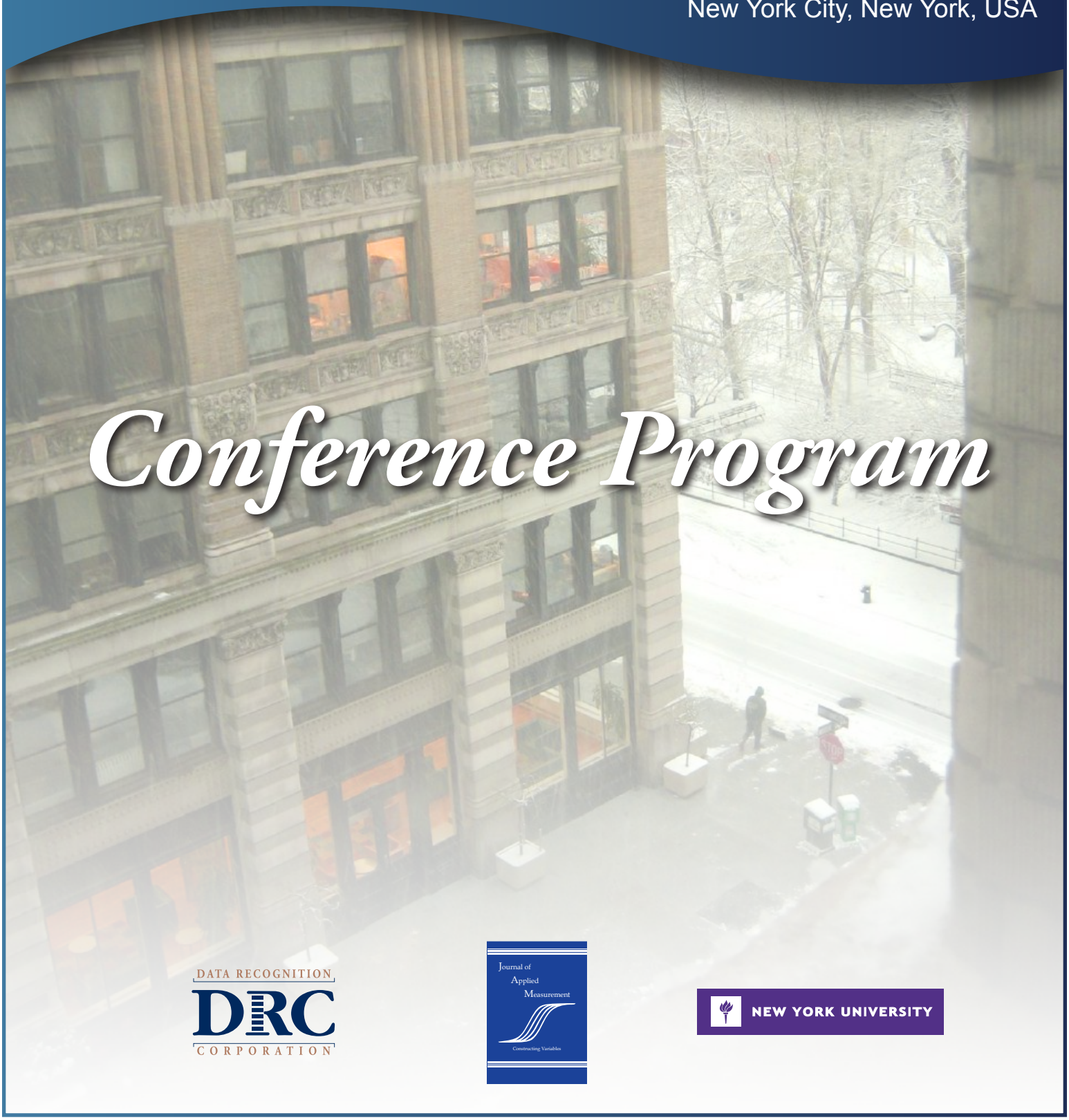
IOMW 2008

"Constructing Variables"

Saturday, March 22 to Sunday, March 23, 2008
New York City, New York, USA



Conference Program

DATA RECOGNITION
DRC
CORPORATION

Journal of Applied Measurement
Constructing Variables

NEW YORK UNIVERSITY

IOMW 2008 Conference Program

Saturday, March 22

Registration and Breakfast (8:00 - 8:30)

Conference Welcome: (Room 714, 8:30 - 8:45)
David Chayer, Data Recognition Corporation (dchayer@datarecognitioncorp.com)

Session 1 (Room 714, 8:45 - 10:15)

The Rasch Model as a Power Series Distribution for Equating Tests Which Measure the Same Proficiency

Presenting Author: David Andrich, The University of Western Australia (david.andrich@uwa.edu.au)

Abstract: Test equating is integral to modern psychometrics. Leunbach (1976) proposed the application of the power series distribution (Noack, 1950) for equating tests. The statistic of this distribution is an integer which is the power of the parameter of the distribution, where the distribution also has a parameter associated each integer. If this integer is taken as a personal test score without identifying scores on items, and the distribution parameter is taken as a personal parameter, the distribution is a general form of the class of Rasch models. Using the data illustratively, this paper relates Leunbach's work to recent advancements in Rasch models and confirms that the two tests can be equated with just the first two moments of the test parameters. Implications for test equating using test scores directly are considered. In particular, the focus is on each test as a whole than on each item individually. In general, this reduces substantially the number of parameters that need to be estimated and can lead to greater stability of the parameters of those that are estimated. For example, it is shown that in equating the two Leunbach tests, shown effectively just three parameters need to be estimated.

Historical View of Theories of Measurement and Language Proficiency within the Context of Second Language Testing

Presenting Author: George Engelhard, Jr., Emory University (gengelh@emory.edu)

Abstract: The purpose of this study is to briefly explore the interactions among measurement theories, theories of language acquisition, and measurement practices from a historical perspective. The assessment of English as a second language provides the framework for examining how theories influence, and in some cases fail to influence, the practice of English proficiency assessment as operationalized in a variety of tests. The first section describes a conceptual framework for examining the assessment of English as a second language. Next, a description of major research traditions in measurement theory that have dominated measurement practices during the 20th century is provided. The next section, introduces the major theories of language acquisition. Next, the findings from the previous two sections are used to examine the adequacy of the proposed conceptual framework for examining the assessment of English proficiency. This section includes criticism of measurement theory by selected language theorists. It also provides a brief history of the use of Rasch measurement theory to calibrate the English proficiency tests. Finally, the main points of the study are summarized and discussed. It should be recognized that this study represents a preliminary analysis of these issues.

