and practice. Needed next steps to ensure successful clinical outcome measurement in the future include: 1) testable psychometric theories; 2) testable construct theories, and 3) an experimental hypothesis-testing approach. We consider in this presentation whether these steps can effectively be implemented on a wide scale only via methods of

persuasion and regulation, or if more compelling practical arguments can be effected via scientific and economic consensus standards alliances.

Stefan Cano
Individual vs. Group Level Measurement: Implications for Health Care Outcome Economics

Tarde (1903), whose economic theories have lately been rediscovered and newly appreciated, understood that quantification in psychology and economics would have to be based on something more than what can be easily counted (Barry and Thrift, 2007, p. 516; Tarde, 1903, p. 107), and that accomplishing it would be a superlative cultural achievement (Latour, 2010). Tarde projected "a statistics penetrated by interpsychological spirit" (1902, pp. 141-142; cited in Hughes, 1961, p. 558) provoking the question as to whether Rasch's probabilistic models for individual-level measurement might indicate a productive path in the direction suggested by Tarde (Fisher, 2014). For this path to prove worth exploring, two preliminary methodological conditions have to be met. One involves the alignment of Rasch measurement practice with the definition of the central theoretical problem of the social sciences, tracing the imitative patterns of individuals whose decisions and behaviors are informed by incomplete knowledge. The second involves the evaluation of within-individual variation as a basis for inter-individual comparisons. An example illustrating this difference is drawn from recent research in neurorehabilitation.

Laurie Burke
On Systematically Measuring the Right Things Well Enough vs. Locally Measuring the Wrong Things Really Well

"It is often much worse to have good measurement of the wrong thing—especially when, as is so often the case, the wrong thing will in fact be used as an indicator of the right thing—than to have poor measurement of the right thing" (Tukey, 1979, p. 486). The papers in this session all speak to the importance of measurement as a systematically constructed shared language distributed throughout communities of research and practice. Despite the fact that all models are wrong (Rasch, 1973/2011; Box, 1979), just as the laws of physics lie (Cartwright, 1983) and are unrealistic (Butterfield, 1957; Heidegger, 1967), models and laws are nonetheless eminently useful for coordinating and aligning practical decision processes. Recent work at the FDA on systematically measuring the right things will be reviewed in this context.

10:00 am – 11:00 am **Paper session 2A** (Pennsylvania)

Nikolaus Bezruczko, Teresa Stanley, Maureen Battle, Cynthia Latty, and Shu-Pi Chen
Pathological consequences of evaluating simulation caregiver training with nonlinear, ordinal ratings: Measuring caregiver response to tracheostomy emergencies

This proposal describes research being conducted to compare hospital and special simulation training on caregivers' confidence to care for infants assisted with medical technology. Central questions in this presentation are the effects of measurement properties, ordinal versus linear, on evaluation of caregiver confidence after simulation training, homecare quality, and consequences of measurement properties on children's hospital readmission, morbidity, and mortality.

Stefanie A. Wind
Evaluating the Quality of Multiple-Choice Assessments with Mokken Scale Analysis

The concept of invariant measurement for social science research is typically associated with Rasch measurement theory (Rasch, 1960/1980; e.g., Engelhard, 2013). Mokken (1971) recognized the value of invariance for measurement in the social sciences. However, he expressed some concern with the parametric transformation upon which the Rasch model is based within the context of social science research. Motivated by these concerns, Mokken proposed a nonparametric procedure for evaluating the quality of social science measurement that can be viewed as a nonparametric analogue to the Rasch model. Mokken's nonparametric scaling procedure can be used to evaluate the quality of dichotomous and polytomous items in terms of useful

psychometric properties, including scalability, monotonicity, and invariance. In previous research, Mokken scaling has been applied as a systematic framework for developing affective scales within the context of political science, public health, economics, and public health. However, the use of Mokken scaling to examine the properties of multiple-choice (MC) items in education has not yet been fully explored. A nonparametric approach to evaluating MC items is promising in that this approach facilitates the investigation of assessment instruments in terms of useful psychometric properties without imposing potentially inappropriate requirements related to the level of measurement obtained from these instruments. The purpose of this study is to demonstrate the use of Mokken's nonparametric scaling procedure to examine the psychometric properties of MC items in education. Data from an eighth-grade physical science assessment are used to illustrate and explore Mokken-based techniques for evaluating the quality of MC items, using Rasch-based indices of measurement quality as a frame of reference. Implications for research, theory, and practice are discussed.

10:00 am – 11:00 am **Paper session 2B** (Independence)

Alexandra Rouquette, Jean-Benoit Hardouin, and Joël Coste
Differential Item Functioning (DIF) and subsequent bias in group comparisons using a multi-item scale: A simulation study

Objective: To determine the conditions in which the estimation of a difference between groups for a construct evaluated using a multi-item scale is biased if the presence of Differential Item Functioning (DIF) in the group under study is not taken into account.

Methods: Datasets were generated using the Partial Credit Model to simulate various realistic conditions. Seven factors potentially affecting the bias on the estimation of difference between groups were considered: sample size, true difference between groups, number of items in the scale size, proportion of items with DIF, DIF size on these items, level of difficulty of these items and type of DIF (uniform or non-uniform)

Results: For 121 620 (22.5%) of the 540 000 simulated datasets analyzed, all factors considered were found to be significantly associated with the mean bias when studied all together in a MANOVA model. The mean bias was lower for higher sample size, number of items or true difference between groups whereas it was higher if the proportion of items with DIF or the DIF size were higher.

Conclusion: This simulation study allowed the measurement bias resulting from DIF to be quantified in various realistic conditions of multi-item scale use.

James J. Thompson
What are you measuring? Dimensionality and reliability analysis of ability and speed in medical school didactic examinations

Summative evaluation of an educational program typically involves examinations, often in multiple-choice format. In order for these exams to be valid, they must have some relationship to the topic tested (dimensionality) and be sufficiently reproducible between persons (reliability) to justify student ranking. Evaluation of dimensionality is difficult and is complicated by the classic observation that didactic performance involves a generalized component ("G") in addition to subtest specific factors. Bifactor analysis explicitly addresses these concerns. I have analyzed 183 students over two academic years in 13 courses with 44 exams and 3352 multiple choice questions for both accuracy and speed. G comprises most explained variance over persons, exams, and courses (62-89%) thereby rendering the person analyses as "essentially unidimensional". Person speed was consistently "more unidimensional" (higher omegaH and explained common variance) than person accuracy. Reliability at the person, exam, and course levels was uniformly acceptable (>0.95 for Cronbach's alpha and McDonald's omegaT). Effect sizes with respect to content domains (approximated by course exams) or course disciplines (e.g. anatomy, biochemistry, etc.) are small. Thus, application of conventional response analysis methods, e.g Rasch modeling, is warranted.

