Journal of Applied Measurement Abstracts

Volume 7, Number 1 (2006)

Expansion of a Physical Function Item Bank and Development of an Abbreviated Form for Clinical Research

Rita K. Bode
Rehabilitation Institute of Chicago/Northwestern University
Jin-shei Lai
Kelly Dineen
Evanston Northwestern Healthcare/Northwestern University
Allen W. Heinemann
Rehabilitation Institute of Chicago/Northwestern University
Daniel Shevrin
Evanston Northwestern Healthcare/Northwestern University
Jamie Von Roenn
Northwestern Memorial Hospital/Northwestern University
David Cella
Evanston Northwestern Healthcare/Northwestern University

We expanded an existing 33-item physical function (PF) item bank with a sufficient number of items to enable computerized adaptive testing (CAT). Ten items were written to expand the bank and the new item pool was administered to 295 people with cancer. For this analysis of the new pool, seven poorly performing items were identified for further examination. This resulted in a bank with items that define an essentially unidimensional PF construct, cover a wide range of that construct, reliably measure the PF of persons with cancer, and distinguish differences in self-reported functional performance levels. also developed a 5-item (static) assessment form ("BriefPF") that can be used in clinical research to express scores on the same metric as the overall bank. The BriefPF was compared to the PF-10 from the Medical Outcomes Study SF-36. Both short forms significantly differentiated persons across functional performance levels. While the entire bank was more precise across the PF continuum than either short form, there were differences in the area of the continuum in which each short form was more precise: the BriefPF was more precise than the PF-10 at the lower functional levels and the PF-10 was more precise than the BriefPF at the higher levels. Future research on this bank will include the development of a CAT version, the PF-CAT.

Using Rasch Analysis to Test the Cross-Cultural Item Equivalence of the Harvard Trauma Questionnaire and the Hopkins Symptom Checklist Across Vietnamese and Cambodian Immigrant Mothers

Yoonsun Choi
Amy Mericle
University of Chicago
Tracy W. Harachi
University of Washington

A major challenge in conducting assessments in ethnically and culturally diverse populations, especially using translated instruments, is the possibility that measures developed for a given construct in one particular group may not be assessing the same construct in other groups. Using a Rasch analysis, this study examined the item equivalence of two psychiatric measures, the Harvard Trauma Questionnaire (HTQ), measuring traumatic experience, and the Hopkins Symptom Checklist (HSCL), assessing depression symptoms across Vietnamese- and Cambodian American mothers, using data from the Cross-Cultural Families (CCF) Project. The majority of items

were equivalent across the two groups, particularly on the HTQ. However, some items were endorsed differently by the two groups, and thus, are not equivalent, suggesting Cambodian and Vietnamese immigrants may manifest certain aspects of trauma and depression differently. Implications of these similarities and differences for practice and the use of IRT in this arena are discussed.

Using Rasch Measurement to Investigate Volleyball Skills and Inform Coaching

Ryan P. Bowles
Nilam Ram
University of Virginia

This paper illustrates some of the ways that Rasch modeling techniques can be used to inform coaching of volleyball. Volleyball game statistics for 11 volleyball players collected over 27 matches on 3 skills were analyzed using a multifaceted Rasch model (Lincare, 1999) incorporating the partial credit model (Masters, 1982). Rating scale analyses and model fit statistics were used to derive interval level scales that provided more objective information regarding player ability and consistency than is usually available to coaches. Detailed results illustrate that Rasch analyses can provide very specific information that coaches might use in drill, practice, and game strategy design.

The Assessment of Diabetes Knowledge and Self-Efficacy in a Diverse Population Using Rasch Measurement

Ben S. Gerber
Maria Pagcatipunan
Everett V. Smith, Jr.
Semonti S. Basu
Kimberly A. Lawless
Louanne I. Smolin
Michael L. Berbaum
Irwin G. Brodsky
Arnold R. Eiser
University of Illinois, Chicago

