

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Kohei Adachi  
Ritsumeikan University

## Exploring Nonlinear Inter-Variable Relations By A Variant Of Nonmetric Principal Component Analysis

Let  $\mathbf{x}_j$  denote the  $j$ th column of an objects by variables matrix and let the mutually different  $K_j$  values in  $\mathbf{x}_j$  be collected into  $\mathbf{y}_j = [y_{j1}, \dots, y_{jK_j}]'$  (in the increasing order): for example, if  $\mathbf{x}_1 = [6, 30, 14, 6, 30]'$ , then  $K_1 = 3$  and  $\mathbf{y}_1 = [6, 14, 30]'$ . Nishisato (2001) has shown that nonlinear correlations between variables  $\mathbf{x}_j$ , which cannot be captured by linear PCA (principal component analysis), can be found by treating value  $y_{jk}$  as just a nominal category  $j-k$  to be given score vector  $\mathbf{w}_{jk}$  with MCA (multiple correspondence analysis). That is, the trajectories connecting the resulting  $\mathbf{w}_{jk}$  and  $\mathbf{w}_{j, k+1}$  may allow us to capture nonlinear correlations. However, if  $K_j$  is large and scores to be obtained are too many, results may become unstable so that trajectories are too zigzag. To deal with this difficulty, we propose a nonmetric PCA procedure, in which an objective function is defined as the sum of the loss function for MCA and a penalty function requiring trajectories to be smooth splines. We report results of a simulation study and discuss the difference between our method and the existing nonmetric PCA which is a constrained MCA technique with rank-one restriction.

---

Kohei Adachi  
Ritsumeikan University

## A Quasi Tucker2 Method Based on Principal Component Analyses of Slices of a Three-Way Array

We present a quasi Tucker2 procedure for analyzing semantic differential rating data to capture perceptual and semantic structures that subjects have. In the method,  $\mathbf{F}_i \mathbf{G}_i' \mathbf{A}_i'$  is fitted for a stimuli by adjectives matrix  $\mathbf{X}_i$  for subject  $i$ , where  $\mathbf{F}_i = \mathbf{F} \mathbf{G}_i$  and  $\mathbf{A}_i = \mathbf{A} \mathbf{B}_i$  represent perceptual and semantic structures for  $i$ , respectively. For identification purpose, we consider the following constraint and requirements. (1)  $\mathbf{F}_i$  is column-orthonormal. (2)  $\mathbf{F}_i$ 's are similar to  $\mathbf{F}$  (common percepts) and  $\mathbf{A}_i$ 's are similar to  $\mathbf{A}$  (common semantics). (3)  $\mathbf{A}$  has simple structure. Without requirements (2) and (3), the optimal  $\mathbf{F}_i$  and  $\mathbf{A}_i$  fitting best, which we express as  $\mathbf{F}_i^{(0)}$  and  $\mathbf{A}_i^{(0)}$ , are obtained with the principal component analysis of  $\mathbf{X}_i$ . We can further let  $\mathbf{F}_i = \mathbf{F}_i^{(0)} \mathbf{T}_i \mathbf{S}$  and  $\mathbf{A}_i = \mathbf{A}_i^{(0)} \mathbf{T}_i \mathbf{S}$  meet (2) and (3) without changing the goodness of fit, where  $\mathbf{T}_i$  and  $\mathbf{S}$  are orthonormal rotation matrices. That is, requirement (2) is met by obtaining  $\mathbf{T}_i$  with a kind of Procrustes method and (3) is achieved by obtaining  $\mathbf{S}$  with an orthomax method. The resulting  $\mathbf{F}$ ,  $\mathbf{A}$ ,  $\mathbf{F}_i$ , and  $\mathbf{A}_i$  give  $\mathbf{G}_i$  and  $\mathbf{B}_i$ . The presented method is illustrated with real data analysis.

---

Carolyn J. Anderson and Hsiu-Ting Yu  
University of Illinois

## Relationships Between Item Response Theory Models and Log-Multiplicative Association Models

Log-multiplicative association models, which are extensions of log-linear models and generalizations of Goodman's (1981, 1985) RC(M) association model, can be interpreted as latent variable models. In this talk, we discuss the relationship between log-multiplicative association models and standard IRT models. Theoretical relationships between these models are made using statistical graphical models, Holland's (1990) Dutch Identity, and conditional specification of models. Empirical demonstrations of equivalencies and similarities between the two classes of models are presented to backup the theoretical claims. The connections yield insights into the models themselves and provide potential methodological contributions to measurement, including a way to estimate models without the use of numerical integrations, models with regressions on the latent ability, estimation of individuals' value on the latent variable as a by product of fitting the model to data, and generalizations to multidimensional models.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Judit Antal  
American Institutes for Research

Tamas Antal  
The Ohio State University

Assessing Rater Behavior in Large Scale Assessment

This proposal introduces a new method for assessing rater behavior in a Generalizability theory framework in cases when typical multi-facet designs (either classical test theory or item response theory based) are not fully available due to the nature of large-scale testing situations. That is, for example, in a situation when a small sample of student's work is randomly assigned to raters, which makes it impossible to apply meaningful fully-crossed or nested multiple facet designs directly. Scoring plans like this are designed for cost efficiency and commonly used by testing firms.

Based on the new method all available maximal fully-crossed subsets of the response matrix are selected to perform a two-facet fully-crossed G-study for each of them. Loosely speaking, maximal fully-crossed subsets are those for which a fully crossed G-study design is applicable. Data for this study was obtained from the State of Ohio third grade Reading Achievement Test administered in 2003 October, which included seven constructed-response items, out of which a small portion was double-scored by 100 trained, randomly assigned raters. Generalizability analyses were carried out for 98 subsets. Variance components and generalizability coefficients were estimated and compared for each of them.

Results show consistently excellent rater performance. The analysis also implies that the same conclusion on rater behavior could have been obtained from a much smaller (and more economical) double-scored sample.

---

Tamas Antal  
The Ohio State University

Judit Antal  
American Institutes for Research

Stability of Some Item Response Theory Monte-Carlo Goodness Fit Indices

The goal of the paper presentation is to report the results of an extensive stability study of some Item Response Theory Monte-Carlo (MC) goodness of indices introduced earlier by the authors. By the nature of their definitions the values of the MC indices in question contain a small ambiguity. The natural question is then the dependence of this variability of the MC indices on the specific circumstances. More precisely, we investigated the dependence of this uncertainty on the size of the response matrix, on the actual value of the goodness of fit index and, most importantly, on the size of the MC sample used to calculate the fit indices. The different fit indices are also compared with one another.