Generally Objective Measurement of Human Temperature and Reading Ability: Some Corollaries

Presenting Author: Jack Stenner, MetaMetrics (jstenner@lexile.com)

Abstract: We argue that a goal of measurement is general objectivity: point estimates of a person measure (height, temperature, and reader ability) should be independent of the instrument used and independent of the sample in which the person happens to find herself. In contrast, Rasch's concept of specific objectivity requires that only differences (i.e. comparisons) between two person measures are independent of the instrument. We present a canonical case in which there is no overlap between instruments and persons: each person is measured by a unique instrument. We then show what is required to estimate measures in this degenerate case. The canonical case forces a simplification and reconceptualization of validity and reliability. Not surprisingly, this reconceptualization looks a lot like the way physicists and chemometricians think about validity and measurement error. We animate this presentation with a technology that blurs the distinction between instruction and assessment and results in generally objective measures of reader ability.

A Study of the Influence of Labels Associated with Anchor Points of Likert-Type Response Scales in Survey Questionnaires

Presenting Author: Jean-Guy Blais, University of Montreal (jean-guy.blais@umontreal.ca)

Co-Author: Julie Grondin, University of Montreal

Abstract: Survey questionnaires serve to obtain data at relatively low cost. However, several factors can influence the responses given and have an impact on the validity of the data thus collected. Among these factors are the introductory text at the beginning of the questionnaire, the formulation of items, the order of item presentation, the categories and response labels of the proposed scale used. The format of the response scale has been flagged many times as significant in certain context. This influence is traced to two sources: the number of anchor points used and the labels chosen to name each of these points (numeric value and wording). Though researchers know that the scale used for recording the response to an item can influence the responses given, there is relatively little literature leading to a better understanding of the scope of this influence. Thus, the objective of the presentation is to illustrate how a measurement model from the Rasch model family, the model for ordered response categories (i.e. Andrich's Rating scale model), can serve to study the influence of response scale labels on the responses' distribution and on the modelled position of anchor points.

Break (10:15 - 10:30)

Session 2a (Room 714, 10:30 - 12:00)

Symposium Title: Partial Credit Model Analyses of Psychological, Social, and Cultural Factors and Relationships with Individual Experiences of Chronic Pain: A Symposium

Symposium Organizer: Karen M. Schmidt, University of Virginia (kschmidt@virginia.edu)

Abstract: Many millions of individuals suffer from chronic pain (Harstall, 2003). Unlike acute pain, chronic pain persists, often leading to physical and emotional suffering. The prevalence and debilitating effects of chronic pain in our society emphasize the importance of research to discover successful methods to aid chronic pain sufferers. Recently, biopsychosocial and biocultural models (Gatchel, 2004), which address treatment in a holistic manner, have emerged in the health care arena. These models suggest that a person's experience of illness and pain is a conglomeration of biological, psychological, social, and cultural factors. This symposium explores these ideas, specifically investigating how the role of personality dimensions (Big Five), social factors (significant other support) and culture (India vs. America) influence a chronic pain sufferer's experience of pain intensity, psychological distress, and locus of control. We are hopeful that the results of this research will be used to illustrate the importance of assessing each chronic pain sufferer in an individualized manner, and be extended to the successful treatment of the millions of individuals seeking pain treatment in clinical settings.

Never Getting a Break: Persistent High Pain Intensity Relationships with Personality in Chronic Pain Sufferers

Presenting Author: Karen M. Schmidt

Pain Intensity, Catastrophizing, and Affect in Chronic Pain Sufferers

Presenting Author: Monica K. Erbacher

Influence of Significant Other Response on Pain Intensity and Psychological Distress in Chronic Pain Sufferers

Presenting Author: Katie J. Ameringer

Rasch Partial Credit Model (PCM) and Differential Item Functioning (DIF) Analysis of the Impact of Culture on Locus of Control in Chronic Pain Sufferers

Presenting Author: Juliana R. Schroeder

Pain Coping and Significant Other Relationships for Chronic Pain Sufferers

Presenting Author: David J. Lick

Session 2b (Room 713, 10:30 - 12:00)

Using Rasch Analysis to Construct a Trust in Medical Technology Instrument

Presenting Author: Enid Montague, Virginia Tech (enid.nicole@vt.edu)

Co-Authors: Edward W. Wolfe, Virginia Tech
Brian M. Kleiner, Virginia Tech
Woodrow Winchester II, Virginia Tech

Abstract: Understanding trust in medical technology provides insight to how medical technologies are used, misused, disused or abused by patients and physicians. Understanding how the two user groups negotiate technology usage may also provide insight into health care issues such as medical error, malpractice, system adoption and satisfaction. A 72-item instrument was developed to measure a persons trusting attitudes towards medical technology. Each item uses a 5-point likert type scale. The items for the instrument were derived from literature about trust in technology and the healthcare domain. Our analyses were conducted within an item response theory framework. Specifically, we scaled the data using the Rasch ratings scale model, evaluated dimensionality by examining residuals of observed item responses and the estimated parameters, we evaluated the manner in which the respondents employed the rating scale as intended. Characterized the quality of the items as indicators of the construct, and assessed the

reliability of measures from the instrument. The study results imply that the Rasch measurement model is a useful method producing a Unidimensional instrument to assess trust in medical technology.

Assessing Student Perceptions of High School Science Classroom Environments: A Validation Study

Presenting Author: Christine D. Luketic, Virginia Tech (cluketic@vt.edu)

Co-Authors: Edward W. Wolfe, Virginia Tech
Kusum Singh, Virginia Tech
Erin Dolan, Virginia Tech

Abstract: This validation study examines measures from the Science Laboratory Environment Inventory (SLEI), an existing classroom environment measurement instrument, for the purpose of providing up-to-date norms and validation evidence for a U.S. secondary school population. Our structural equation modeling analyses revealed that a multi-dimensional model encompassing five distinct factors and excluding negatively-worded items best characterized the SLEI measures. Multidimensional measures created by scaling the data to the Multidimensional Random Coefficients Multinomial Logit (Rasch) Model exhibited suitable rating scale structure, item quality, and reliability of separation. In addition, comparison of gender and ethnicity groups revealed no differences in SLEI measures.

Measuring Positiveness Towards Educational Policy

Presenting Author: Jinnie Choi, University of California, Berkeley (jinnie@berkeley.edu)

Abstract: In this paper, I aimed to illustrate the construction of the latent variable that measures peoples' positiveness towards an educational policy. Particularly, a thirty-five year old governmental educational policy in Korea was chosen as a target policy, in recognition that the most controversial research results around it lacked valid development and implementation of the policy-related attitude variables. It was explained how Wilson (2004)'s four building blocks (i.e. construct map, items design, outcome space, and measurement models) provided the procedural basis to build a sound instrument for the pilot study. Both classical test theory and item response modeling methods were implemented to examine the performance of the scale. The numerical and graphical results from ConQuest and ConstructMap software and the interpretations of them demonstrated various ways to collect reliability and validity evidence of the variable. In conclusion, I addressed the issues regarding future refinement, interpretation, and utilization of the measure.