The purpose of this research was to develop survey instruments to evaluate diabetes knowledge and self- efficacy in a diverse population, and investigate the psychometric properties of data obtained with these instruments using Rasch measurement. Two-hundred and fifty-five urban-dwelling participants with diabetes were recruited to complete surveys through independent interviews. To evaluate the association of health literacy on metabolic control, formal literacy and hemoglobin A1c fingerstick testing were per-formed. Rasch analysis of the data yielded item and person calibrations for self-efficacy and knowledge, with variable maps created to provide both norm and criterion-referenced interpretations. Knowledge scale person separation reliability was 0.50 and item separation reliability was 0.98; while self-efficacy scale person separation reliability was 0.72 with item separation reliability of 0.92. Statistically significant par-tial correlations were observed between knowledge and health literacy ($r = 0.41$, $p<0$ .001), and self- efficacy and hemoglobin A1c ($r = –0.33$, $p<0.001$). However, there was no correlation between diabetes knowledge and hemoglobin A1c ($r = 0.035$, $p = 0.29$), or health literacy and A1c ($r = 0.022$, $p = 0.36$). Diabetes knowledge varied, with non-English speaking individuals having lower measures than English speakers ($t(252) = –4.86$, $p<.001$). Non-English speaking individuals also had lower self-efficacy measures than English speakers ($t(251) = –2.68$, $p = .008$). Current knowledge deficits and perceptions of self- management may be estimated visually through variable mapping, which may help in individualizing informational needs for people with diabetes.

Using LinLog and FACETS to Model Item Components in the LLTM

Tracy L. Kline
Karen M. Schmidt
Ryan Bowles
University of Virginia

The current study investigates the performance of two Rasch measurement programs and their parameter estimations on the linear logistic test model (LLTM; Fischer, 1973). These two programs, LinLog (Whitely and Nieh, 1981) and FACETS (Linacre, 2002), are used to investigate within-item complexity factors in a spatial memory measure tool. LinLog uses conditional maximum likelihood to estimate person and item parameters and is an LLTM specific program. FACETS is usually reserved for the many-facet Rasch model (MFRM; Linacre, 1989), however in the case of specifically designed within-item solution processes, a multifaceted approach makes good sense. It is possible to consider each dimension within the item as a separate facet, just as if there were multiple raters for each item. Simulations of 500 and 1000 persons expand the original data set (114 persons) to better examine each estimation technique. LinLog and FACETS analyses show strikingly similar results in both the simulation and original data conditions, indicating that the FACETS program produces accurate LLTM parameter estimates.

Development of a Comprehensiveness of Rasch Measurement Application Scale

Iasonas Lamprianou
University of Manchester

A large number of papers and technical reports are published every year describing researches where Rasch models are used. It has been observed, however, that not all the authors describe the application of the Rasch measurement with the same thoroughness. Some authors may leave behind important bits of information e.g. they may fail to investigate the person or item fit or may even fail to discuss the reliability of measurement. As a result, editorial guidelines have been published in order to suggest an informal 'minimum' of thoroughness with which the authors may describe the application of Rasch measurement in their papers. This study presents stages for the development of a scale to investigate the comprehensiveness with which individual papers describe the application of Rasch models in practical settings. The scale is used to evaluate how comprehensively the papers published by the *Journal of Applied Measurement* present the application of Rasch models.

Issues in Multi-Item Scale Testing and Development using Structural Equation Models

Shaun McQuitty
James W. Bishop
New Mexico State University

Employing a structural equation model to evaluate a measurement scale can be challenging, especially for a multidimensional scale that contains many items. We describe two issues that can contribute to the poor fit of such models: the statistical power associated with the test of a large measurement scale; and the degree of correlation between items and factors within the scale. These issues are not well understood, so our purpose is to explain them at an applied level, clarify their practical implications for tests of measurement scales and other large structural equation models, and discuss potential strategies for addressing them.

Rasch Analysis of Rank-Ordered Data

John M. Linacre
University of Sydney, Australia

Theoretical and practical aspects of several methods for the construction of linear measures from rank-ordered data are presented. The final partial-rankings of 356 professional golfers participating in 47 stroke-play tournaments are used for illustration. The methods include decomposing the rankings into independent paired comparisons without ties, into dependent paired comparisons without ties and into independent paired comparisons with ties. A further method, which is easier to implement, entails modeling each tournament as a partial-credit item in which the rank of each golfer is treated as the observation of a category on a partial- credit rating scale. For the golf data, the partial-credit method yields measures with greater face validity than the paired comparison methods. The methods are implemented with the computer programs *FACETS* and *WINSTEPS*.