The results indicate, that the MC fit indices show a remarkable stability with respect to the size of the MC sample. This observation makes it possible to use the proposed indices in real test situations, since using an extremely small MC sample of 1000 response matrices already gives acceptable precision and the computational time are not significantly longer than that of the only asymptotically understood "usual" goodness of fit indices.

---

Ron Armstrong  
Rutgers University

Dmitry Belov  
Psychometric Research Division, Law School Admission Council

A Method for Determining Multiple Non-Overlapping Linear Test Forms

A valuable resource for any testing agency is its item pool. This paper considers the problem of assembling multiple non-overlapping linear test forms from an item pool. Knowing the maximum number of linear test forms supported by an item pool assists with monitoring the pool and guides the development of new items. Test forms are assembled by solving a sequence of integer programming problems. If specified conditions are satisfied, the procedure assures the assembly of the maximum number of linear forms supported by the pool. If the conditions are not satisfied, multiple linear forms are assembled and an upper bound and a lower bound on the maximum number of forms are provided. All integer programming problems are solved with commercially available software.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Dmitry Belov and Alexander Weissman  
Psychometric Research Group, Law School Admission Council

## Combinatorial Analysis For Determining Item Pool Usability In Computerized Adaptive Testing

In a computerized adaptive test (CAT), the extent to which test specifications and information constraints are met is necessarily dependent on the characteristics of items in the pool. Thus, the number of non- or partially-overlapping CATs that meet such constraints for examinees within a specified ability range is a critical determinant of item pool usability. The usability question may be addressed through combinatorial analysis of a CAT item pool, where individual CAT forms are assembled through Monte Carlo sampling. The approach demonstrated here involves two steps for a given ability range: (1) generate a large set of overlapping CAT forms satisfying all test specifications and the particular information constraints associated with this ability range; (2) extract the maximum subset of non-overlapping or partially-overlapping CAT forms. This second step is formulated and solved as a maximum set packing problem. Applying this approach to a collection of ability ranges will estimate how well each ability range is represented in the pool. In addition to providing lower bounds on the maximum number of available CATs, the resulting estimates can guide test developers in the design of new items, a crucial component for maintaining an item pool. Computer experiments with operational item pools and test specifications are presented.

Keywords: computerized adaptive testing; combinatorial optimization; set packing; test assembly; item pool analysis

---

Coen A. Bernaards  
AMC Cancer Research Center

Thomas R. Belin  
UCLA

Joseph L. Schafer  
The Pennsylvania State University

## Robustness of A Multivariate Normal Approximation For Imputation of Incomplete Binary Data

Multiple imputation has become easier to perform with the advent of several software packages that provide imputations under a multivariate normal model, but imputation of missing binary data remains an important practical problem. Here, we explore three alternative methods for converting a multivariate normal imputed value into a binary imputed value: (1) simple rounding of the imputed value to the nearer of 0 or 1, (2) a Bernoulli draw based on a "coin flip" where an imputed value between 0 and 1 is treated as the probability of drawing a 1, and (3) an adaptive rounding scheme where the cutoff value for determining whether to round to 0 or 1 is based on a normal approximation to the binomial distribution. We perform simulation studies on a data set of 206,802 respondents to the California Healthy Kids Survey, where complete cases are viewed as a finite population and incomplete cases provide a basis for imposing missing-data patterns. Frequently, we found satisfactory bias and coverage properties, suggesting that approaches such as these that are based on statistical approximations are preferable in applied research than either avoiding settings where missing data occur or relying on complete-case analyses.

---

Shelley A. Blozis  
University of California, Davis

## A Second-Order Structured Latent Curve Model for Normal Repeated Measures

This paper proposes a second-order structured latent curve model for the study of change in a latent variable. In a structured latent curve model, the mean response of an observed variable measured at multiple time points is assumed to follow a smooth, nonlinear function. A first-order Taylor polynomial taken with regard to the mean function defines elements of a factor matrix that may include parameters that enter nonlinearly. Stochastic coefficients combined with the factor matrix produce individual latent curves that need not follow the same form as the mean curve. This paper considers a model for which multiple indicators of a construct are available, and the latent variable is assumed to follow a nonlinear function that is invariant to a constant scaling factor. An example is provided.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Todd E. Bodner  
Portland State University

## The Level-of-Imputation Question for Massively Missing Composite Variable Data

The present paper explores the use of multiple imputation for handling missing composite variable data when the items constructing the composite variable are completely missing for some units (but data from other variables not in the composite are observed). In this case, one can either impute scores at the item level or the composite variable level. Does one's choice make a difference when the composite is used in analyses (e.g., multiple regression) with other data? A simulation study was conducted to explore this question. We explored the relationships among three variables (X, Y, and Z) in the presence of completely missing data for composite variable X (10 items) for some cases. From a known covariance matrix, we drew random multivariate normal data samples of size 100, deleted between 5% and 50% of the data for the composite variable under MCAR, and then imputed the missing data either at the item level or at the composite level to create "complete" data sets which were then analyzed. Our results did not find any important differences (e.g., in statistical power) across the two imputation strategies. However, as the percentage of cases with massively missing data increased, sizable drops in statistical power were observed.

---

Daniel Bolt and Andrew Mroch  
University of Illinois, Urbana-Champaign

## Application of a Mixture IRT Model for Cognitive Diagnosis

A system of constraints is applied to a mixture IRT model based on the cognitive skills required by test items so as to permit student-level diagnosis of item response patterns. A Bayesian approach to estimation of the model subject to these constraints is presented and applied using a mixed number subtraction dataset. Using several model comparison criteria, the current approach is evaluated against previous models for diagnosis that have been applied to the same dataset. Extensions of the model to incorporate collateral information are discussed.

---

Derek Briggs  
University of Colorado, Boulder

Alicia Alonzo, Cheryl Schwab and Mark Wilson  
University of California, Berkeley

## Modeling Partial Information in Multiple Choice Items

In this presentation we describe the development, analysis and interpretation of a novel item format we call Ordered Multiple-Choice (OMC). A unique feature of OMC items is that they can be linked to a model of student cognitive development for the construct being measured. Each of the possible answer choices in an OMC item is linked to developmental levels of student understanding, facilitating the diagnostic interpretation of student item responses. OMC items seek to combine the validity advantages of open-ended items with the efficiency advantages of multiple-choice items. On the one hand, OMC items provide information about the developmental understanding of students that is not available with traditional multiple-choice items; on the other hand, this information can be provided to schools, teachers and students quickly and reliably, unlike traditional open-ended test items.