11:15 am – 12:15 pm **Paper session 3A** (Pennsylvania)

Chih-Ying Li and Craig A. Velozo
Using Rasch analysis to generate Medicare G-Code Modifiers and develop a treatment framework in the attention domain

Objective: This study demonstrates a practical application in generating CMS G-Code Modifiers in the domain of attention for outpatient healthcare reimbursement, and an effective treatment framework for individuals with traumatic brain injury (TBI).

Methods: The ICF-AM Attention domain consists of 52 self-report items measuring impact of cognitive deficits on daily functioning. Ninety individuals with TBI completed the ICF-AM with inclusion criteria of moderate/ severe TBI, 18-85 years old, no previous psychiatric disorders and English speaking. Forty-seven people were recruited from outpatient rehabilitation centers and 43 people were recruited 1 or more years after TBI. Winsteps 3.75 was used to generate item-person maps to define G-Codes Modifiers with task analysis techniques. Keyforms were used to show person response patterns.

Results: The 52 items divided participants into 5 statistically distinct strata. We proposed these 5 strata to be directly connected to the 5 out of 7 G-Code Modifiers. Keyforms showed distinct ability patterns that could be used for treatment plans and goal setting.

Conclusions: The ICF-AM Attention score connects to G-Code attention Modifiers for CMS reporting. Keyforms identify patient abilities and limitations for individualized treatment planning.

Important to Practice/Science: This study provides an evidence-based approach for CMS reporting and daily practice decision making.

Craig Velozo , Leigh Lehman, Ickpyo Hong, and Chih-Ying Li
Use of Rasch analysis to generate G-Code Modifiers for CMS outpatient reimbursement

Purpose: To provide a Rasch-based method for generating G-Code modifiers (percent of impairment) (e.g., 0%, 1-19%, 20-39%. 40-59% and 60-79%, 80-99% and 100%).

Participants: Secondary analysis of 960 patients treated for upper extremity injuries from the Focus on Therapeutic Outcomes (FOTO) database.

Measure: Gross Motor Measure derived from the Symptom and Disability subscale of the Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire

Methods: The Rasch analysis rating scale model was applied to the 13-item Gross Motor Measure of the DASH to generate a person separation measure. Person strata were used to create five strata on person-item maps.

Results: Person separation was 3.7 and person separation reliability was 0.93 resulting in 5.3 person strata. Classifications with thresholds three standard errors apart were plotted on person-item maps representing five G-Code modifier classifications of impairment (0-19%, 20-39%, 40-59%, 60-79%, and 80-100%).

Conclusions: Rasch-based methodologies can be used to generate G-Code modifiers that are empirically supported and conceptually connected to levels of impairment.

Paper session 3B (Independence)

Matthew W. Grady, Haiqin Chen, Chien-Lin Yang, and David Waldschmidt
Comparing the Rasch and Rasch testlet model for a health care licensure examination

In healthcare licensure test contexts, testlets are attractive because they require test-takers to demonstrate a broad rather than atomistic understanding of the same sets of patient-related stimuli they will likely encounter as healthcare providers. Despite their attractiveness, studies have shown that test-takers' responses to testlet items sometimes show local item dependence (LID; Bradlow et al., 1999; Wainer and Wang, 2001; Yen, 1993). This study used real data from a healthcare licensure examination that contains testlets to determine the extent to which the testlet-based items showed LID and to determine the impact of the LID on test-taker ability estimates and their standard errors. The study results indicated that, while the testlet-based

items did show LID, the impact of the LID on test-taker ability estimates is quite small. The effect of the LID on standard errors of ability estimates was also small and in the expected direction.

Haiqin Chen, Matthew W. Grady, Chien-Lin Yang, and David Waldschmidt
Application of the multilevel Rasch testlet model for dual local dependence to empirical data in the health care field

The standard Rasch model is commonly used, but can lead to biased parameter estimates if the local independence assumption is violated. In this study, a multilevel Rasch testlet model is used to account for the coexistence of local item and person dependence. The performance of Rasch, multilevel Rasch, Rasch testlet, and multilevel Rasch testlet models are compared under the generalized linear mixed model framework. Results show that the multilevel Rasch testlet model performs best among four models. Local item dependence did not affect parameter estimation. However, ability estimation was affected by person dependence.

1:30 pm – 3:00 pm **Paper session 4A** (Pennsylvania)

Ngadiman Djaja, Monika Janda, Catherine Olsen, and David Whiteman
Diagnostic discrimination of the Skin Cancer Risk (SCR) scale: Application of item response theory

Aims: Queensland, Australia has the world's highest incidence of skin cancer. Self-administered scales are commonly used to measure risk factors such as phenotype, sun exposure and sun protection, or overall skin cancer risk (SCR). We sought to develop new scales for measuring skin cancer risk and calibrate it using PCM.

Subjects: Prospective skin cancer risk cohort of 43794 men and women aged 40–69 years randomly sampled from the population of Queensland, Australia.

Analysis: Dimensionality of the scale and calibration of items were studied using the partial credit model. Receiver operating characteristics (ROC) curves analyses were used to assess how well the final items predicted future development of skin cancer.

Results: Four of twenty nine items had mean square values outside acceptable boundaries, indicating item misfit. Item calibration found that item measures between -2.800 and +1.950 logits on the SCR scale. Diagnostic discrimination showed area under the curve (AUC) statistics of .753 ($p < .000$), .530 ($p < .000$) and .487 ($p=0.093$), for the phenotype (PE), sun exposure (SE) and sun protection (SP) subscales, respectively.

Conclusion: The results show unidimensional structure of each SCR subscale. Item calibration shows they are distributed along the continuum. Only the PE subscale shows good predictive discrimination.