Volume 7, Number 2

Rasch Analysis of a New Construct: Caregiving for Adult Children with Intellectual Disabilities

Shu-Pi C. Chen
St. Xavier University
Nikolaus Bezruczko
Chicago, IL
Sheila Ryan-Henry
Seguin Retarded Citizens Association

This research examined empirical evidence for a new construct, Functional Caregiving, which is a theory about mothers' caregiving of their adult children with intellectual disabilities. A sample of 108 biological mothers and primary caregivers rated survey items about their confidence to perform caregiving tasks. Rasch rating scale analysis found 61 items defined an empirical construct with three caregiving levels: Advocacy, Personal Caregiving, and Community. Results show item separation was 3.11 with high reliability, .91, and mother separation was 2.93 and reliability, .90. Both items and mothers showed adequate INFIT and OUTFIT values. Item invariance was confirmed between older and younger mothers, and principle components analysis of residuals did not reveal any major dimensionality threats. Item decomposition analysis showed FC content theory to account for 58 percent of item calibration variance (R^2 = .58, F = 42.3, p < .001). These results have important practical implications for health and social services, as well as family caregiving, interdisciplinary practices, and health policy development.

Whose Criterion Standard Is It Anyway?

Gregory Ethan Stone
The University of Toledo

A criterion-referenced standard is an important element of most successful professional testing programs. A growing body of evidence suggests that judge decisions are influenced by characteristics related to the normative experience of the individual judge (e.g. gender, age, etc.). This investigation used two health-care related boards to explore the effects of judge characteristics on the standards established. Two judge panels (com-posed of 26 and 30 members respectively) were used in a simplified Objective Standard Setting exercise to define examination cutoff points. Multi-faceted Rasch analyses were employed to detect and explore differences in judgment making. Significant but not necessarily consistent differences were found between panel judges on several examined characteristics. Results suggest that criterion-referenced standards defined by judge panels are inexorably connected to their normative experiences and are therefore wholly sample dependent. While stratification of judge panels is clearly an important element in defining standards, if they are ever to achieve the goals of Glaser (1963) and Majer (1962) including meaningful independence, more must be done to investigate these and other concerns. The case for the predictive validity of criterion-referenced standards has not thus far been made in any convincing fashion.

Adjusted Rasch Person-Fit Statistics

Dimiter M. Dimitrov
George Mason University
Richard M. Smith
Data Recognition Corporation

Two frequently used parametric statistics of person-fit with the dichotomous Rasch model (RM) are adjusted and compared to each other and to their original counterparts in terms of power to detect aberrant response patterns in short tests (10, 20, and 30 items). Specifically, the cube root transformation of the mean square for the unweighted person-fit statistic, *t*, and the standardized likelihood-based person-fit statistic Z_3 were adjusted by estimating the probability for correct item response through the use of symmetric functions in the dichotomous Rasch model. The results for simulated unidimensional Rasch data indicate that *t* and Z3 are consistently, yet not greatly, outperformed by their adjusted counterparts, denoted *t** and Z3*, respectively. The four parametric statistics, *t*, Z_3, *t**, and Z_3*, were also compared to a non-parametric statistic, H_T, identified in recent research as outperforming numerous parametric and non-parametric person-fit statistics. The results show that H_T substantially outperforms *t*, Z^3, *t**, and Z^3* in detecting aberrant response patterns for 20- item and 30-item tests, but not for very short tests of 10 items. The detection power of *t*, Z3, *t**, and Z3*, and H_T at two specific levels of Type I error, .10 and .05 (i.e., up to 10% and 5% false alarm rate, respectively), is also reported.

From Rasch Scores to Regression

Karl Bang Christensen
National Institute of Occupational Health, Denmark

Rasch models provide a framework for measurement and modeling latent variables. Having measured a latent variable in a population a comparison of groups will often be of interest. For this purpose the use of observed raw scores will often be inadequate because these lack interval scale properties. This paper com-pares two approaches to group comparison: linear regression models using estimated person locations as outcome variables and latent regression models based on the distribution of the score.

The Stability of Marker Characteristics Across Tests of the Same Subject and Across Subjects

Iasonas Lamprianou
University of Manchester

This research investigates the stability of marker characteristics within a very short period of time for both tests on the same subject as well as tests on different subjects. It reports on the scoring of the scripts of the whole cohort of students that took three high stakes tests in 2003 in a European country: a Language test consisting of a Literacy and a Literature paper and a History test. The many-facets Rasch model was used to study marker severity and marking consistency and it was found that some markers had more stable characteristics than others. Although the stability of marker characteristics was generally weak, it was non-negligible (correlation indices as indicators of stability ranged up to 0.707). This study, however, is not absolutely accurate due to the small sample sizes employed and it can be added that more research is needed to reach definite results.