Using the Standardized Letters of Recommendation in Selection: Results from a Multidimensional Rasch Model

Presenting Author: Ou Lydia Liu, Educational Testing Service, (lliu@ets.org)

Co-Authors: Jennifer Minsky, Educational Testing Service
Guangming Ling, Educational Testing Service
Patrick Kyllonen, Educational Testing Service

Abstract: To standardize academic application procedures, the Standardized Letter of Recommendation (SLR) was developed to capture important cognitive and noncognitive applicant characteristics. The SLR consists of seven scales (Knowledge, Analytical Skills, Communication, Organization, Professionalism, Teamwork, and Motivation) and was applied to an intern-selection scenario. Both professor ratings ($N = 414$) during the application process and mentor ratings after the internship was completed ($N = 51$) were collected using SLR. A multidimensional Rasch investigation suggests that the SLR displayed satisfactory psychometric properties in terms of reliability, model fit, item fit statistics, and discrimination. Two scales (Knowledge and Analytical Skills) were found to be the best predictors for intern selection. Furthermore, the professor ratings were systematically higher than the mentor ratings. Possible reasons for the rating discrepancies are discussed. Also, implications for how the SLR can be used and improved in other selection situations are suggested.

Lunch (12:00 - 1:30)

Session 3a (Room 714, 1:30 - 3:00)

Measuring Student Proficiency with the Constructing Measures Framework

Presenting Author: Brent Duckor, University of California, Berkeley (bduckor@berkeley.edu)

Co-Author: Mark Wilson, University of California, Berkeley

Abstract: This paper examines the learning progression among students in measurement education who are struggling with understanding what constitutes knowledge of measurement as defined by Wilson (2005) and others (National Research Council, 2001; Mislevy, Almond, & Lucas, 2003). It addresses the problem of "how do we know what our students of measurement know" by showing how a pre-post test instrument developed for a graduate course in higher education can measure change in student understanding and also meet nationally recognized standards of quality control (APA, AERA, NCME, 1999) for instrument design. The results of this paper are relevant to those

measurement ecducators who are interested in the variability in understanding and use of the 4 building blocks in the CM framework (Wilson, 2005) and those who seek to better understand how students' progress with skills and knowledge related to course topics and themes in related graduate courses.

A Comparison of Structural Equation and Multidimensional Rasch Modeling Approaches to Confirmatory Factor Analysis

Presenting Author: Edward W. Wolfe, Virginia Tech (edwolfe@vt.edu)

Co-Author: Kusum Singh, Virginia Tech

Abstract: Researchers have applied the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM) to confirmatory factor analyses (CFA). However, to date, there have been no evaluations of the suitability of that model for recovering the dimensional structure of item response data. This paper compares the results of CFA applications of the MRCMLM to those obtained in comparable structural equation modeling (SEM) analyses. Specifically, we apply MRCMLM and SEM CFA techniques to three real datasets to determine the comparability of the identified dimensional structure, interfactor correlations, and factor reliabilities. The results indicated that the two methods perform comparably in identifying dimensional structure, but they provide different depictions of factor reliabilities and intercorrelations in some situations.

A Multilevel Item Response Theory Analysis of Health-Related Quality of Life: An Illustration with 11,158 Healthy and Chronically Ill Children using the PedsQL™ Emotional Functioning Scale

Presenting Author: Prathiba Natesan, University of Miami (prathibachaj@gmail.com)

Co-Authors: Christine Limbers, Texas A&M University James W. Varni, Texas A&M University

Abstract: This study demonstrates the use of Graded Response Multilevel Item Response Theory (GRMLIRT) analysis in estimating item parameters and the effect of a specific covariate, health status, on the latent trait, health-related quality of life. In so doing, the present study demonstrates the evaluation of Samejima's Graded Response Model using multilevel analysis. WINBUGS, freeware software, was used for this analysis. Data from the Emotional Functioning Scale on 11,158 healthy and chronically ill children and adolescents was utilized from the PedsQL™ 4.0 Generic Core Scales Database to demonstrate the methodology. With health status as the covariate, the R2 effect size indicates that healthy children have a better emotional quality of life than their chronically ill counterparts, as can be expected. This paper serves as an illustration of MLIRT in pediatric health outcomes research and as a tutorial to the novice researcher in IRT.

A Summary Index of Multidimensionality in Scales Composed of Subscales: Applications to Traditional and Rasch Measurement Theory

Presenting Author: Barry Sheridan, RUMM Laboratory (rummlab@arach.net.au)

Co-Author: David Andrich, University of Western Australia

Abstract: Many scales in social measurement constructed to measure a single variable are nevertheless composed of subscales which measure different aspects of the variable. Although the presence of subscales increases the validity of the scale, it compromises its unidimensionality. Many studies have considered this compromise in from both the traditional and modern test theory perspectives. This paper derives a formula for readily calculating a summary index of multidimensionality among all subscales from a traditional test theory perspective and applies it using the Rasch model of modern test theory. The formula rests of the change in the value of the traditional reliability index when calculated at two levels. First at the level of the original items and second at the level of the subscales. The formula's structure confirms that an analysis at the level of subscales accounts for multidimensionality among subscales. A set of simulation studies, generated according to the Rasch model, illustrates the effectiveness of the formula in recovering the degree of multidimensionality. The application of the formula is shown with a read data set from the Hospital Anxiety/Depression Scale.

Session 3b (Room 713, 1:30 - 3:00)

ConstructMap: A Software Demonstration

Presenter: Andy Maul, Berkeley Evaluation and Assessment Research (BEAR) Center

Co-Presenter: Cathleen Kennedy, Berkeley Evaluation and Assessment Research (BEAR) Center

Abstract: In this workshop we will introduce several new features of the ConstructMap model fitting program (formerly called GradeMap). This workshop will focus on setting up and fitting unidimensional and multidimensional one-parameter models and generating and interpreting commonly used reports such as Wright Maps, Item Fit Reports, and Item Plots. If time and interest permit, we will also discuss advanced features such as the analysis of alternate forms and instrument alignment through common item equating.

Break (3:00 - 3:30)

Session 4a (Room 714, 3:30 - 5:00)

A Bootstrap Approach to Evaluating Person and Item Fit to the Rasch Model

Presenting Author: Edward W. Wolfe, Virginia Tech (edwolfe@vt.edu)

Abstract: Historically, rule-of-thumb critical values have been employed for interpreting fit statistics that depict anomalous person and item response patterns in applications of the Rasch model. Unfortunately, prior research has shown that these values are not appropriate in many contexts. This article introduces a bootstrap procedure for identifying reasonable critical values for Rasch fit statistics and compares the results of that procedure to applications of rule-of-thumb critical values for three example datasets. The results indicated that rule-of-thumb values may over- or under-identify the number of misfitting items or persons.