Sherri L. LaVela, Sara Locatelli, Carol Kostovich, and Megan Gosch
Developing the Respirator Comfort, Wearing Experience, and Function Instrument using Rasch partial credit model analysis

The objective of this study was to develop and validate a measure of comfort and tolerability of filtering face-piece respirators (FFRS). Items were developed through literature reviews and focus groups healthcare workers (HCWs). The draft instrument was completed by 12 HCWs, who reviewed item clarity and relevance; an additional 12 HCWs ranked items in order of comfort and tolerability. The final instrument was completed by 165 HCWs, and data were analyzed using Rasch partial credit model analysis. Items were removed if they 1) violated the assumption of independence, 2) were misfitting, or 3) were deemed not relevant. Pivot anchoring was used to specify the threshold defining item difficulty; in our analyses, this was the point that participants moved from possessing none of the trait to some of the trait. Category function analysis demonstrated that all categories progress monotonically. Principal components analysis (PCA) demonstrated the existence of three subscales (Discomfort, General Wearing Experience, and Function). Final reliability analyses showed that the scale had moderate to high person reliability and high item reliability. The final instrument contains 21 items. This measure is ideal for assessing comfort and tolerability of current FFRs, as well as in developing the next wave of FFRs.

Benjamin Fox
The application of Rasch measurement theory to dementia research

Physical function declines with advanced ageing. However, no comprehensive psychometric assessment has been completed in dementia specific population samples. Furthermore, Rasch measurement theory is not readily used nor understood in psychomotor fields. The aim of this paper is to describe the methodology of pilot work in the assessment of common measures of physical function for older adults with dementia under a Rasch framework. While some preference the use of activities of daily living scales, physical function is the underlying, basic qualities required to remain independent as the population ages. The three measures are the Short Physical Performance Battery, the Berg Balance Scale and the Performance Orientated Mobility Assessment. Unidimensionality and model fit will be assessed and a variable map of items and persons will be created. The future direction of this work is to develop a dementia specific item bank of measures of physical function for use in residential and community based aged care assessments.

1:30 pm – 3:00 pm **Paper session 4B** (Independence)

Nick Marosszeky
Not a fan of Fan (1998)! Item response theory and classical test theory: An empirical comparison of their item / person statistics

Streiner and Norman (2003) "*Health Measurement Scales: A practical guide to their development and use*", now in its fourth edition, is one of the foundational texts of the health outcomes movement. It states, that the differences between scales constructed with IRT and CCT are trivial." This statement is representative of the view about the equivalence of classical and item response theory techniques. This view is widely held and has been one factor in limiting the use of modern psychometric techniques in the development of health outcome measures.

However, the equivalence view relies heavily on a paper by Fan (1998) which examined the item statistics derived from CTT and IRT for a large educational dataset. While subject to a number of theoretical and practical criticisms from a Rasch measurement perspective this paper has not been replicated.

This conference paper by replicating the paper by Fan (1998) challenges the finding that item difficulty indexes derived from high and low ability samples using classical test theory derived techniques are invariant. This secondary data analysis, by working through the methods used by Fan (1998) step by step, demonstrates that a reliance on correlational methods cannot be used to determine the invariance of item statistics.

Peter Hagell and Albert Westergren
Sample size and statistical conclusions from tests of fit to the Rasch measurement model according to the RUMM2030 program

Sample size is a major contributor to statistical null hypothesis testing, which is the basis for many approaches to testing Rasch model fit. To allow for taking this into account, the RUMM2030 Rasch analysis software has the ability to adjust n in the calculation of its chi-2 based fit statistics. This paper examines the effects of such post-hoc adjustments on the statistical conclusions, and explores the occurrence of type I errors with Rasch model fit statistics implemented in RUMM2030. Data simulations of Rasch model fitting 25-item dichotomous scales with sample sizes ranging from n=50-2500 were generated an analysed regarding fit with and without adjusted sample sizes corresponding to the same n values as those simulated. Results suggest that post-hoc downward sample size adjustment is a useful procedure to avoid type I errors when working with relatively large data sets ($n \geq 500$). The value of upward adjustment with small data sets is less clear, particularly regarding the total item-trait chi-2 test, which tends to falsely signal misfit. Under the assumption of Rasch model fit, our observations suggest that a sample size around 250 (up to about 500) provides a good balance for the statistical interpretation of RUMM2030 fit statistics.

Robert Furter
Test speededness: Collecting evidence to support form length decisions

Test speededness can have serious validity implications in the form of construct-irrelevant variance and content underrepresentation if the time limit set for an exam administration is not appropriate given the exam form length. This study demonstrates a body of evidence-style approach to evaluating test speededness, focusing on omitted item responses, response latencies, and examinee performance throughout the exam form from both CTT and IRT perspectives, in order to inform decision-making regarding future form lengths and time limits. Using this information, measurement professionals and testing program administrators can more consciously negotiate the balance between critical aspects of the exam form (length, coverage, exposure, reliability, etc.) and the operational constraint of testing time. In doing so, the result is a more valid test score reflecting an examinee's medical knowledge, rather than an amalgamation of the construct of interest and extraneous variables, promoting public protection and healthy outcomes. Data from the 2012 administration of a medical certification exam is used to demonstrate the body of evidence approach.

3:30 pm – 5:00 pm **Paper session 5A1** (Pennsylvania)

Melissa Hofmann
Assessment of acute trauma exposure response for FIRE-EMS personnel

Purpose. To develop an instrument that measures response to acute trauma exposure for firefighter and emergency medical service (EMS) personnel.

Methods. Data were analyzed for 97 firefighter and EMS personnel employed by a fire department in a city in the Western region of the United States. Rasch analyses assessed dimensionality, person and item reliability, scale use and function, and construct validity including person-item fit statistics.

Results. Rasch analyses showed the ATERS (Acute Trauma Exposure Response Scale) performed well with three distinct subscales. Reliability of person separation and item reliability was .81and .96 for Emotional Psyche, .66 and .95 for Coping Ability, and .70 and .97 for Support Systems. Scale use and function was appropriate for each subscale. Validity was supported by Rasch evidence through an even spread of person ability to item difficulty for each distinct subscale.

Conclusion. The ATERS performed well as a measure of acute trauma exposure response for constructs of Emotional Psyche, Coping Ability, and Support Systems with good Rasch person reliability and structure. Items were deleted for each construct following Rasch analyses due to misfit and low item-measure correlation. Further research is recommended to optimally represent each construct in regard to person-item fit.

Jane Summer and William P. Fisher, Jr.
Ontological midwifery of caring in nursing: Practical measures for management

Situating caring in a "transpersonal framework of conscious intentionality," Watson (1999, pp. 237, 243-259) notes that postmodern analyses and concepts often fail to move beyond critical evaluations of failed philosophies or methods to new ways of constructing possibilities for healing environments. This suggests that an unmodern (Dewey, 2012) or amodern (Latour, 1990, 1993) perspective on caring may open new doors to productive concepts, theories, and methods. If, following Watson, nurses are to be able to assume roles as "ontological architects" designing healing spaces, they will need a new science of ontological engineering to support them. Just as it is said that "science is measurement," so, also, "engineering is metrology," i.e., the discipline that creates and maintains invariant reference standard units. Nursing has not yet adopted to a significant degree relevant advances in measurement theory and practice made over the last several decades. A new basis for theoretical and practical developments in caring in nursing follows from close consideration of the experience of *physis*, spontaneously self-organizing natural processes. Probabilistic Rasch models making use of sufficient statistics and requiring separable parameters have been shown to be of significant value in constructing these measures.