Development of a Money Mismanagement Measure and Cross-Validation Due to Suspected Range Restriction

Kendon J. Conrad
Michael D. Matters

University of Illinois at Chicago
Daniel J. Luchins
Patricia Hanrahan
University of Chicago
Danielle L. Quasius
University of Illinois at Chicago
George Lutz
North Chicago VA Hospital

A measure of the tendency to mismanage money was developed in an evaluation of a representative payee program for individuals with serious mental illnesses. A conceptual model was composed to guide item development, and items were tested, revised, added, and rejected in three waves of data collection. Rasch analyses were used to examine measurement properties. The resulting Money Mismanagement Measure (M3) consisted of 28 items with a Rasch person reliability at .72. Restriction of range was likely responsible for the low Rasch reliability. Validity analyses supported the construct validity of the M3. Subsequently, a cross- validation study was conducted on an untreated sample not as susceptible to range restriction. The M3 produced a Rasch person reliability = .85 with good validity. The M3 fills a gap that can facilitate research in the understudied area of money mismanagement.

The Mixed-Rasch Model: An Example for Analyzing the Meaning of Response Latencies in a Personality Questionnaire

Michaela M. Wagner-Menghin
University of Vienna

The present study uses the Mixed-Rasch Model to analyze questionnaire data. It is hypothesized that more than one latent sub-population (classes) can be identified. It is also hypothesized that response latencies and social desirability can be used to describe and discriminate the latent classes. Using the software WinMira 2001, data from the Eysenck-Personality-Profiler 'German-version' (EPP-D) was analyzed. A so-called nonscaleable and a scaleable group of subjects were identified which differed with regard to the use of the response category 'I don't know,' however, there was no difference between these groups with regard to response latencies and social desirability. Results will be discussed with regard to the EPP-D scoring scheme as suggested in the manual.

Volume 7, Number 3

Measuring Teaching Ability with the Rasch Model by Scaling a Series of Product and Performance Tasks

Judy R. Wilkerson
Florida Gulf Coast University
William Steve Lang
University of South Florida, St. Petersburg

Rasch measurement can provide a much needed solution to scaling teacher ability. Typically, decisions about teacher ability are based on dichotomously scored certification tests focused on knowledge of content or pedagogy. This paper presents early developmental work of a partial-credit teacher-ability scale of 42 tasks (performances and products) with 348 rated items or criteria. The tasks and criteria are aligned with national and state standards for expected teacher knowledge and skills. These tasks are being used in about two-thirds of Florida school districts and are spreading to colleges of education. Over time there will be many variations in both tasks and criteria, but here we focus on the initial system and the Rasch model as part of the plan for development of the system.

An Introduction to the Theory of Unidimensional Unfolding

Andrew Kyngdon
University of New South Wales, Sydney, Australia

Despite its 55 year presence in the field of mathematical psychology, the theory of unidimensional unfolding remains an enigma for many psychometricians and applied practitioners. This paper is the first of a three part series; and it aims to introduce unidimensional unfolding theory. The paper begins with a simple hypothetical example presenting an idealised distinction between responses to cumulative and unfolding dichotomous items. This followed by an accessible presentation of the theory of unidimensional unfolding as first articulated by Clyde H. Coombs (1950, 1964). The concept of the single peaked preference function (Coombs and Avrunin, 1977) which underpins unfolding theory is then presented. The article then progresses to the class of Rasch (1960) based IRT models developed by Andrich (1995) and Luo (2001). It was shown these models propose arguments not inconsistent with Coombs's (1964) original theory. The presumption of additive structure in psychological attributes was concluded to be the key weakness of the theories of unidimensional unfolding discussed.

Estimating Person Locations from Partial Credit Data Containing Missing Responses

R. J. De Ayala
University of Nebraska-Lincoln

Certain assessment situations produce partial credit data. For instance, performance assessment items may utilize a rubric that assigns partial credit for some not completely correct responses. In some cases examinees may choose to not answer each question. This study investigated the effect of various strategies for handling these missing responses for estimating a respondent's location. These methods included ignoring the omitted response, selecting the "midpoint" category score, treating the omitted response as incorrect, hotdecking, and a likelihood-based approach. A simulation study was performed to examine the efficacy of these methods with the partial credit and generalized partial credit models. Expected a posteriori (EAP) ability estimation was used. Results showed that the Midpoint and Likelihood procedures performed the best of methods examined. In contrast, omitted responses should not be treated as incorrect nor ignored when estimating an examinee's proficiency using EAP. Implications for practitioners are discussed.