Are QOL and Spirituality Separate Constructs? A Lesson from Breast Cancer Patients

Presenting Author: Nikolaus Bezruczko, Measurement and Evaluation Consulting (nbezruczko@msn.com)

Co-Authors: Kyle Perkins, Florida International University
David Cella, Northwestern University

Abstract: A measurement investigation was conducted into the joint calibration of Functional Assessment of Cancer Therapy-General (FACT-G), a 28 item cancer quality of life (QOL) measure, and Functional Assessment of Chronic Illness - Spiritual Well-Being (FACIT-Sp), a 12 item spiritual well-being measure. Both FACT and FACIT-Sp are validated for cancer patients, and separate scales have respectable measurement properties. Their measurement properties after consolidation, however, have never been examined. A Rasch model for rating scales was implemented with WINSTEPS for data from 532 breast cancer patients. Results show coherent consolidation of QOL and Spirituality on a common dimension. Item map shows items to intermingle on the measurement dimension rather than form separate blocks, which suggests successful dimensional integration. Moreover, overall item distribution is continuous without major breaks or interruptions. Only a few items arguably show larger than expected fit values. Both patient separation (2.66) and reliability (.88) for a three category rating scale improve after joint calibration. Residual analysis does not show major threats to dimensionality, and joint calibration explains item variance comparable to separate calibration (51.9%). However, comparison of patient measures for separate and consolidated QOL scales is unusual because overall patient distribution does not significantly change, while ethnic subgroup means shift significantly. These results suggest that inclusion of spiritual values requires serious reconsideration of conventional ideas about QOL.

Measures That Count, Measures That Matter

Presenting Authors: William P. Fisher, Jr., Avatar International (wfisher@avatar-intl.com)
Brent Duckor, University of California, Berkeley (bduckor@berkeley.edu)

Abstract: The story of modern measurement theory is tied up with the history of science, mathematics and philosophy. In this paper, we place the search for theoretically tractable constructs in the context of principles and systems that can yield a set of shared artifacts to unify discourse in the Rasch community. Taking two cases—the BEAR system and the Lexile framework—we argue that important lessons can be drawn from these approaches to measurement in education and the social sciences. Whether motivated by epistemological, practical or ethical concerns, science in general aspires to the establishment of a common currency, a root metaphor, a gold standard, a proportionate rationality, a sufficient reason, or a reference standard metric of measurement—in short, a philosophical logos—that unifies the discourse of a field as the shared object of ongoing conversations. This paper explores the first principles of these aspirations while also considering the possibility that some domains, topics, skills, etc. are more amenable to construct-theory driven measurement than others.

Exploration of a Taxonomic Framework to a New Instrument Development and Item Types: Dimensional Disaster or Informed Instrument?

Presenting Authors: Judy R. Wilkerson, Florida Gulf Coast University (jwilkers@fgcu.edu)
W. Steve Lang, University of South Florida St. Petersburg (ws.lang@knology.net)

Abstract: The construct of dispositions is well defined in national standards, and colleges of education are required to assess candidate dispositions to meet accreditation requirements. Measurement, however, is virtually nonexistent. On-line reviews of college accreditation reports indicate that colleges are attempting to assess dispositions without the use of sound measurement techniques or adequate definitions of the construct. The end result, of course, is a reliance on face validity. Rasch measurement provides a much needed solution to scaling the dispositions needed for good teaching. This paper presents early work in the development of six related instruments measuring ten national principles related to dispositions. The instruments use different item structures and response formats,

which will eventually be combined into a single disposition scale. More specific to this paper is the construct mapping and analysis of the items by Krathwolh's Taxonomy and INTASC Principles. If valid construct planning dictates ordered dimensions, how will the items calibrate? This paper explores the initial results of a pilot study.

Session 4b (Room 713, 3:30 - 5:00)

ConQuest: A Software Demonstration

Presenter: Ou Lydia Liu, Educational Testing Service (lliu@ets.org)

Co-Presenter: Jinnie Choi, University of California, Berkeley

Abstract: In this ConQuest workshop, we will introduce some basic features of this program. We will demonstrate the required commands to run analyses in ConQuest. We will provide some example analyses to illustrate how ConQuest deals with Rasch analyses, rating scale analyses, and partial credit analyses. We will explain outputs produced by ConQuest.

Sunday, March 23

Registration and Breakfast (8:00 - 8:30)

Session 5 (Room 714, 8:30 - 10:15)

Ben Wright's "Life" in the Village from 1933 to 1947

Presenting Authors: Ed Bouchard (ed@edbouchard.com) Nikolaus Bezruczko, Measurement and Evaluation Consulting (nbezruczko@msn.com)

Abstract: In 1990, Ben attended his elementary school graduation fiftieth reunion. Ben was adamant that The Little Red School House, a school that creatively used "Life" in Greenwich Village itself as the core curriculum, was his "best school," that he "learned more science" there than elsewhere. Given excellent teachers at Ben's secondary school, studies at Cornell with Nobel Laureates Hans Bethe and Richard Feynman, work at Bell Laboratories with Nobel Laureate Charles Townes, and studies at the University of Chicago with Nobel Laureate Robert Mulliken, we explore whether Ben's assessment was an exaggeration. Or whether this school's teachers and experiential curriculum, in fact, were what prepared him to work with Nobel Laureates and to make profound contributions to measurement. He began at Little Red in 1933, at age seven. After a year hiatus, he returned to Little Red in 1936, remaining until graduating in 1940. Except for summers at a farm in Stroudsburg, PA in the early 1940s, the Village remained his home base through boarding school, years at Cornell, and work at Bell Laboratory. Studying at this school, and his life as a Village denizen, subsequently informed whatever he would do in his life and career.

Functional Assessment in a Wellness Program for the Frail Elderly

Presenting Author: Dr. Carl V. Granger, SUNY at Buffalo, Department of Rehabilitation Medicine, Uniform Data System for Medical Rehabilitation (cgranger@udsmr.org)

Abstract: Objective: To incorporate functional assessment into wellness program evaluation and care management for frail elderly living in an adult home. Participants: 45 residents were evaluated over 18 months at 3- month intervals. All residents were over 70 years of age. Interventions: The wellness program included group and individual exercise classes, nutrition counseling, current event discussions, memory circles and other cognitive training, and a broad range of recreational activities. Main Outcome Measures: The LIFEware® assessment includes measures of physical and cognitive function (18-item FIMTM instrument, memory measure), behavior, affect/mood, satisfaction with life, participation in social activities, physical limitations, pain level, anxiety/depression, and quality of sleep. LIFEware® ratings are between 0 and 100; 70 is the threshold for clinical significance; higher number is better state of function. Results: Physical limitations, affect, sleep, and satisfaction with life were well maintained (ratings higher than 70); whereas memory, motor function and social level were at risk (ratings close to or under 70). Lower ratings and more variation in ratings over time were associated with transfer to long-term care. Conclusions: Behavior, cognition, memory, and FIMTM instrument motor items were particularly suitable for monitoring the effects of a wellness program for the elderly. Regular assessments prompted earlier interventions to help residents maintain functional ability and allow them to remain in an adult home setting rather than skilled nursing.