Test forms with OMC items were administered to 140 and 156 fifth and eighth grade students, respectively, in the spring of 2002. Both samples were also administered test forms with traditional multiple-choice (MC) items covering subjects such as earth science, life science, physical science and scientific inquiry. We model the data with both the Ordered Partition Model (Wilson & Adams, 1993) and an extension of the nominal categories model (Thissen & Steinberg, 1986) that we call the Latent and Nominal Categories Model (LNCM). We illustrate two ways of interpreting student responses to the OMC items. In the first, one can evaluate the pattern of an individual student's understandings across all items on the test. In the second, one can evaluate classroom (or school) understandings for each item on the test.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Derek Briggs  
University of Colorado, Boulder

Mark Wilson  
University of California, Berkeley

Generalizability Theory in Item Response Modeling

On the surface, Generalizability Theory (G Theory) and Item Response Theory (IRT) appear incompatible. Robert Brennan, for example, writes “Generalizability Theory is primarily a sampling model, whereas Item Response Theory is principally a scaling model.” In this presentation we introduce an approach we call Generalizability Theory in Item Response Modeling (GIRM). The GIRM approach essentially incorporates the sampling model of G Theory into the IRT “scaling model” by making distributional assumptions about the relevant measurement facets. Given these assumptions, and taking advantage of the flexibility of Markov Chain Monte Carlo (MCMC) estimation methods, it becomes possible to estimate G Theory variance components concurrently with traditional IRT parameters. We show how G Theory and IRT can be linked together, in the context of a single facet measurement design with binary items, and a multi-faceted design with polytomous items. Using simulated data and the software WinBUGS, we demonstrate how the GIRM approach can produce results equivalent to those from a standard G Theory analysis. Some advantages of the GIRM approach relative to IRT alone, or the sequential use of IRT and G Theory, are discussed.

---

Michael W. Browne and Guangjian Zhang  
The Ohio State University

Exploratory Factor Analysis of Lagged Correlation Matrices

A model for the factor analysis of lagged correlation matrices that incorporates a stationary VARMA time series for common factors is presented. It allows rotation of the factor matrix together with appropriate transformations of parameter matrices associated with the time series. Initial starting values and an iterative process for obtaining ordinary least squares estimates are provided. An example is given.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Claus H. Carstensen  
Christian-Albrechts-Universität zu Kiel

Andreas Voss  
Universität Hamburg

Wilfried Bos  
Universität Hamburg

## Reading Comprehension of Primary School Students - A Unified Cognitive Process or Interacting Component Processes?

For a mature reader's point of view, reading appears to be an instantaneous and unified cognitive process. Under the scrutiny of research, however, this impression is not accurate. From a cognitive psychology perspective, several interactive component processes including word, sentence and text levels form the act of comprehension rather than one single unified process.

In 2001, the International Association for the Evaluation of Educational Achievement assessed the reading achievement of fourth grade students in the Progress in International Reading Literacy Study (PIRLS). The framework of this study distinguishes between four different processes of reading comprehension:

- (1) to focus on and retrieve explicitly stated information,
- (2) to make straightforward inferences,
- (3) to interpret and integrate ideas and information, and
- (4) to examine and evaluate content, language and textual elements.

These processes may not be observed independently of each other in an examination. Process (1) "retrieving information" rather defines a precondition for a successful application of one of the other three processes. The second process may as well be seen as a prerequisite for the third one. In this paper we present the analysis of the interrelations between these comprehension processes for the German primary school students using ConQuest (Wu, Adams & Wilson, 1998, 2004) as tool for multidimensional IRT modelling.

The favoured model distinguishes two dimensions, one that combines (1) "retrieving information", (2) "straightforward inferences" and (3) "interpreting and integrating ideas" and a second dimension that combines (1) "retrieving information" and (4) "examine and evaluate content, language and textual elements". It is assumed that the items load on different dimensions at the same time (Within Item Multidimensionality WIM). Alternative models analysed are a) a unidimensional model, b) a three dimensional WIM model based on the processes (2), (3) and (4), each combined with process (1) and c) a four-dimensional subtest model.

---

Hua-Hua Chang  
University of Texas, Austin

Jiahe Qian  
Educational Testing Service

Pei-hua Chen and Ying Cheng  
University of Texas, Austin

## Adjustment of Bib Data for Dif Testing

National Assessment of Educational Progress (NAEP) is a giant project and requires a variety of techniques of a complexity beyond any simple survey. Differential Item Functioning (DIF) analysis is one of the most important steps in NAEP data analysis. Due to the complex balanced incomplete block (BIB) design, the current DIF procedures are based on matching examinees with their total scores on pooled booklets. Since difficulty levels often vary among different booklets, examinees with same booklet scores may not be the same in ability. Such matching may cause misplacement and increase measurement errors. Clearly, establishing a common matching criterion across pooled booklets becomes important. In this project we propose a simple approach of using the classical *common item equating* to ensure compatibility in forming matching variables for the NAEP complex BIB design. We will focus on three specific objectives:

1. Revising the method of forming matching variable so that the matching will be based on transformed booklet scores after common block equating/linking.
  2. Evaluating the modified NAEP SIBTEST procedure and the NAEP Standardization procedure by using both simulated and real NAEP data.
  3. Deriving some analytic results to provide theoretical evidence about these improvements.
-

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Po-Hsi Chen  
National Yunlin University of Science and Technology

Wen-chung Wang  
National Chun Cheng University

The Influence of the Number and Magnitude of the Testlet on the Computerized Adaptive Testing

This research used the multidimensional random coefficients multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997) to deal with the testlet effect in the computerized adaptive testing (CAT). Two variables, number and magnitude of the testlet, had been used in this research in order to know their influences on the ability estimation of CAT. Results indicated that the ignorance of the testlet effect will make the variance of the latent trait smaller incorrectly and cause a fake and higher reliability. These influences are more serious when the number and the magnitude of the testlet are large. It suggested that when testlet effect is existed in CAT condition, testlet response model, but not item response model, should be used in the ability estimation and item selection procedure in order to prevent incorrect inference about the latent trait. Applications of the use of MRCMLM on the CAT with testlet effect had been addressed in this research.