Validation of a Questionnaire Used to Assess Safety and Professionalism among Arborists

Steven G. Viger
Michigan State University
Edward W. Wolfe
Virginia Tech
Hallie Dozier
Krisanna Machtmes
Louisiana State University

This article summarizes a validation study of an instrument designed to measure safety and professionalism practices of arborists. A sample of 386 arborists from the State of Louisiana responded to the 58-item questionnaire. Analyses focused on several aspects of Messick's validation framework. Structural validity evidence was provided by analyses that indicate that the measures are unidimensional. Content validity evidence was supported by generally high positive values of the biserial correlations and optimal values of standardized mean-square item fit indices. Substantive validity evidence was provided by analyses that support the use of the two-point rating scale and a rank ordering of item means that is consistent with substantive theory. Person fit indices indicated little misfit among measures. Support for the generalizability aspect of validity was provided by an acceptable level of internal consistency and fairly tight error bands around estimated arborist measures. Additionally, few items exhibited DIF. Finally, with respect to the external aspect of validity, group differences between arborist measures were consistent with substantive theory.

How Accurate Are Lexile Text Measures?

A. Jackson Stenner
Hal Burdick
Eleanor E. Sanford
Donald S. Burdick
Metametrics, Inc.

The Lexile Framework for Reading models comprehension as the difference between a reader measure and a text measure. Uncertainty in comprehension rates results from unreliability in reader measures and inaccuracy in text readability measures. Whole-text processing eliminates sampling error in text measures. However, Lexile text measures are imperfect due to misspecification of the Lexile theory. The standard deviation component associated with theory misspecification is estimated at 64L for a standard-length passage (approximately 125 words). A consequence is that standard errors for longer texts (2,500 to 150,000 words) are measured on the Lexile scale with uncertainties in the single digits. Uncertainties in expected comprehension rates are largely due to imprecision in reader ability and not inaccuracies in text readabilities.

Rasch Modeling of the Structure of Health Risk Behavior in South African Adolescents

Elias Mpofu
Linda Caldwell
Edward Smith
The Pennsylvania State University
Alan J. Flisher
Catherine Mathews
University of Cape Town
Lisa Wegner
Tania Vergnani
University of the Western Cape

The study used Rasch analysis to investigate the presence of a syndrome of health risk behavior in South African adolescents. A total of 2186 in-school adolescents participated in the study (males = 1077; females = 1119; age range = 12-16 years; median = 13 years). The data are baseline from a longitudinal study of a leisure-based drug abuse and HIV/AIDS prevention program at Mitchell's Plain in Cape Town, South Africa. The adolescents completed a self-report measure on various health risk vulnerabilities, including use of alcohol, tobacco and other drugs (ATOD), co-occurrence of penetrative sex with use of ATOD, health related self- efficacy, personal beliefs about health, peer perceptions, and use of contraceptives. The Rasch analysis calibrated data on 50 items from the aforesaid conceptually distinct health risk domains. Infit and Outfit mean square statistics and principal components analysis of the standardized residuals suggested a fit of the data to the unidimensional Rasch measurement model. The findings support a syndrome view of health risk in teen-agers as proposed by problem behavior theory.

Multicomponent Latent Trait Models for Complex Tasks

Susan E. Embretson
Georgia Institute of Technology
Xiangdong Yang
University of Kansas

Contemporary views on cognitive theory (e.g., Sternberg and Perez, 2005) regard typical measurement tasks, such as ability and achievement test items, multidimensional, rather than unidimensional. Assessing the levels and the sources of multidimensionality in an item domain is important for item selection as well as for item revision and

development. In this paper, multicomponent latent trait models (MLTM) and traditional multidimensional item response theory models are described mathematically and compared for the nature of the dimensions that can be estimated. Then, some applications are presented to provide examples of MLTM. Last, practical estimation procedures are described, along with syntax, for the estimation of MLTM and a related model.

Volume 7, Number 4

Standard Systems: The Foundational Element of Measurement Theory

Marion S. Aftanas
University of Manitoba

All measurement involves some system or mechanism for deriving metric information and yet definitions and theories of the process in psychology have not emphasized this element. The theory of *standard systems*, on the other hand, introduces the system as the foundational element in the measurement process. When combined with a categorization of types of standard systems and other elements of the measurement process, the theory highlights, and provides a meta-theoretical framework for integrating the historically important and heuristic contributions to measurement theory. The early positive contributions that focused on the development of systems of measurement and models for deriving metric information, were deflected by the requirements outlined in the *physical* addition theory, and in the limiting theory of scales. Some of these requirements were liberalized by subsequent theoretical thrusts. Future research should bolster the promise of the Rasch solutions with emphasis on the provision of standard systems with empirically anchored magnitude points.