Validity and Objectivity in Health Related Scales: A Second Look at SF36

Presenting Author: Svend Kreiner, University of Copenhagen (s.kreiner@biostat.ku.dk)

Abstract: Many summated scales developed for health research suffer from two problems. Reliability is poor because the number of items is limited and measurement is neither valid nor objective since items are locally dependent and

function differentially relative to several person covariate. The subscales of the much used SF36 questionnaire is a particular case in point. The physical functioning (PF) subscale containing have three different subsets of locally dependent items and several items that function differentially relative to gender, age and self reported health. It is possible to eliminate item such that the remaining items fit a pure partial credit model, but the resulting objective subscale has only four items and consequently poor reliability and large standard errors of measurement. Instead one may fit a loglinear Rasch model with uniform DIF and uniform local item dependence. In this model, the total score is sufficient for the person parameter and inference concerning person parameters may be separated from inference on parameters describing item properties in precisely the same way as in ordinary Rasch model. The difference in terms of validity and objectivity between measurement by loglinear Rasch models and measurement by ordinary Rasch models will be discussed.

Measuring Mental Health Problems among Adolescents– the Youth Self Report Examined with the Rasch Model

Presenting Author: **Curt Hagquist, Karlstad University** (curt.hagquist@kau.se)

Co-Authors: David Andrich, University of Western Australia
Sven R. Silburn, Curtin University of Technology
Stephen R. Zubrick, Curtin University of Technology

Abstract: The purpose of this paper is to use the unidimensional Rasch model to examine the psychometric properties of two broad YSR-scales (Youth Self Report) intended to measure internalising and externalising mental health problems respectively. The study makes use of data from Western Australia concerning adolescents 12-16 years old, collected among 788 boys and girls. The present paper highlights the issue of targeting. The results provided by the Rasch-analysis reflect the construction of the YSR as a clinical instrument and confirm that many of the items may not be useful for constructing latent variables in general populations. As a whole the original scales on internalising and externalising mental health problems seem to be of doubtful values in separating adolescents in a general population across the whole continuum of a latent scale. The results show that neither the set of items intended to measure internalising problems nor the one on externalising problems conform to the Rasch model. In order to achieve accordance between the items and the Rasch model a great number of items of each scale were removed. These final sets of items show acceptable internal consistency but, because the number of items removed is large, capitalising on chance effects is also large.

Session 6a (Room 714, 10:30 - 12:00)

Solving Incomplete Paired Comparison Matrices

Presenting Author: Ronald Mead, Data Recognition Corporation (rmead@datarecognitioncorp.com)

Abstract: The estimation algorithm for paired comparisons (Rasch, Choppin, Andrich) is as old as the Rasch model itself. This paper provides a new presentation of the method intended to be instructional to those unfamiliar with Rasch's principles. The method is so logically connected to Rasch's formulation that understanding this approach to estimation can illuminate Rasch's method for test development and variable construction. When the log-odds matrix of pairwise observations is complete, the solution is well-known: the logit item difficulty is the sum of its row divided by the number of items. When the matrix is incomplete, a solution has typically been obtained iteratively, which raises the question of when to stop. A simple, non-iterative solution is suggested and tried out, which can be readily applied in today's testing environments. This paper presents a general solution based on a simple algorithm for determining a coefficient matrix that accounts for the missing elements. The solution is demonstrated with simulated data and standard errors are discussed. The algorithm may also be applied directly to the polytomous models, when the log-odds matrix is necessarily incomplete.

Measurement of Student Nurse Performance in the Safe Administration of Medication

Presenting Author: Deborah Ryan, Emory University (dryan@emory.edu)

Abstract: Medication errors are a significant concern in today's healthcare environment. Patient safety is dependent on nurses who consistently demonstrate behaviors fundamental to safe administration of medication. This study describes the development and performance of the Safe Administration of Medication Scale (SAM Scale) to objectively measure student nurse ability in identifying medication errors. The SAM Scale is a paper and pencil instrument, with a series of five clinical cases of hospitalized adults and children. Each case provides patient information, chief complaint, history & physical, diagnosis, physician orders and medication orders. Two or three associated vignettes describe the actions taken by the nurse. The principles of Rasch measurement, as proposed by Wilson (2005) provide the theoretical framework for this study. Data analysis was conducted using FACETS: Rasch Measurement Program. Nursing students enrolled in associate degree nursing programs (N=137) and baccalaureate students (N=130) participated in this study. Associate degree student nurses made more errors than baccalaureate degree student nurses. While both groups identified a majority of

the errors, 115 students did not identify a potentially lethal error made by the pharmacist. The ongoing development of the SAM Scale will include a significantly larger population of students and incorporate additional challenging medication errors.

Measurement of Visual Disability in Low Vision Patients: Does DIF for Health Status Matter?

Presenting Author: Lohrasb Ahmadian, Johns Hopkins University (lahmadi2@jhmi.edu)

Abstract: PURPOSE: To examine the effects of distortion from health status on Rasch model-based measurements of visual function, and establish measurement equivalence across different health status groups in low vision patients. METHODS: Self-reported data were obtained from 1746 low vision patients who completed the Activity Inventory (AI) and an intake health-related questionnaire prior to their first visit to the low vision rehabilitation service. Differential Item Functioning (DIF) analysis by health status and separate Rasch analyses adjusted by health were performed on the responses to both a DIF-free scale and the full scale of the AI. RESULTS: Of 48 Goal-level items, only 15 items showed significant DIF (p<0.001). Comparing the vision-related estimates from the original full set of items with those from the DIF free scale; we found that only 25% of the person measure estimates differed by 0.5 logits or more and there was a strong intraclass correlation between the two scales in measuring the patients' visual ability (IC=0.75). Patients' health influenced the Rasch model based estimation of visual ability by the AI (ANOVA, p= 0.005), but this effect was within accepted range of MISFIT statistics. CONCLUSION: Despite confounding effects of health status on visual ability, we can still regard visual ability to be a single theoretically constructed variable for the low vision population. It appears that self-perceived comorbidities add to vision-related disability, but do not distort its measurement.

Proficiency at Scoring and Preventing Touchdowns: Pairwise Comparisons

Presenting Author: Vincent Primoli, Data Recognition Corporation (vprimoli@datarecognitioncorp.com)

Co-Authors: Christie Plackner, Data Recognition Corporation
Ronald Mead, Data Recognition Corporation

Abstract: The National Football League includes 32 teams, arranged in eight groups of four. Each team plays 16 regular season games, half at its own home field, typically against 13 different opponents. Success including qualification for the post-season playoff games is based primarily on a simple count of the number of wins without consideration of quality of the opponents. This paper analyzes the 2008 NFL season as 512 pairwise comparisons. The analysis yields an alternataive ranking of teams based on the log ratio of points scored versus points allowed. This scaling of teams correctly predicted the winner for 74% of the games. The paper also examines relative strength of schedule, possible home field effects, the week 16 effect, and other anomolies. Of particular interest to Rasch analysts, the paper presents a straightforward, non-iterative algorithm for dealing with a partially-filled pair-wise matrix.