Keywords: computerized adaptive testing, item response model, testlet, testlet response model

---

Ying Cheng and Hua-hua Chang  
The University of Texas, Austin

Yi Qing  
Harcourt Assessment, Inc.

Two-Phase Item Selection with Realistic Content Balancing Constraints in Computer Adaptive Testing

Computerized adaptive testing (CAT) has become a popular mode of educational assessment in the United States. The major advantage of CAT is that it provides more efficient latent trait estimates ( $\theta$ ) by selecting items that fit examinees' trait levels for administration. Therefore, item selection is one of the core issues of CAT, which encompasses topics such as measurement precision, test security, item pool utilization, and content balancing etc. A CAT item-selection procedure, BAS, which is characterized by a-stratification with b-blocking, can improve exposure control and item pool utilization while maintaining high measurement precision. It can also be readily applied when the content balancing requirement is simplified, i.e. when the number of items from each content area is fixed. In real practice, however, the number of items from a certain content area is often constrained by lower-and upper-bounds. In this paper, three methods are proposed to address the realistic content balancing problem: (1) BASRC, which can be viewed as an extension of the STR\_C method; (2) Two-Phase content balancing with Multinomially distributed content specification at the second stage (TPM); (3) Two-Phase item selection with "Free" content specification at the second stage (TPF). Monte Carlo studies are conducted using the three methods and the results are compared. It is demonstrated that all the three methods can achieve realistic content balancing. Among them TPF is the best in terms of exposure control and item pool utilization.

Index terms: realistic content balancing, constraints, exposure control, a-stratify, item selection

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Christopher W.T. Chiu  
Law School Admission Council (LSAC)

Visualizing Standard Error of Equating Using the Bootstrap and the SEER Methods

Test equating is a technique to adjust for minor differences in scores appearing in different test forms used in psychological and educational testing. One of the goals in test equating is to ensure that test scores are comparable across test forms and/or test administrations. The quality or the accuracy of test equating can be evaluated through the standard error of equating. While analytical methods are commonly used to derive or to estimate standard error of equating, the bootstrap method is an alternative when analytical solutions are too complicated to be derived or are nonexistent. Kolen and Brennan (1995) stipulated the general steps for using the bootstrap method to examine standard error of equating. In the current study, the researcher takes a step further and develops a graphical procedure (SEER for Equating) to visually estimate standard error of equating. The graphical procedure can (a) dynamically show the intermediate steps of the bootstrap method and thus makes the bootstrap resampling process transparent to aid the equating process for computerized adaptive testing, (b) depict multiple sources of statistical information in one display ---- including, for example, the score distributions, the equated and the unequated scores of individual test takers, conversion tables, and conditional standard error of equating, (c) provide a systematic scheme to convert data displayed in graphics into statistical indices by ways of elementary matrix manipulations, and (d) visually compare the results and impact of multiple equating procedures. A simulation based on the Law School Admission Test (LSAT) score scale is used to illustrate the graphical procedure.

---

Sy Miin Chow and John R. Nesselroade  
University of Virginia

A Monte Carlo Comparison of Methods for Fitting Nonlinear Dynamic Models

Over the last few decades, developments in the field of dynamical systems have inspired researchers to formulate dynamic models that could map onto observable psychological phenomena. However, most contemporary nonlinear analytic tools are exploratory in nature, and researchers are very much in need of model fitting tools that can provide a direct linkage between a mathematical model, and a set of empirical data. Even though maximum likelihood is by far one of the most dominant approaches to model fitting, it is not well suited for fitting nonlinear dynamic models. In most instances, complex nonlinear constraints will have to be imposed on the model to yield accurate parameter estimates (see e.g., Jöreskog & Yang, 1996). In this study, I evaluate the applicability of an extended Kalman filter approach to fitting nonlinear dynamic models. Monte Carlo simulations are used to generate data with different degrees of missingness, and different magnitudes of measurement errors. I then assess the strengths and weaknesses of this approach using data simulated from (1) a static (no change) model, (2) a linear change model, and (3) a nonlinear predator prey model with a quadratic and an interaction term. The accuracy and computational efficiency of this approach in comparison to other conventional model fitting approaches are discussed.

---

Hans Colonius  
Oldenburg University

Ehtibar N. Dzhafarov  
Purdue University

Multidimensional Scaling of Fechnerian Distances

Fechnerian metrics (whether on discrete or continuous sets of objects) belong to a very general class of so-called intrinsic metrics. They need not be Euclidean, city-block, or other Minkowskian metrics traditionally used in MDS. Once Fechnerian distances are computed from discrimination probabilities for a finite set of objects, however, they can be used to isometrically immerse the configuration of objects in a low-dimensional Euclidean space, by applying a procedure of metric MDS to Fechnerian distances. The results of such analyses are compared with those of nonmetric MDS applied directly to discrimination probabilities. Nonmetric MDS of discrimination probabilities can only be used as a crude approximation, due to the violations of the basic assumptions of MDS: the discrimination probabilities are generally asymmetric, and the probability of discriminating an object from itself (more generally, from its point of subjective equality) is not a constant.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Samuel Copt and Maria-Pia Victoria-Feser  
University of Geneva

## High Breakdown Inference in the Mixed Linear Model

Mixed linear models are used to analyse data in many settings. These models have in most cases a multivariate normal formulation. The maximum likelihood estimator (MLE) or the reweighted MLE (REML) are usually chosen to estimate the parameters. However, the latter are based on the strong assumption of exact multivariate normality. Welsh and Richardson (1997) have shown that these estimators are not robust to small deviations from the multivariate normality. This means in practice for example that a small proportion of data (even only one) can drive the value of the estimates on their own. Since the model is multivariate, we propose in this paper a high breakdown robust estimator for very general mixed linear models that include for example covariates. This robust estimator belongs to the class of S-estimators (Rousseeuw and Yohai, 1984) from which we can derive the asymptotic properties and inference tools. In particular, we propose robust procedures for testing contrasts and multivariate hypothesis. We also use the robust estimates as a diagnostic tool to detect outlying subjects. We discuss the advantages of this estimator compared to other robust estimators proposed previously and illustrate its performance with simulation studies and real data.