Jonathan Comins
Forward-backward translation vs. dual-panel translation methods for adaption of patient-reported outcome measures - how can we determine which is best?

What is the best way to ensure the transfer of the original content and meaning of items in cross-cultural adaptation of patient-reported outcome measures? Forward-backward translation techniques, as described by Guillemin et al., 1993, are considered the most appropriate methods for translating patient-reported outcome measures (PROMs). However, an inherent weakness is that the translator at each step is blinded from the originator of the PROM, the patients that contributed to the original version of the PROM, and all other translators. We believe we better can ensure that we retain the meaningful content of the original version of a condition-specific Danish PROM (KNEES-ACL) by using a dual-panel bilingual translation process (Swaine-Verdier et al., 2004). We hypothesize further that this dual-panel technique will yield superior psychometric properties in the US version of KNEES. In this talk, I will present and compare both translation techniques.

3:30 pm – 5:00 pm **Paper session 5A2** (Independence)

Brett Berg, Karen Atler, and Anne G. Fisher
Constructing a health outcome measure of occupational experience: An application of Rasch measurement methods

This presentation reports the results of an application of the Rasch rating scale model to develop a health outcome measure of three different dimensions of subjective experience related to everyday activity from an existing assessment, The Daily Experiences of Pleasure, Productivity, and Restoration Profile (PPR Profile). Analyses of PPR Profiles from a sample of 263 mountain state university students in the United States focused on rating scale structure, dimensionality, and reliability of the three scales of the PPR Profile. Results indicate that, in general, rating scales are functioning reasonably well, but one rating scale category (*Extreme displeasure*) did not fit the Rasch rating scale model of the PPR Profile. Items of the PPR Profile appear to work together to form three unidimensional scales, but some item misfit requires further attention. Standard errors of measures are acceptable, but person separation may indicate that the tool did not separate the sample into as many distinct groups as anticipated. Further investigation into rating scale functioning, dimensionality, and reliability are warranted to lay the groundwork for the next steps in the development of a health outcome measure of subjective experience.

Chris Wera
Development of a brief screening measure for depression and problem drinking

This study examined the psychometric characteristics of a brief 15-item assessment measure for depression and alcohol use prior to primary care medical office visits. The measure was adapted from two widely used measures: the Patient Health Questionnaire (PHQ-4) and the Alcohol Use Disorder Identification Test (AUDIT-C). Items from the UPPS-P measure were examined in order to determine if they could enhance screening for depression and alcohol consumption.

Exploratory factor analysis identified three factors. Rasch analysis showed the PHQ-4, the AUDIT-C, and the impulsivity questions as unidimensional. Rasch analysis of AUDIT-C showed poor item fit and significant differential item functioning and was determined to inadequate as a scale, and so, individual items were used in subsequent analyses.

Hierarchical regression revealed a significant contribution of the impulsivity measure in explaining variance for the PHQ-4, but was lacking in explaining additional measure variance when used with the AUDIT-C individual items. Latent class analysis identified three classes with this scale.

The 15-items scale was unsuccessful is improving identification of problematic drinking, however the impulsivity items could be useful in helping to better identify depression among this population. The results also questioned the effectiveness of the AUDIT-C in screening for excessive alcohol consumption.

Ickpyo Hong, Annie N. Simpson, Chih-Ying Li, and Craig A. Velozo
Development of an upper extremity function measurement model

Purpose: To identify the critical measurement mechanisms underlying upper extremity function. Participants: 203 outpatients with musculoskeletal disorders. The participants' average age was 48.3±17.9 years.

Measure: The ICF-Gross Upper Measure (ICF-GUE) self-report of upper extremity function

Methods: The dependent variable for the multiple regression was the Rasch linear calibration of the ICF-GUE 27 test items. The independent variables were four task-analysis derived measures: object weight (lb), lifting distance from floor (in.), carrying (yes/no), lifting (yes/no). The independent variable contribution to the model was determined using adjusted variance.

Results: Object weight and lifting distance were the only significant independent variables in the regression model, accounting for 83% of the variance ($p < 0.01$). The regression model indicates that with one pound increased in object weight, item challenge increased by 0.16 logits ($p < 0.00$) and with one inch increased in distance lifted from floor, item challenge increased by 0.02 logits ($p < 0.03$).

Conclusions: The results indicate that the Rasch and multiple regression analyses are useful in developing a measurement model for gross upper-extremity activities. The findings suggest that the measurement model for the ICF-GUE can be explained by object weight and distance lifted from the floor.

5:00 pm **End of Day 1**

6:30 pm – End **Optional Conference Dinner**
Dao Thai Restaurant
230 E. Ohio St.
Chicago, IL
Menu located at: http://www.daothai.com/pdf/Dao_menu_dinner140721.pdf

Day 2

8:00 am – 8:05 am **Welcome**

8:05 am – 9:45 am **Symposium 1 (continued)** – Plenary session (Pennsylvania)

Health Care Outcome Metrology (Part 1): "Build It and They Will Come"
William P. Fisher, Jr., Organizer

Health care research and practice tend to define measurement in terms of ordinal, instrument- and sample-dependent units. The contrast of these changeable units—usually of unknown magnitude—with the linear and invariant unit standards developed in the natural sciences is rationalized on the grounds that the concrete objects available in physics and chemistry provide epistemological and ontological advantages unavailable for more abstract psychological and social objects. Mathematical proofs, inspired teaching, readily available software, simplified methods, and empirically obtained and theoretically explained invariant units in thousands of research investigations have neither dispelled these preconceptions nor led to new metrological standards. This symposium proposes an alternative approach—the "Field of Dreams" approach, which asserts "Build It and They Will Come"—to improving the quality of outcome measurement in health care. The distinguishing characteristic of this approach is its focus on beginning from the idea of a standard unit that is translated into terms facilitating the realization of multiple allied social sectors' scientific and economic interests. The presentations begin from a statement of principles, work through several examples of standards in process, and conclude with reflections on FDA regulatory policy development regarding the use of patient-reported outcomes (PROs) in labeling and advertisements.