Session 6b (Room 713, 10:30 - 12:00)

Using the Rating Scale Model to Examine the Angoff Ratings of Standard-Setting Panelists

Presenting Author: Jade Caines, Emory University (jcaines@emory.edu)

Co-Author: George Engelhard, Emory University

Abstract: Some approaches to standard setting recommend including broad representation on the panel from varying stakeholders. In education, specifically, the panel may consist of constituents who have a stake in the cut scores set for high stakes educational assessments. There is concern, however, that panelists without the necessary content knowledge may differentially influence the results of the standard-setting process. Also, panelists that lack direct knowledge of student academic performance may be unable to perform the cognitive tasks required by certain standard-setting procedures. The primary purpose of this study is to examine the quality of judgments obtained from standard-setting panelists who represent various demographics and stakeholder roles. In other words, do the ratings of standard-setting panelists differ by gender, ethnic background, geographic region, and stakeholder role? A secondary purpose is to illustrate the use of the Rating Scale Model (Andrich, 1978) to provide a methodological framework to systematically examine these judgments. These questions are addressed with ratings from an English language arts standard-setting panel over 3 rounds of judgments. The results show no difference in final cut scores based on gender, ethnic background, and stakeholder role. Differences, however, do exist in final cut scores based on geographic regions.

Examining the Bookmark Ratings of Standard-Setting Panelists: An Approach Based on the Multifaceted Rasch Measurement Model

Presenting Author: Rubye Sullivan, Emory University (rsulli4@emory.edu)

Co-Authors: Jade Caines, Emory University
Courtney Tucker, Emory University
George Engelhard, Jr., Emory University

Abstract: The purpose of this study is to describe a new approach for evaluating the judgments of standard-setting panelists within the context of the bookmark procedure (Mitzel, Lewis, Patz, and Green, 2001). The bookmark procedure is widely used for setting performance standards on high-stakes assessments. A multifaceted Rasch measurement (MRM) model is proposed for evaluating the bookmark judgments of the panelists, and its use illustrated with standard-setting judgments from the Michigan Educational Assessment Program (MEAP). Panelists set three performance standards to create four performance levels (Apprentice, Basic, Met, and Exceeded). The content areas used to illustrate the model are mathematics and reading in Grades 3, 4, 5, 6, 7, and 8 and science in Grades 5 and 8. The analyses will be conducted to examine the bookmark ratings based on the model described in Engelhard (2007). Implications for research, theory and practice regarding standard-setting procedures will be discussed.

The Construction of the Malaysian Educators Selection Inventory (MedSI): A Large Scale Assessment Initiative

Presenting Author: Joharry Othman, International Islamic University (proff9@hotmail.com)

Co-Authors: Taib Harun, UKM; Norlena Salamuddin, UKM; Syed Mohamed Shafeq Syed Mansor, UTM; L. Borhan, UM; W.M.W Jaafar, UPM; N. Abdullah, UiTM; S. M. Noah, UPM; A.M Abdul Karim, UUM; A.M Abdul Rahman, UPSI; S. Ibrahim, UTHM; J.Ahmad, UPM; T. Sulaiman, UPM; Z. Mohamed, UMS; M.Y. Ab.Hadi, UTHM; M. Ghazali, USM

Abstract: This paper will present the development and creation of a nation wide assessment instrument for use in the selection process of candidates taking education courses in Malaysian public universities. The background and rationale for such a large scale instrument will be discussed. One of the main rationale is to make sure that the right candidates are being screened and selected to enter the teaching profession, in line with the requirements of the New Malaysian Educational Policy to ensure that the teaching profession produces competent teachers and educators. This instrument named MedSI consisted of 300 likert scale items covering 4 constructs, namely personality, interest, integrity and EQ. Its construction processes and psychometric properties will be elaborated and preliminary results on the first cohort of candidates taking education courses for the year 2007 shall be explained. Since this is a national endeavor by the Malaysian Ministry of Education, it is hoped that MedSI would helped better identify and screen the right candidates (according to the 4 constructs) to the teaching programs. MedSI will become the first and main screening instrument followed by face to face interviews for those who passed MedSI.

Construct Development for Linear Measurement of Accessibility to Education in Regions of the Russian Federation

Presenting Author: Anatoli Maslak, Slavyansk-on-Kuban State Pedagogical Institute (anatoliy_maslak@mail.ru)

Co-Authors: Tatyana Anisimova Slavyansk-on-Kuban State Pedagogical Institute
Nikolaus Bezruczko, Measurement and Evaluation Consulting

Abstract: This research addresses a requirement to establish an overall, latent trait measure of accessibility to education for pupils in 89 regions of the Russian Federation. Consequently, a Rasch model for rating scales was implemented to estimate difficulty parameters of 10 indicators on a Unidimensional scale, and then assess their conformability to a common theoretical construct. Results show these indicators defining an equal interval, latent trait scale ranging from the easiest item "Percent of children and teenagers age 7-15 years not trained in educational establishments" to hardest "Percent of children in short-term preschool establishments". Based on a Chi-Square fit analysis, all qualitative indicators conformed to a common latent trait construct and are useful for measuring accessibility to education.

Lunch (12:00 - 1:30)

Session 7a (Room 714, 1:30 - 3:00)

Defining a Measurement Scale for Curriculum Evaluation

Presenting Author: Ronald Mead, Data Recognition Corporation (rmead@datarecognitioncorp.com)

Co-Authors: Julie Korts, Data Recognition Corporation
Kyoungwon (Kei) Lee, Data Recognition Corporation

Abstract: The current emphasis of school accountability has lead to extensive student level testing. However, to be truly accountable, the schools also need an economical and reliable method to document the match between the local curriculum and the content standards that are the basis of the student test design. The current study created an instrument that allows districts to evaluate their curri-

cula. The results reported here are from an analysis of that created for district use. The instrument used Likert-type response to a series of prompts concerning the appropriate of the design and content of the instructional units. The data reported here involve four content areas, with 15 to 20 courses each, covering pre-school through high school, with 10 to 15 instructional units for each. The questionnaire was completed by locally trained content specialists. Three or four judges rated each unit. This analysis uses a multifaceted rating scale model to explore the question of whether the questionnaire is equally effective across grade level and across content areas. There is an expectation that appropriate instruction and material will become more complex in high grade levels. The study also considers how the increased complexity is reflected in the responses, if at all.

Inferring an Experimentally Independent Response Space for the Rasch model for Ordered Categories from its Experimentally Dependent Subspace

Presenting Author: David Andrich, The University of Western Australia (david.andrich@uwa.edu.au)

Abstract: Assessments in ordered categories are ubiquitous in the sociological and biomedical sciences. Following R.A. Fisher, we consider it mandatory that the ordering of the categories is an empirical property of the assessments. This paper shows the logistic Rasch model for ordered categories provides necessary and sufficient evidence that the categories are ordered empirically. It provides such evidence because an experimentally independent response space can be inferred from a subspace in which the responses in categories are necessarily dependent. A simulated and a real research example using social mobility data illustrate this interpretation of the model. It is also argued that because important research, individual, group or policy decisions can rest on the putative order of the category of a response, statisticians and psychometricians have a responsibility to identify malfunctioning category structures.