An Empirical Study into the Theory of Unidimensional Unfolding

Andrew Kyngdon
University of New South Wales, Sydney, Australia

This article is the second in the series on unidimensional unfolding. Its aim was to test the quantitative component of Coombs's (1964) theory via an empirical application to subjective control in gambling behavior (Dickerson and Baron, 2000). It was found that approximately 96% of judgments upon bilateral stimulus pairs were as predicted by the theory of unidimensional unfolding. The double cancellation axiom of the theory of axiomatic conjoint measurement (ACM) (Krantz, Luce, Suppes and Tversky, 1971) was satisfied by the interstimulus midpoint order obtained from these judgments. These results supported previous unfolding studies on attitudes (Johnson, 2001; Michell, 1994). Exponential and linear relationships were found between the transformed scaling solutions of Coombs's (1964) theory and the SHCMpp (Andrich, 1995). The implications of these results were discussed. Additionally, the article presented both a formal theory of item construction (Michell, 1994) and an accessible demonstration of the Goode's algorithm scaling procedure.

Expanding an Existing Multiple Choice Test with a Mixed Format Test: Simulation Study on Sample Size and Item Recovery in Concurrent Calibration

Insu Paek
Michael J. Young
Harcourt Assessment

When a new set of mixed format items is augmented with a previous old multiple-choice (MC) test, those mixed format items should be linked to the existing old MC test. This study used simulation to investigate sample size effect on recovery of known item parameter from the concurrent calibration in the context of horizontal equating, where the new mixed format tests are equated to the existing MC test which acts as the common linking items. In the partial credit model following the Andrich style parameterization, item location and item step parameters were

differentially affected by the sample size. Item location parameters were recovered better than item step parameters at the individual item, the sub-test, and the total test level. This study also shows the outward bias for the item location parameter estimated by the maximum likelihood estimator.

Fitting Polytomous Rasch Models in SAS

Karl Bang Christensen
National Institute of Occupational Health, Denmark

The item parameters of a polytomous Rasch model can be estimated using marginal and conditional approaches. This paper describes how this can be done in SAS (V8.2) for three item parameter estimation procedures: marginal maximum likelihood estimation, conditional maximum likelihood estimation, and pairwise conditional estimation. The use of the procedures for extensions of the Rasch model is also discussed. The accuracy of the methods are evaluated using a simulation study.

The Development and Validation of the Self-Directed Learning Scales (SLS)

Magdalena Mo Ching Mok
Cheng Yin Cheong
Phillip John Moore
Kerry John Kennedy
The Hong Kong Institute of Education, Hong Kong

This article describes the development and validation of the Self-directed Learning Scales (SLS) using data from 14,846 secondary students. Self-directed learning refers to a process whereby the learner consciously and actively directs his/her actions in the learning process. The SLS comprised a battery of subscales measuring students' goal setting, planning, academic motivation, academic self-efficacy, inquiry and information processing, strategic help-seeking, management of learning resources, and self-monitoring. Rasch analysis following factor analyses provided evidence in support of the validity of SLS for use with secondary students. Two original subscales were merged with other subscales on the basis of the analyses, resulting in 19 subscales with strong psychometric properties in the Self-directed Learning Scales.

Using Paired Comparisons to Create the Semantic Construct of Frequency

Thomas R. O'Neill
National Council of State Boards of Nursing, Inc.

Using a paired comparison data collection procedure and a one-faceted Rasch model, a stable construct of frequency is derived using 43 non-numeric quantitative descriptors, such as never, almost never, rarely, sometimes, often, etc. This construct is acontextual in that only the words or phrases were compared. The raters were not supplied with any particular context. To make the paired comparison survey more manageable certain assumptions were made that would result in a sparse, but connected 43x43 data matrix. Three hundred ninety-six pairs were selected and each pair was assigned to one of two forms. The quality control procedures to detect pair-order effects, fatigue effects, aberrant raters, model misfit, etc. are discussed. The results were mapped onto a continuum, which seems to represent the common understanding of frequency using non-numeric quantitative descriptors. These results have implications for how many different strata of frequency people can reliably differentiate in spoken language. This study can also serve as a primer for using paired comparison data collection procedures with the Rasch model and the methodology can easily be applied to other similar semantic continua. This study used a small sample and the pool of raters was not very diverse, therefore future studies should further consider those issues.