---

Anna Villa T. Dagohoy and Cees A.W. Glas  
University of Twente

## Lagrange Multiplier Person Fit Tests for Polytomous IRT Models

A class of person fit tests based on Lagrange multiplier tests is presented for three item response theory models for polytomous items: the generalized partial credit model, the sequential model and the graded response model. It is shown that these tests can also be used in the framework of multidimensional ability parameters. A simulation study of the power and Type 1 error rate is presented. Further it is investigated to what extent the three models can be distinguished using person fit statistics. Finally, an example using data from NEO Personality Inventory-Revised is presented. This test battery consists of a set of sub-scales to which the multidimensional versions of the three IRT models for polytomous items are fitted. Person fit in to the subscales is assessed with and without using auxiliary information from the other subscales.

Key words: item response theory, person fit, model fit, multidimensional data

---

Tim Davey and Elizabeth Stone  
Educational Testing Service

## A trend model for monitoring item security under continuous testing

Although continuous testing offers practical advantages to both examinees and testing organizations, it also entails considerable risk. The concern is that examinees can take advantage of the fact that items are exposed repeatedly across testing occasions and so breach security. Ideally, monitoring all items for changes in their performance characteristics over time could identify dangerously over-exposed items. Unfortunately, effectively measuring gradual, subtle change presents some formidable statistical challenges.

Most monitoring statistics take either of two basic approaches, both of which have their drawbacks. The first is to work continuously and cumulatively, basing results at any time on all data collected to date. The down side is that change will not be recognized until the proportion of more recent, nonconforming responses outweighs the proportion of older, more conforming responses. The second approach is to work discretely, conducting a series of analyses on batches of data gathered within a given time window. Although this should allow gradual change to be more readily detected, the smaller samples available at each point rob the statistics of power.

We propose an approach to monitoring change that is based on modeling trends in item performance over time. The trend model makes the reasonable assumption that with repeated exposure and widespread prior knowledge, an item will gradually become easier and less discriminating. However, the rate and manner in which items manifest these expected changes is more difficult to assert. As a starting point, we will assume that items change in the simplest possible way, as a linear function of the number of occasions on which an item has been administered. More formally, we will assume that items generally accord with a logistic IRT model, but that the parameters of the model change over time.

A Bayes approach will be described for estimating the "trend" parameters for each item across its multiple administration occasions. The trend model will be evaluated through application to both simulated and actual datasets. Strategies for testing whether a given item has changed sufficiently to warrant action will also be discussed.

---

Paul de Boeck  
K.U.Leuven

#### Double Structure Item Response Models

Structural equation models are single structure models because they explain the structure of the variables based on one mode of the data (the person mode). For three-mode data, for example, person  $\times$  situation  $\times$  variable data, two different modes are available to describe the structure of the variables (the person mode and the situation mode), so that a double structure model can be defined. The model we have formulated for such a double structure is a logistic mixed model for ordered-category data (an item response model). For each of the observed variables ( $k=1, \dots, K$ ), two different latent variables are defined: a latent person variable,  $\theta_k$ , and a latent situation variable,  $\beta_k$ , and structural relations are formulated between the latent person variables and between the latent situation variables. These structures explain the relations between the variables over persons ( $p=1, \dots, P$ ), and over situations ( $s=1, \dots, S$ ), respectively. The basic formula for binary data is  $\text{logit}(Y_{psk}=1) = \theta_{pk} + \beta_{sk}$ .

A  $K \times K$  matrix  $\Gamma$  defines the latent relations between the  $\theta$ s, and a  $K \times K$  matrix  $\Lambda$  defines the latent relations between the  $\beta$ s. The exogenous latent person variables are normally distributed random variables, and the endogenous latent person variables include a normally distributed error term. The latent situation variables can be treated in the same way, which would require random item effects, but in this presentation only fixed effects will be used. The endogenous latent situation variables include an intercept. The model is based on a multidimensional extension of the MIRID (Butter, De Boeck & Verhelst, 1998), and a new view on that model. The double structure models (DSMs) will be illustrated with a study on anger and irritation. The results show that in both structures ( $\Gamma$  and  $\Lambda$ ) anger and irritation are based on both frustration and an antagonistic action tendency, but that anger relies more on the latter, whereas irritation relies more on the former.

---

Paul de Boeck  
K.U. Leuven

#### Explanatory Measurement: A Case Study of Modeling Coping With Stress

Coping with stress is often studied from a dispositional point of view, and is as such measured through questionnaires and sum scores derived from these. Either the subscales are constructed a priori (deductive method) or they are based on factor analysis (inductive method). On the other hand, coping with stress is also considered to be situation dependent, and therefore questionnaires have been constructed on how people cope with stress in specific situations—these questionnaires lead to individual-differences measures, but of a situation-specific kind. Whether a dispositional or a situational perspective is taken, the measures are of a descriptive kind. They do not explain coping responses. Explanation often follows in a next step when the measures are correlated with external variables.

As an alternative we constructed a questionnaire with two side-by-side questions for each situation. For each of 12 situations it was asked how stressful the experience would be, and what the coping response would be, so that the coping response can be related to the degree of stress. For example, based on a biological theory on stress (Taylor et al., 2000) we expect that, for females but not for males, a tending response (caretaking) is positively related to the degree of stress. The data are analysed with a rating scale model, complemented with two linear functions: one that links the situational tending parameter to the corresponding situational stress parameters, and one that links the two underlying person parameters. Using this model we do not just describe the degree of tending in response to stress (as expressed in the situation and person parameters) but the degree of tending is also linked, over both situations and persons, to how stressful situations are experienced. This linkage is what makes the measurement explanatory.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Jimmy de la Torre  
Rutgers University

Improving the Accuracy of Ability Estimates through Simultaneous Estimation and Incorporation of Ancillary Variables

In typical testing situations, multiple tests that measure correlated abilities are administered to examinees at the same time. In addition to item responses, background information about the examinees is routinely available. It has been shown that more accurate estimates can be obtained when all the tests are used simultaneously in estimating the abilities, and more precise inferences can be made when background information are used in estimation. The primary purpose of this paper is to integrate both sources of additional information into one single model in improving ability estimation. Using a simulation study, the impact on ability estimates of the number of tests, the length of tests, the strength of relationship between the background information and the underlying abilities, and the degree of correlation between the abilities were studied. Compared to a baseline method where abilities are estimated using a single test without considering background information, this study shows that better ability estimates are obtained when multiple short tests that measure highly correlated abilities, and background information that correlate highly with the abilities are available. Moreover, results indicate that using both sources of information simultaneously yields better estimates than using either source singly.