Optimizing Response Categories in a Measure of Health Care Quality Perceptions

Presenting Author: William Fisher, Jr., Avatar International (wfisher@avatar-intl.com)

Co-Authors: Geoffrey A. Nagle, Tulane University
Clayton Williams, Louisiana Public Health Institute

Abstract: Objective. Public hospital emergency department clients' sense of quality health care is hypothesized quantitatively measurable. An initial test of that hypothesis focused on the rating response structure. Setting. Three urban academic health sciences center public hospital emergency rooms. Participants. Clients (total $N = 444$) visiting one of the emergency rooms on one of 10 consecutive days. Each ER was sampled in a volume proportionate with its portion of the total visits, and clients were approached in every range of hours in the entire 24-hour period. Main outcome measure. A 13-item survey concerning perceptions of quality in health care was designed with client participation. Participants rated their sense of quality health care using a six-point scale ranging from Very Strongly Disagree to Very Strongly Agree.
Analyses. The rating structure was experimentally evaluated 11 times, fitting a series of related probabilistic additive conjoint measurement models, with the aim of satisfying each of six criteria associated with construct-valid data quality and interpretability. RESULTS: In the first two analyses, none of the six criteria were met; in the last two, all six were. CONCLUSIONS: Establishing the ordinal consistency of survey-based observations in this way sets the stage for evaluations of the construct's invariance over respondent subsamples.

Is the Partial Credit Model a Rasch Model?

Presenting Author: Robert W. Massof, Johns Hopkins University (rmassof@lions.med.jhu.edu)

Abstract: Rasch models must satisfy the requirement of statistical sufficiency of person and item raw scores and separability of model parameters. The Andrich rating scale model (RSM) satisfies these requirements, but because of the interaction between items and response category thresholds, the Masters partial credit model (PCM) does not. When attempting to prove statistical sufficiency of the item raw score and separability of parameters, Masters implicitly made assumptions that allowed the item-dependent step measures to be factored out, thereby making the item-dependent step raw score vector appear to be a sufficient statistic for the step measures. However, his arguments in effect require the trivial condition that there is only a single item. The relationship between person measures and the expected response is free to vary across items, therefore, Masters' implicit assumptions are not valid and the item raw score is not a sufficient statistic. However, the PCM can be viewed as a scale equating device, thereby preserving scale invariance post hoc.

Session 7b (Room 713, 1:30 - 3:00)

Development of the "Chinese Character Difficulty Scale" for Primary students

Presenting Author: Magdalena Mo Ching MOK, The Hong Kong Institute of Education (mmcmok@ied.edu.hk)

Co-Author: Dr Yee Man Cheung, The Hong Kong Institute of Education

Abstract: In Chinese language development, characters serve as the basic building blocks in sentence construction and in reading comprehension for pupils aged from 6 to 9. Consequently, measures on Chinese character knowledge can be used as a proxy for the language ability of young children. A Revised Chinese Character List, based on the original developed in 1990 and comprising 2376 characters, was recently published by the Hong Kong government. Nevertheless there was no guideline as to when during the first three years of primary schooling should these characters be taught. This study aims to develop a Rasch scale to measure the difficulty levels of the Chinese characters in the Revised List, so that teachers can make evidence-based decisions on their teaching. The sample for this study comprised 30 primary school teachers and about 2000 students in primary 1 to 3. Thirty-six versions of test booklets were designed to provide comprehensive coverage of all characters in the Revised List. Each booklet was made up of a list of 130 to 132 items (characters), with 50% linkage to the next booklet in the series. Teachers were also invited to identify reasons for sources of difficulty in relation to attributes of the characters.

Constructing the Variable "Explanation in Mathematics"

Presenting Author: Brian Doig, Deakin University (badoig@deakin.edu.au)

Co-Author: Susie Groves, Deakin University

Abstract: This paper reports on upper primary and junior students' explanations to mathematics questions in a large-scale project in Victoria, Australia. These questions asked students to write, or draw, explanations that we believe are susceptible to analysis with the tools described by earlier researchers, but are enhanced through the use of IRT analyses, that allows the construction of a 'map' developing capability in, or sophistication of, explanation in mathematics not hitherto possible. The methodology of this research is detailed and the outcomes described.

Testing the Assumption of Sample Invariance of Item Difficulty Parameters in the Rasch Rating Scale Model

Presenting Author: Joseph A. Curtin, Brigham Young University (joseph_curtin@byu.edu)

Co-Authors: Richard R. Sudweeks, Brigham Young University
Richard M. Smith, Data Recognition Corporation
Joseph A. Olsen, Brigham Young University

Abstract: This study tests the invariance property of Rasch item difficulty estimates using non-simulated, polytomously scored data. The data used in this study was obtained from a university alumni questionnaire that was collected over a period of five years. The analysis tests for significant variation in item difficulty estimates between (a) small samples taken from a larger sample, (b) a base sample and subsequent (longitudinal) samples and (c) variation over time controlling for confounding variables. The confounding variables identified include: (a) the gender of the respondent and (b) the respondent's major at the time of graduation. Two methods are used to assess variation: (a) the between-fit statistic, and (b) a general linear model. The general linear model uses the person residual statistic from the WINSTEPS person output file as a dependent variable with year, gender and type of major as independent variables. Results of the study support the Rasch invariance property of item difficulty estimates in polytomously scored data. The analysis found consistent results when comparing the between-fit statistic and the general linear model methods. An advantage of the general linear model is its ability to identify both the source and the effect size of observed variations in item difficulty estimates.

Break (3:00 - 3:30)

Session 8a (Room 714, 3:30 - 5:00)

Calibrating Instruments for Improving What We Do: Establishing Rasch Measurements for Self-Theories of Intelligence

Presenting Author: Sharon Solloway, Bloomsburg University (sharon@solloway.net)

Abstract: Implicit self-theories of intelligence influence an individual's motivation when faced with challenges in learning. Making self-theories explicit has been shown to influence changes in a student's goals and concerns and thus offer possibilities for enhancing the potential for at-risk students' academic success. Published items in the self-theories of intelligence literature suggest eight categories. This paper describes a project to develop linear scales for these items using a Rasch rating scale model. These items sets were modified to maximize the likelihood of provoking responses that would meet fundamental measurement theory's requirements for additive invariance. The difficulties of beginning from within a theory not oriented toward measurement-theory versus beginning from a qualitatively-informed approach are outlined. The results demonstrate the possibility of measuring the effects of curriculum designed to make self-theories explicit across time. Suggestions propose a refined self-theories instrument that includes conceptually valid but non-fitting items refined through collection of qualitative data.