Jack Stenner
Reading Measurement in Education as a Model Metrology Network for Health Care

Though there are no publicly developed and maintained unit standards for constructs measured with tests, assessments, and surveys in either education or health care, education does possess a de facto unit standard for measuring and managing reading ability, text complexity, and comprehension rates. This standard was created via a social process uniting different groups' educational and economic interests in common cause. Lessons of potential value in health care were learned in the course of understanding how to realize each group's interests more fully and efficiently via a joint effort than could be realized by acting alone. The challenges of coordinating and aligning the interests of care providers, patients, payers, advocates, researchers, investors, and others via the medium of a shared language for measuring and managing outcomes are complex. Our experience in education suggests these challenges might best be met in a context where care providers and payers are already in close relationships that can be bootstrapped into a more efficient leveraging of the information they already have at their fingertips.

Rob Cavanagh
An Unmodern Perspective on the Role of Educational Measurement in Globalization

How might the five distinguishing characteristics of the modern global environment (neo-liberalism, globalization, marketization, decentralization, and accountability and performativity) and its effects on education change when viewed from the perspective of an amodern (Latour, 1999) or unmodern (Dewey, 2012) philosophy incorporating Rasch metrology? The problem is one of understanding how to coordinate local behaviors and decisions over a variety of different kinds of decisions across wide swaths of society in an uncoerced way that respects individual rights and liberties. An unmodern, amodern or postpositivist neo-liberalism extends the application of the ideals of classical liberalism but modifies the organizing principle of the market to take advantage of contemporary Rasch measurement technologies shown over the last several decades able in theory and practice to make all essential forms of capital fungible.

Robert Massof
Lions Low Vision Rehabilitation Network (LOVRNET): A System that Uses Outcome Measures for Quality Improvement through Continuous Professional Education

The Lions Low Vision Rehabilitation Network (LOVRNET) is a demonstration project in Maryland, Delaware, and District of Columbia that trains and supports local service providers in the provision of low vision rehabilitation services in the patient's community and in the patient's home. High quality measurement of patient reported outcomes is achieved via an online item bank and system of equated visual functional assessments; these will inform continuous provider education to improve the quality and effectiveness of low vision rehabilitation services. LOVRNET is a partnership between Lions Club members in the community and community healthcare providers to assure that adequate low vision rehabilitation services are available to the community. Lions LOVRNET is supported by Lions and by private foundations, and hopes to gain the support of the low vision device industry, which stands to profit from the increased volume of service.

Maureen K. Powers, William P. Fisher, Jr., Robert Massof, and Mark Wilson
Integrating visual symptoms and visual skills to model and measure functional binocular vision

Obtaining a clear image of the world depends on good eye coordination ("binocular vision"). Yet no standard exists by which to determine a threshold for good vs poor binocular vision, as exists for the eye chart and visual acuity. We asked whether data on the signs and symptoms related to binocular vision are sufficiently consistent with children's self-reported visual symptoms to substantiate a construct model of Functional Binocular Vision (FBV), and then whether that model can be used to aggregate clinical and survey observations into a meaningful diagnostic measure. Data on visual symptoms from 1,100 children attending school in Los Angeles were obtained using the Convergence Insufficiency Symptom Survey (CISS); and for more than 300 students in that sample, 35 additional measures were taken, including acuity, cover test near and far, near point of convergence, near point of accommodation, accommodative facility, vergence ranges, tracking ability, and oral reading fluency. A preliminary analysis of data from the 15-item, 5-category CISS and 15 clinical variables from 96 grade school students who reported convergence problems (CISS scores of 16 or higher) suggests that the clinical and survey observations will be optimally combined in a multidimensional model.

10:00 am – 11:00 am **Paper session 7A** (Pennsylvania)

Myriam Blanchin, Elodie De Bock, Gildas Kubis, Tanguy Le Néel, Véronique Sébille, and Jean-Benoit Hardouin
Rasch and CTT-based approaches for joint analysis of group and time effects of longitudinal Patient Reported Outcomes: impact of informative and non-informative missing data

Patient Reported Outcomes (PRO) which are assessed through questionnaires are increasingly used in health sciences. Several approaches exist to handle such data, and the most commonly used are the Classical Test Theory (CTT) and the Rasch model. The evaluation of the evolution of the PRO is often the major concern of the study and this raises some issue regarding the choice of a method of analysis adapted to correlated data. The analysis has also to take account of missing data to which longitudinal studies are also frequently faced with. The most adequate strategy to analyze longitudinal latent variables remains to be identified. Our aim was to compare a Rasch and a CTT-based method to analyze such longitudinal PRO data through simulation studies on data prone to dropout. Both methods presented comparable and acceptable type I error and comparable power for time effect. Due to biased time effects estimations and a loss of power for group effect observed for both methods, results from analysis where data are subject to informative dropout have to be used carefully.

Véronique Sébille, Myriam Blanchin, Alice Guilleux, Mohand-Larbi Feddag, and Jean-Benoit Hardouin
Methods of power and sample size determination of clinical studies based on Rasch measurement

Despite the widespread use of patient-reported Outcomes (PRO) in clinical studies assessing quality of life, depressive symptoms, anxiety, or pain, their design remains a challenge. Justification of study size is hardly provided, especially when Rasch model is planned for analysis. The classical sample size formula for comparing

normally distributed endpoints between two groups has shown to be inadequate in this setting (underestimated study sizes). A method (Raschpower) for sample size and power computations based on Rasch models for evaluating PRO in cross-sectional and longitudinal studies has been developed. It provides the power for a given sample size during the planning stage of a study in the framework of Rasch models. Its performance has been studied in simulation studies in different settings: i) when the model used for planning cross-sectional or longitudinal studies for dichotomous or polytomous items is well-specified, ii) when some assumptions of the Rasch model are not met (normality of the latent trait, local independence of items) or misspecifications are made on the expected values of some parameters. The Raschpower method seems to be valid and relevant for power and sample size determination in cross-sectional and longitudinal studies aiming at assessing PRO data using dichotomous or polytomous items.

10:00 am – 11:00 am **Paper session 7B** (Independence)

Michael R. Peabody, Kelly D. Bradley, and Melba Custer
Assessing the validity of a continuum-of-care survey: A Rasch measurement approach

Satisfied patients are more likely to be compliant, have better outcomes, and are more likely to return to the same provider or institution for future care. The Satisfaction with a Continuum of Care survey (SCC) was designed to improve patient care using measures of patient satisfaction and facilitate a cultural shift from a "silos-of-care" to a "continuum-of-care" mentality by fostering inter-departmental communication as patients moved between environments of care at a Midwestern rehabilitation hospital. This study provides a Rasch measurement framework for investigating issues related to survey reliability and validity. The results indicate that although certain aspects of the survey seem to function in a psychometrically sound manner, the questions are too easy to endorse and provide little information to help improve patient care. Suggestions for future revisions to this survey instrument are provided.