---

Jimmy de la Torre  
Rutgers University

Jeffrey A. Douglas  
University of Illinois

Model Evaluation and Selection in Cognitive Diagnosis: An Analysis of Fraction Subtraction Data

Three models for cognitive diagnosis are studied, and are illustrated with an application to fraction subtraction data. The objective of each of these models is to classify examinees according to their mastery of skill assumed to be required for fraction subtraction. We consider the DINA model and two versions of the NIDA models. For each of these models, the joint distribution of the indicators of skill mastery is modeled using a single continuous higher-order latent trait, to explain the dependence on the mastery of distinct skills. Several techniques for comparing models and assessing goodness-of-fit are discussed and are implemented with the fraction subtraction data with an aim of selecting the best of three models for this application. Markov chain Monte Carlo algorithms for parameter estimation and model fitting are extended and developed. Simulation results are presented to examine the performance of these algorithms.

---

Mark de Rooij  
Leiden University

The Analysis of Change, Newton's Law of Gravity and The RC(M) – Association Model

Newton's law of gravity states that the force between any two objects in the universe is equal to the product of the masses of the two objects divided by the square of the distance between the two objects. It will be shown that this law is very well applicable to the analysis of social change where the number of people changing their behavior from category a to b is a measure of force and the goal is to obtain estimates of mass for the two categories and an estimate of the distance between them. In order to provide a better description of the data dynamic masses and dynamic positions will be introduced. It will be shown that this generalized law of gravity is equivalent to Goodman's RC(M) -association model.

Keywords: Euclidean distance; two mode distance; distance association model; longitudinal data; square contingency tables.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Hui Deng  
The College Board

Tim Ansley  
The University of Iowa

An Investigation of Stratified and Maximum Information Item Selection Procedures in Computer Adaptive Testing

The maximum Fisher information procedure (F) is a commonly used CAT item selection procedure; it leads to great test efficiency with very unbalanced item usage. The  $\alpha$ -stratified multistage CAT (STR) was developed to remedy the item usage problem with F, and was found to effectively balance item usage with reduced test efficiency. To address the efficiency loss, a refined stratification procedure was proposed that allows unequal item exposure across strata (USTR). This study compared the three procedures, along with completely random item selection (RAN) with respect to test efficiency and item usage through CATs simulated under test conditions varied in terms of content and exposure constraints and item selection space.

The results showed that F had an apparent efficiency advantage over STR and USTR under unconstrained item selection with poor item usage. USTR reduced error variances relative to STR under various conditions, with small compromises in item usage. Compared to F, USTR achieved a similar efficiency level with improved item usage when items were selected under exposure control and the item selection space was restricted by long tests or a stringent security criterion. The results provide implications for choosing an appropriate item selection procedure in applied settings.

---

Lou DiBello  
Educational Testing Service

Jon Templin and Bob Henson  
University of Illinois

Applications to Operational Tests—Next Generation TOEFL

A description of the evidence centered design of the Next Generation TOEFL, development of a framework for skills to be reported, and application of Arpeggio for modeling and scoring.

---

John R. Donoghue  
Educational Testing Service

Estimation of Proficiency Distributions from Matrix Sample Assessments: Context and Current Procedures

Large-scale educational surveys differ from more traditional student assessments in a variety of ways, including the main target of inference. Accurate estimates of group and population statistics are the focus of the analysis of large-scale educational survey data and the methodologies used to achieve that goal differ from the more traditional methods focused on precise estimates of individual proficiency. This presentation will provide a technical discussion of the current methodology used to estimate proficiency distributions in these assessments.

---

Jeff Douglas, Louis Roussos, and Bill Stout  
University of Illinois

Introduction to Profile Scoring

A general framework is presented for student profile scoring. A short description is provided of the use of Bayes Inference Networks, and Evidence Centered assessment design. The focus is on the Arpeggio system for profile scoring.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Amy R. Dresher, John R. Donoghue, and John Mazzeo  
Educational Testing Service

Comparing Bias, Precision, and Stability of Estimation of Group-Level Statistics Based on High Dimensional and Low Dimensional Population Models

Traditionally, all variables and principal interactions are included in a single population estimation model, extracting the subgroup performance of all subgroups (e.g. females, Hispanics, etc) from that single model. Alternatively, each subgroup level could be estimated in its own population estimation model, thus providing unbiased estimates of all variables, not just those specified in the conditioning model. A simulation study is currently being conducted which compares the estimates of both methods to the true sample values, thus allowing the bias and RMSE to be compared across the estimation procedures.

---

Ehtibar N. Dzhafarov  
Purdue University

Hans Colonius  
Oldenburg University

Fechnerian Scaling of Discrete Object Sets

Fechnerian Scaling of Discrete Object Sets (FSDOS) is a method that imposes a metric on a discrete set of objects based on the probabilities with which these objects are distinguished from each other (i.e., judged to be different) by an individual or group of respondents. The method does not presuppose a monotone relation between discrimination probabilities and distances, the discrimination probabilities may be asymmetric, and the probability of discriminating an object from itself (more generally, from its point of subjective equality) may not be the same for all objects. The only requirement is Regular Minimality, introduced by present authors in the context of Generalized Fechnerian Scaling as the fundamental property of discrimination probabilities. In FSDOS Regular Minimality means that each row in the matrix of pairwise discrimination probabilities should contain a minimal entry which should also be minimal in its column.

---

Michael C. Edwards and David Thissen  
The University of North Carolina, Chapel Hill

Defining and Finding Optimal Designs for uMFS CATs

In a multi-stage CAT, the examinee responds to some small fixed set ("block") of items, after which adaptation takes place in the choice of the next block of items to be administered. An advantage of this type of CAT design over traditional item-selection systems is that the blocks of items are assembled in advance, so many aspects of test assembly can be done by test specialists (likely computer-assisted) at leisure instead of by real-time item-selection algorithms. In addition, it is possible to avoid the issue of exposure control by designing a multi-stage testlet CAT that yields (nearly) uniform item exposure. Accomplishment of this goal would control "absolute exposure for expected candidate volumes," as opposed to exposure rates.