Constructing a Variable: Hotel Satisfaction

Presenting Author: Trevor Bond, Hong Kong Institute of Education (tbond@ied.edu.hk)

Abstract: The Partial Credit Rasch measurement model was used to construct a general Guest Satisfaction Scale for major aspects of the hotel operations of a major international group in Townsville, Australia. In spite of the difficulties inherent in working with data collected via an instrument of another's design (i.e. the one in use by the chain at the time) and the rather haphazard nature of those collected data, the constructed Scale showed robust psychometric properties. In principle, the scale should transfer to other Hotel properties in the group to measure the underlying satisfaction variable. Moreover, it is likely that the Rasch measurement principles should apply equally well to other instruments designed to collect hotel guest satisfaction data. The hotel group was provided with graphical feedback based on the ubiquitous Wright map principles. The Guest Satisfaction Scale worked like a ruler - equal distances between levels reflecting equal differences in Guest Satisfaction and between Satisfying Feature locations. Least satisfied guests and least satisfying features of the Hotel were easily identified on the graphic displays and the amount of dissatisfaction could be measured. Hotel management was then better informed about whom or what to target in terms of its own cost-benefit estimates.

The Effect of Assessment Context on Construct Definition in Direct Writing Assessment

Presenting Author: Sharon E. Osborn Popp, Arizona State University (osbornpo@asu.edu)

Abstract: This study investigated the impact of assessment contexts on the definition of the construct of writing, as outlined in an analytic rubric. Features that characterize the assessment context, but are not explicit in the scoring rubric, such as discourse mode and grade level, may impact writing assessment score outcomes and alter the trait-based representation of the construct of writing. Observed ratings from writing responses in two discourse modes for each of two grade levels were analyzed using a many-faceted Rasch model. The contexts of mode and grade level were compared through student responses to a common prompt in both grades. Findings included variance in the order and degree of difficulty of analytic traits between modes and grade levels. Adapting scoring procedures to explicitly incorporate identifiable context-dependent scoring elements, such as expectations related to grade-level or discourse mode, is recommended.

Déjà vu: The Rasch Measurement Model and Google's PageRank Algorithm

Presenting Author: Mary Garner, Kennesaw State University (mgarner@kennesaw.edu)

Abstract: "PageRank" is a widely-acclaimed algorithm used for determining the ranking or ordering of web pages by web search engines. The PageRank algorithm has much in common with one particular algorithm for estimating the parameters of the Rasch model – the eigenvector method (EVM) described by Garner and Engelhard (2000; 2002; in press). The purpose of this paper is to explain the similarities between the EVM and the PageRank algorithm, and explore how these similarities could enhance our understanding and use of the Rasch Measurement Model.

Session 8b (Room 713, 3:30 - 5:00)

The Conjoint Additivity of the Lexile Framework for Reading

Presenting Author: Andrew Kyngdon, MetaMetrics (akyngdon@lexile.com)

Abstract: Reading is one of the most important skills a person acquires. The Lexile Framework for Reading is a system for the measurement of reading which matches continuous prose text to reader ability. It argues that comprehension of continuous prose text is a non-interactive, additive function of reader ability and text difficulty, consistent with measurement as argued by the Rasch (1960) model. However, the Framework has been publicly criticised by Joseph Martineau as not providing interval scale measurement of reader ability. In part to address this criticism, the Framework was test with the theory of conjoint measurement (Luce & Tukey, 1964). Given the ordinal constraints imposed by the cancellation axioms of this theory, an order restricted inference approach must be taken (Iverson & Falmagne, 1984). Karabatsos' (2001) Bayesian Markov Chain Monte Carlo methodology was used to analyse data taken from 130 North Carolina 2nd grade children responding to 200 native Lexile text passage items. It was found both the single and double cancellation axioms of conjoint measurement were probabilistically satisfied. Such results support the conceptual argument of the Lexile Framework and that it provides measures of reader ability and text difficulty on the same interval scale.

A Developmental Framework for the Measurement of Writer Ability

Presenting Author: Harold Burdick, MetaMetrics (hburdick@lexile.com)

Co-Authors: Jack Stenner, MetaMetrics
Donald Burdick, MetaMetrics
Carl Swartz, MetaMetrics

Abstract: This study provides the empirical foundation for a developmental scale of writer ability that generalizes across rater, prompt, rubric, and occasion. A developmental scale of writing ability allows educators, researchers, and policy-makers to monitor student growth in written language expression both within and across grades from elementary through post-secondary education. The proposed developmental scale is a significant advance over human and machine-scoring systems that employ raw scores based on holistic and analytical scoring rubrics. Participants (N = 589) enrolled in grades 4, 6, 8, 10, or 12 in a rural/suburban school district in north-central Mississippi wrote one essay in response to each of six prompts; two narrative, two informative, and two persuasive. Three prompts, one from each genre, were shared with the adjacent grade thus ensuring the connectivity needed to detect a developmental scale. Many-Facet Rasch Measurement (MFRM; Linacre, 1994) was used to create the developmental scale while adjusting for genre, prompt, and rubric variation. Results provide strong evidence that the goal of constructing a developmental scale of writing across grades 4 to 12 was achieved. Implications for research on writing assessment and instruction are discussed.

Using Confirmatory Factor Analysis and Rasch Measurement Theory to Assess Measurement Invariance in a High Stakes Reading Assessment

Presenting Author: Jennifer Randall, University of Massachusetts, Amherst (jrandall@educ.umass.edu)

Co-Author: George Engelhard, Jr., Emory University

Abstract: The psychometric properties and multigroup measurement invariances of scores across subgroups, items and persons on the Reading for Meaning items from the Georgia Criterion Referenced Competency Test (CRCT) were assessed in a sample of 997 seventh grade students. Specifically, we sought to determine the extent to which score-based inferenced on a high stakes state assessment hold across several subgroups within the population of students. To that end, both multi-group confirmatory factor analysis (to examine factorial invariance) and Rasch (1980) measurement theory (to examine item-level invariance) were used. Results revealed a Unidimensional construct with factorial-level measurement invariance across disability status (students with and without disabilities), but not across test accommodations (resource guide, read-aloud, and standard administrations). Item level analysis also revealed some differential item functioning suggesting the need for closer examination of items by content experts for some student groups.

An Examination of Fairness of Math Word Items

Presenting Author: Xuejun (Ina) Shen, Stanford University (xuejuns@stanford.edu)

Co-Authors: Xiaohui Zheng, University of California, Berkeley
Edward Haertel, Stanford University

Abstract: Great efforts have been made over past many years to ensure tests are constructed in ways that treat people equally regardless of their backgrounds. However, even carefully written test questions may still be particularly difficult for certain subpopulations due to factors, such as students' cultural, linguistic, racial and gender backgrounds, other than what the questions are intended to measure. This study offers a case study of the fairness of math word items to English language learner (ELL), ethnic minority, low socio-economic status (SES) and female students. Logistic regression-based differential item functioning (DIF) with multiple group membership variables and SIBTEST bundle analysis techniques are used to examine these subgroups' performances on math word items versus numeric items addressing the same or similar arithmetic concepts and skills.