Sarah Thomas, Karen M. Schmidt, Monica Erbacher, and Cindy Bergeman
Sliding scales and changing rulers: Anchoring the longitudinal measurement of positive affect

This study investigated three methods of conducting partial credit model (PCM) analyses in a longitudinal study of Positive Affect in older adults. Unanchored analyses were compared to two types of anchored analyses, one fixed item parameters on each day to equal the average of item parameters for three occasions and one used the average of all 56 occasions. It is important to investigate how scaling affects observed day-to-day variation in longitudinal analyses to avoid assessing measurement error due to shifting scales rather than true variation in Positive Affect. This research builds on Erbacher et al. (2012) who compared several anchoring methods to unanchored analyses and concluded that desirable results were obtained by averaging item parameters over three days. However, utilizing information from all occasions may further improve the anchoring item parameters. The anchoring parameters in the current study were obtained by conducting a PCM analysis on data from all items on all 56 occasions and then averaging the resulting item difficulties and transition locations (three or 56 days). These average item parameters were then used to fix item parameters for the analysis of each occasion. Properties of the day-to-day variation in Positive Affect under unanchored and anchored analyses are compared.

11:15 am – 12:15 pm **Paper session 8A** (Pennsylvania)

Carol T. Kostovich, Beyza Aksu Dünya, Lee A. Schmidt, and Eileen G. Collins
A Rasch rating scale analysis of the Presence of Nursing Scale-RN

Paterson and Zderad's Humanistic Nursing Theory (1976) emphasizes the emotional connection between nurse and patient, in addition to the technical skills performed by the nurse. The Presence of Nursing Scale-RN Version (PONS-RN) was developed to measure nurses' perceptions of their ability to be present to their patients. Therefore, the purpose of this study was to examine the psychometric properties of the Presence of

Nursing Scale-RN using Rasch rating scale analysis. The results of this study have shown that the thirty-one item Presence of Nursing Scale-RN yielded scores with high item and person reliability and validity.

Christophe Chénier, Gilles Raîche, Céline Gélinas, Nadine Talbot, and Bianca Carignan
Rasch analysis of the Critical-Care Pain Observation Tool (CPOT)

The ability to assess pain matters, because pain is associated with negative health outcomes (Kastrup et al., 2009). One of the most widely used instrument to assess pain in intensive care unit patients is the *Critical-Care Pain Observation Tool* (CPOT: Gélinas et al., 2009). Previous validation studies having used classical techniques, several psychometrics properties of the CPOT remain unknown. The objectives of this research were to assess the instrument's dimensionality, its construct validity and its capacity to distinguish between levels of pain by using exploratory factor analysis and Rasch modelling. A sample of 257 patients was used to collect the data. Results show that, while a unidimensional model is conceivable, concerns remain about the dimensionality of the instrument. A third of the items show misfit and about the same proportion have a difficulty level that disagrees with the clinical guidelines about the CPOT's usage. Finally, while items show a great level of reliability (0.96), persons' measures have a rather low reliability (0.67) and only 1.62 strata of pain can be distinguished. The narrow range of pain levels in the sample could explain this poor performance and further study is needed, with a sample exhibiting a wider range of pain levels.

11:15 am – 12:15 pm **Paper session 8B** (Independence)

Trudy Mallinson
Addressing response dependence in repeated measures of rehabilitation outcomes in the unidimensionsl Rasch model

Repeated measures are common in rehabilitation practice and research studies because capturing the amount of change a patient has made is an important component of rehabilitation, particularly in this era of value-based purchasing. However, including the same patients at multiple time points in the same Rasch analysis violates the assumption of independence. Yet using only baseline or only discharge values is also unacceptable. Studies using simulated data have described the impact of response dependence and reviewers critique studies using Rasch measurement that combine repeated measures in a single analysis. One solution is to randomly select patients across the time periods so that each patient appears once in the analysis but the range of patient "ability" being captured is relatively well represented. Item and step anchors from this "unbiased" run are then applied to an analysis of all the repeated measures of patients. However, examining the extent to which response dependence influences "real world" rehabilitation outcomes has not been thoroughly examined. This study examines the impact of attempting to ameliorate response dependence in two different sets of real rehabilitation patients (community living older adults; post-partum women) and compares the results to minimally clinically important measures of change.

Robert Massof and Judith Goldstein
The role of item filtering in measures of low vision rehabilitation outcomes

Patient reported outcome measures of low vision rehabilitation depend on patients' judgments of their difficulty performing specific activities before and after intervention. The measured person trait is visual ability and the measured item trait is the visual ability required to perform the activity in the item. Low vision rehabilitation has the aim of making daily activities less difficult for the patient with the use of assistive devices and/or adaptations – it does not change the patient's vision. Consequently, successful rehabilitation produces intervention-specific differential item functioning (DIF), not changes in the person's visual ability per se. If item measures are anchored to pre-rehabilitation values, then intervention-specific DIF manifests as changes in the estimated person measure. However, the magnitude of the effect of rehabilitation depends on the choice of items since rehabilitation is goal directed. If items are included that are not addressed by the plan of care, then the effect of rehabilitation will be lowered. Therefore, the calibrated item bank must be filtered to be clinically meaningful for each patient. This paper explores the effects of patient-dependent item filtering on low vision

rehabilitation outcome measures and on estimations of minimum clinically important differences in estimated individual person measures.

1:30 pm – 3:00 pm **Paper session 9A** (Pennsylvania)

Deborah M. Rooney, Bruce L. Tai, Oren Sagher, Albert J. Shih, and Luis Savastano
Validation of performance measures from a Novel Ventriculostomy Simulator using Standards framework

Background: This study extends our previous work to evaluate the *validity* evidence from the ventriculostomy simulator using a newly-developed tool, the Ventricolostomy Procedural Assessment Tool (V-PAT).

Methods: Performance data were collected from 11 novice and 3 expert neurosurgeons (n=14). Participants self-reported their ability to perform tasks on the simulator using the V-PAT, a 15-item instrument with 4-point rating scales ranging from 1 (unable to perform) to 5 (performs easily and smoothly). Performances were videotaped and rated by three neurosurgeons, using V-PAT and OSATS. We evaluated validity evidence using a many-facet Rasch model, supported by Cronbach α, ICC, and Wilcoxon signed rank tests.