Optimal assignment of items to blocks hinges on the definition of "optimal" and the selection of an algorithm to accomplish this assignment. The focus of the current research is on determining components for an objective function and the choice of an optimization algorithm.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Wilco H.M. Emons  
Tilburg University

## Person-Fit Analysis for Polytomous Items in Personality Assessment

Personality indicators increasingly play an important role in individual decision-making and diagnosis in, for example, health risk assessment, jurisdiction, and job selection procedures. Accumulated evidence showed that response faultiness may seriously threaten the validity of the outcomes of a personality assessment. Examples include malingering, faking, and carelessness. Person-fit analysis is a psychometric method for the detection and diagnosis of response faultiness as reflected in the pattern of item scores. Most person-fit approaches assume dichotomous item scoring, and the properties of the approaches are investigated mainly in the context of cognitive assessment and educational testing. For example, the applicability to identify testees who suffered from severe test anxiety. Personality assessment is different from cognitive assessment in that personality items measure typical performance, whereas cognitive items measure maximum performance. In addition, personality questionnaires often use polytomous item scoring (e.g., Likert scales). In this presentation, I will consider generalizations of person-fit approaches to polytomous items in the context of personality assessment. These person-fit approaches are based on nonparametric and parametric item-response theory models, which are increasingly used for evaluating the psychometric properties of personality tests. Results from simulation studies, as well as real data examples from health psychology, will be presented.

---

Elena A. Erosheva  
University of Washington

## Bayesian Estimation of a Latent Trait Model with Bounded Continuous Latent Variables

Given discrete responses, the assumption of a bounded latent continuum can be plausible in many applied problems. Consider, for example, an opinion survey where we expect extreme responses from pro- and anti- groups as well as mixed responses from less opinionated individuals. The Grade of Membership (GoM) model, developed in the context of medical diagnosis (Woodbury, Clive and Garson 1978), assumes that individuals have mixed membership over extreme categories. The latent variables in the GoM model are nonnegative individual membership scores which add up to unity. The GoM model can also be viewed as a constrained latent class model and as a factor analysis model for discrete data.

Membership scores and extreme categories are usually unknown. This paper presents a Bayesian approach for estimation of the GoM model parameters treating membership scores as latent realizations of random variables. Although the standard GoM model has a hierarchical structure, full conditional distributions are intractable. Using a constrained latent class representation of the GoM model, we obtain a posterior distribution of the model parameters by augmentation the data with the latent class indicators. Examples of analyses of simulated and real data are presented.

---

Xin Feng  
CTB/McGraw-Hill

Zhiliang Ying  
Columbia University

## Detection of Differential Item Functioning in Computerized Adaptive Testing Using Measurement Error Models

Differential item functioning (DIF) is an important issue in large scale standardized testing. DIF refers to the unexpected difference in item performances among groups of equally proficient examinees. Its presence could seriously affect the validity of inferences drawn from a test. This article addresses DIF analysis in the context of computerized adaptive testing (CAT). In a CAT, DIF item may be more consequential and more deteriorative.

We propose simultaneous implementations of online calibration and DIF testing. Under any specific parametric IRT model, we can use the (online) estimated latent traits as covariates and fit a nonlinear regression model to each of the two groups. Because of the use of the estimated, not the true ability, the regression fit has to adjust for the covariate "measurement errors". We develop two bias-correction methods for the Rasch and 2PL models using asymptotic expansion and conditional score theory. After correcting the bias caused by measurement error, one can perform a significance test to detect DIF with the parameter estimates for different groups. Extensive simulation studies show that the resulting methods perform well. More general method for 3PL and partial credit models is also discussed.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Tron Foss, Karl G. Jøreskog, Ulf Henning Olsson  
Norwegian School of Management BI

Does the Satorra-Bentler Scaled Chi-Square Statistic Approximate Zero When the Kurtosis Approimates Infinity?

In this study we demonstrate how the Satorra-Bentler Scaled Chi-square statistic is affected by (excessive) kurtosis in the observed data.

More specifically we address how different levels of univariate kurtosis affect the chi-square values (and therefore fit-indices) for misspecified factor models. We also outline a formal proof that the probability limit  $F_0$  of  $\hat{F}$  tends to zero when the kurtosis tends to infinity. This will be valid for a whole class of structural equation models.

Keywords: Scaled Chi-square statistic, asymptotic covariance matrix, kurtosis.

---

Laurence E. Frank and Willem J. Heiser  
Leiden University

Statistical Inference in Feature Network Models and Additive Trees

Feature Network Models (FNM) are a particular class of graphical structures that represent proximity data in a discrete space while using the same formalism that is the basis of least squares methods used in multidimensional scaling. Additive trees are a special case of FNM because the distances between  $n$  objects can be expressed in terms of a specific set of  $2n-3$  distinctive features.

Existing methods to derive a network model from empirical data only give the best fitting network and yield no standard errors for the parameter estimates. The additivity properties of networks make it possible to consider both FNM and additive trees as univariate multiple linear regression problems with positivity restrictions on the parameters, which forms a starting point for statistical inference. Theoretical standard errors as well as empirical standard errors have been obtained for the parameters of the FNM. It will be shown that the same method for statistical inference can be applied to additive trees.

Keywords: Additive trees, distinctive features model, graphs, Monte Carlo simulation, nonnegative least squares, statistical inference.

---

Furong Gao  
CTB/McGraw-Hill

Empty Bubble: Which Test Form Should the Response Be Scored With?

The Problem arises when more than one test form is administered and generic answer sheet is used for all the forms. If a student did not indicate which form he took by filling proper bubble for the test form, method needs to be developed to score the student responses. This paper compares several IRT-based methods to assign test form to students who left the bubble for test form blank, including some commonly used person-fit statistics. Simulation is conducted to calculate error rate for each method.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Cees A.W. Glas  
University of Twente

## Violations of Ignorability in Computerized Adaptive Testing

Using auxiliary information and allowing item review in computerized adaptive testing produces a violation of the ignorability principle for missing data (Rubin 1976) that may bias parameter estimates in IRT models. However the violation of ignorability does not automatically lead to bias. In this paper, two situations are distinguished.

1. Estimation of the proficiency parameters is computerized adaptive testing using auxiliary information about proficiency and allowing item review, where the item parameters are considered known. Both analytically and through simulation studies it is shown that the violation of ignorability does not lead to a gross inflation of bias.
2. Calibration of item and population parameters using maximum marginal likelihood estimation. Through simulation studies it is shown that violation of ignorability due to the use of un-modelled auxiliary information about proficiency and allowing item review does result in bias. An analytical explanation of the result is given.