Results: Validity evidence relevant to test content was supported by positive V-PAT item point-measure correlations [.39, .81] and Rasch Outfit MS values <2.0. Evidence relevant to response processes was supported by reasonable person fit (<2.0) and favorable category function indices. Evidence of internal structure was supported by high α (.95) and ICC=[.10, .93] for most V-PAT items. Overall, novices performed at a lower level than experts for both scales, $p \leq 0.05$, although statistical significance at the item-level was not achieved for the V-PAT. Positive, strong correlation between summed V-PAT and OSATS scores supported evidence relevant to relationships to other variables, r=0.72, p<.001, while rater bias indices indicated no overall differences across expert raters, p=.65, and supported evidence of consequences of testing.

Conclusion: In spite of a small sample, favorable validity evidence supports the use of the simulator for skills training and performance assessment, particularly when used with the OSATS.

Thomas Salzberger and Stefan Cano
Investigating a lack of discrimination between two adjacent response categories in the Rasch model for ordered categories in health measurement

Polytomous response formats are abundant in social measurement in general, and in health measurement in particular. Their scoring with successive integers implies properly ordered categories. Among the possible violations of this requirement, a lack of discrimination between two adjacent categories, which are perceived as being equivalent, may be the most widespread cause of improperly ordered categories. Generally, the evaluation of the response format is based on item fit statistics and the order of empirical threshold estimates. A series of simulation studies shows that, under some conditions, improperly ordered categories may neither result in item misfit nor lead to reversed thresholds. However, threshold probabilities and associated observed frequencies lend themselves as a reliable criterion to identify malfunctioning response scales. Thus, when it comes to testing the response format, we recommend their consideration in any Rasch analysis of polytomous data in addition to item fit statistics and threshold ordering.

Curt Hagquist
Do 7 items provide as good measurement as 13 items? A comparison of a short and long version of Kidscreen

The Kidscreen questionnaire is a generic instrument to measure health-related quality of life among children and adolescents, available as a short (K-27) and a long (K-52) version.
K-52 covers two dimensions of mental health, positive and negative. In K-27 items from each dimension and an additional item are included in the Psychological Well-being dimension.
The purpose of the paper is to examine whether the Psychological Well-being dimension in K-27 may be used as a replacement for the two measures of mental health included in the K-52.
Data from a nationwide survey in 2009 on adolescent mental health in Sweden, including 172 000 students in grades 6 and 9, were used to examine the three measures.

For all measures the person reliability and the item fit was good, and there were no reversed thresholds. Independent t-tests of person measures indicated lack of, or only minor, multidimensionality.
The outcomes from the analysis of the K-27 dimension of Psychological Wellbeing indicate that the two subsets of items (positive and negative) constitute a higher order interaction, which is manifested firstly by non-significant differences between the person estimates of each subset of items, secondly by high correlations between the two subsets of items.

1:30 pm – 3:00 pm **Paper session 9B** (Independence)

Paula Petry and William Fisher
Applying published instrument development research in a workshop evaluation: Practical use of Rasch-calibrated instruments with small samples

A new approach to outcomes evaluation was employed for two workshop series focused on stress reduction and joyful living conducted for health system employees and caregivers since 2012. Rasch-calibrated self-report instruments measuring depression, anxiety and stress, and the joyful living effects of mindfulness behaviors were identified in peer-reviewed journal articles. Items from one instrument were modified for use with a US population, other items were simplified, and some new items were written. Participants provided ratings of their depression, anxiety and stress, and the effects of their mindfulness behaviors before and after each workshop series. The numbers of participants providing both pre- and post-workshop data were low (16 and 14). The item calibration results show that, with some exceptions, the item hierarchies defining the constructs retained the same invariant profiles they had exhibited in the published research (correlations (not disattenuated) range from 0.85 to 0.96). Comparisons of the pre- and post-workshop measures for the three constructs showed substantively and statistically significant changes. Implications for program evaluation comparisons, quality improvement efforts, and the organization of communications concerning outcomes in clinical fields are explored.

Ying Du
Investigating knowledge growth during pediatric residency training using Rasch and linear mixed models

This study uses Rasch and linear mixed models to explore the pattern of medical knowledge growth in general pediatrics residency training. Data from residents in categorical pediatrics training programs who took the in-training examination (ITE) at the beginning of all three training years and the Board certifying examination were included in the study. The ITE is a condensed version of the certifying examination, designed to be similar in difficulty, format and content coverage. Findings indicate that knowledge growth was not linear. The growth rate decreased in the second training year, and then accelerated in the third year. Female residents' growth rate was slightly behind males in the first and second training years, but excelled in the third year and culminated in higher performance on the certifying examination. International medical school graduates maintained a similar growth rate to American medical school graduates during the first two training years, but accelerated at a lower rate by the time of the certifying examination. Future studies will investigate the curriculum and practice in each training year. Exploration of this topic will help programs improve pediatric training, which ultimately will result in better health care for children, adolescents and young adults.

3:15 pm – 4:45 pm **Symposium 3** – Plenary session (Pennsylvania)
The relationship between measurement models and scale properties
Nikolaus Bezruczko and Richard Smith, Organizers

Gunnar Grimby
On the treatment of ordinal scale data in rehabilitation medicine research

Background: There have been a number of publications during near 30 years pointing out that descriptive statistical measures, such as mean and standard deviations, are invalid whenever data are on ordinal scales as may also be summed raw scores. Changes in an observed raw score may reflect a different change in

the underlying metric range depending upon the starting point. In a recent Editorial in J Rehabil Med (2012), it was argued to end the malpractice of using raw scores from ordinal scales. Authors who report IRT (item response theory)-based estimates should justify that their estimates are on an interval scale and the use of Rasch analyses and Rasch-derived instruments was encouraged. User-friendly conversion tables for transforming raw scores to linear measure should be available.

Method: All original articles, including Brief reports and Short communications, published during 2013 in Am J Phys Med Rehabil (AJPMR), Arch Phys Med Rehabil (APMR), Clin Rehabil (CR), Eur J Phys Rehabil Med (EJPRM) and Journal of Rehabilitation Medicine (JRM) were analyzed for the use of ordinal scales and the treatment of their data.

Results: Of totally 634 original articles, 316 (50%, range 31-61%) used multi-item ordinal scales. In 23 (7%, range 0-13%) of these, Rasch-analysis had been performed. None of the articles in AJPMR or EJPRM had used Rasch analysis. Seven (in APMR and JRM) of these articles also reported conversion tables. Other means of treating ordinal scales, instead of summed raw scores, as item for item analysis, dichotomization or other IRT type of analysis were used in 17 articles (5%, range 0-9%).

Conclusion: The use of Rasch analysis or other acceptable modalities for treating ordinal scales is still rare in rehabilitation research. It is to note that Rasch analysis may not be the solution for all ordinal scales used, as many of them may not be unidimensional and may not in their original form fulfill criteria for measures based on Rasch analysis. The use of conversion tables should be encouraged.

Reference: Grimby G, Tennant A, Tesio L. The use of raw scores from ordinal scales: Time to end malpractice? J Rehabil Med 2012;44:97-98.

Nikolaus Bezruczko
Standards and practices to guide health outcomes measurement: A strategy to avoid measurement malpractice

A fundamental proposition that guides American scientific research policy is "to do no harm" as described by the Belmont Report. Health outcomes literature shows increasing concern about measurement "mis-inference" because incorrect implementation of social behavioral measurement technology is raising risk for patients. A recurring issue concerns scale properties during patient reported outcome measurement. This presentation will review measurement practitioner responsibilities during government funded research with special attention to malpractice. Measurement scale properties will be reviewed and their vulnerability to misuse. Typical measurement results will be presented that demonstrate capricious disregard for patient well-being. Alternative methodologies will be summarized to lower patient and institutional risk from outcome measurement, and measurement practitioner strategies will be presented to monitor data quality, document measurement process and validity, and minimize threats of malfeasance.

Richard Smith, Lee McKenna, and Christie Plackner
A comparison of the information available in various measurement models

The precision of estimated is an important characteristic of scales. This is true of both person and item estimates. For persons the amount of information determines the SEM and the confidence interval for the estimate. For items the information impacts the determinations of estimate invariance and fit. With the use of simulated data the differences in item information will be highlighted for three Rasch models (dichotomous, rating scale, and partial credit) and three IRT models that correspond to these three data types.

4:45 pm – 5:00 pm **Closing Remarks**

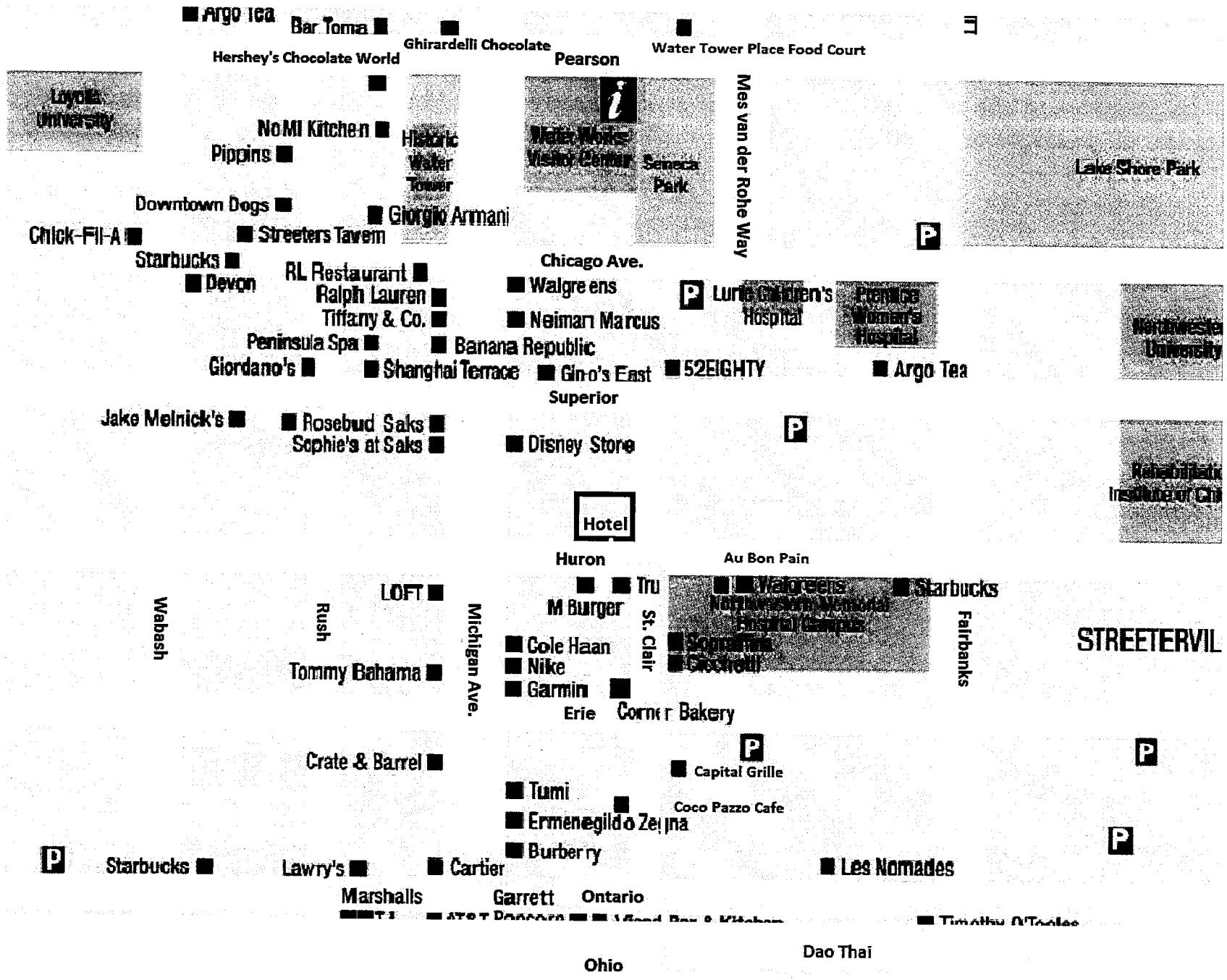
5:00 pm **End of Day 2**

All symposia will be plenary sessions. All paper sessions will be concurrent sessions, sessions **A** and **B** are in adjacent rooms.

A paper sessions generally highlight clinical presentations and **B** paper sessions generally highlight methodological issues.

Paper presentations will allow 20 minutes per presentation, with 20-30 minutes for discussion, depending on the number of presentations.

Revised April 8, 2015



Dao Thai Resraurant is located at 230 E. Ohio St., one block south of Ontario, between St. Clair and Fairbanks, on the north side of the street.