Keywords: computerized adaptive testing; ignorability; item calibration; marginal maximum likelihood; missing data; sampling design.

---

Cees A.W. Glas and Jean-Paul Fox  
University of Twente

## Analysis of Variance and Regression Using Multilevel IRT

It is shown that measurement error in predictor variables can be modeled using item response theory (IRT). The predictor variables, which may be defined at any level of an hierarchical regression model, are treated as latent variables. The approach entails the definition of a multilevel linear model, where latent variables from IRT measurement models are entered either as dependent or as independent variables. The resulting model is the so-called multilevel IRT model (MLIRT model, Fox & Glas, 2001, 2002, 2003). The normal ogive model is used to describe the relation between the latent variables and dichotomous observed variables, which may be responses to tests or questionnaires. It will be shown that the multilevel model with measurement error in the observed predictor variables can be estimated in a Bayesian framework using Gibbs sampling. An example using real data is given (Shalabi, 2002). The data were a cluster sample of 3,384 grade 7 students in 119 schools. At student level the variables were gender, SES, and IQ. At school level: leadership, school climate and mean-student-IQ. The dependent variable was a mathematics achievement test. The 2-parameter normal ogive model was used to model the responses on the leadership and school climate questionnaire and the mathematics test. It is shown that using the MLIRT model leads to a significant increase of the proportion of variance explained.

---

Irina Grabovsky and David Swanson  
National Board of Medical Examiners

## An Application of Integer Programming to Optimal Test Assembly with Psychometric and Scheduling Constraints

Integer programming is one of several theoretical approaches that has been used to construct test forms from item banks according to test specifications. Typically, the test specifications concern the statistical characteristics and content coverage desired for test forms.

This study addresses issues of test assembly in the setting of a performance-based test of clinical skills in which lay persons (standardized patients – SPs) portray patient roles in a series of simulated clinical encounters. Examinees (physicians-in-training) interact with SPs as if they were real patients, and SPs rate examinee performance at the conclusion of each simulated encounter. The test is administered throughout the year at multiple regional sites, and test form construction must take into account the number of examinees to be tested, the work schedules of SPs, and the patient roles available SPs have been trained to portray, as well as the content and statistical specifications for test forms. Thus, the new issue in form construction is to assemble collections of test forms that simultaneously meet content, statistical, and scheduling requirements.

We provide an example of such a scheduling problem solved by the linear programming package CPLEX for one set of sessions for each day of one week. We explore possible addition of controlling case and SP exposure while simultaneously ensuring scheduled SP work hours fall between reasonable minimum and maximum values.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Kevin J. Grimm, John J. McArdle, and Fumiaki Hamagami  
University of Virginia

## Growth Mixture Modeling of Cognitive Abilities in the Berkeley Studies

Research on the development of intelligence has focused on the differential development of multiple cognitive abilities, but has not fully considered models with multiple groups of subjects with different patterns of development (McArdle and Nesselrode, 2003). Most multiple group models are of known groupings (i.e. gender, education) but there may be clusters of individuals with distinct growth patterns that are not well defined in terms of demographics. Averaging across these latent classes could lead to false conclusions about developmental patterns and theory. Recent advances in structural equation modeling, including growth mixture modeling (Nagin, 1999; Muthén & Muthén, 2000), have allowed further investigation of the clustering of individuals based on developmental trends. The clustering of participants in the development of vocabulary and memory ability is investigated in the Berkeley Growth (*BGS*) and Guidance (*GS*) studies using growth mixture models. A three class growth mixture model best characterizes the trends of the participants' development in vocabulary and memory. The vocabulary classes were related to the sampling of participants (*BGS* vs. *GS*), educational attainment of participants and parents, and early childhood IQ. The memory classes were also associated with the sampling of the participants and childhood IQ.

---

Adam R. Hafdahl  
University of Missouri-Columbia

## Refinements for Random-Effects Meta-Analysis of Correlation Matrices

Consider the meta-analytic problem of combining sample correlation matrices across  $k$  independent studies. Available methods accomplish this under fixed- and random-effects models. The conventional fixed-effects approach, which employs generalized least squares, has been improved appreciably by two refinements: using initial estimates of the common correlations ( $\boldsymbol{\rho}$ ) instead of observed correlations ( $\mathbf{r}_i$ ) in the conditional covariance matrices that serve as (inverse) weight matrices, and analyzing Fisher  $Z$ -transformed correlations instead of Pearson correlations. The present paper extends these refinements to the random-effects case where both the between-studies mean ( $\boldsymbol{\mu}_\rho$ ) and covariance-component matrix ( $\mathbf{T}$ ) of study correlations ( $\boldsymbol{\rho}_i$ ) are estimated. One choice concerns using estimates of  $\boldsymbol{\mu}_\rho$  versus  $\boldsymbol{\rho}_i$  in the conditional covariance matrices. These options, along with the Fisher- $Z$  modification, were evaluated in a Monte Carlo study. Other factors included  $k$  and mean within-study sample size. Random-effects data generated from a six-element mean correlation matrix were analyzed by conventional and refined methods, each of which yields estimates of  $\boldsymbol{\mu}_\rho$  and  $\mathbf{T}$ . In terms of element-wise estimation and inference, all five refined methods outperformed the conventional approach, especially with many small studies. The superior method varied, however, according to the focal parameter (i.e.,  $\boldsymbol{\mu}_\rho$  vs.  $\mathbf{T}$ ). Practical recommendations are discussed.

---

Ellen Hamaker and Peter Molenaar  
University of Amsterdam

## Towards an Integration of Intraindividual and Interindividual Techniques

It has been acknowledged that measurements contain both stable, trait-like properties and temporal, state-like properties. Although there are several methods and models to disentangle these properties, all of these approaches are based on two rather restrictive assumptions: (1) the state structure is identical across individuals, implying that the way individuals differ from themselves is universal; and (2) the state structure and the trait structure are identical, implying that the way individuals differ from themselves is the same as the way individuals differ from each other.

A new model is presented called the integrated trait state model, which consists of a combination of intraindividual and interindividual techniques. In this model individuals may differ from each other with respect to their state structure, and as a logical consequence the trait structure and the state structure may differ from each other.

It is shown what the consequences are when data are generated by this model, and analyzed with standard factor analysis.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Jeffery Harring  
University of Minnesota

## Statistical Methods for the Analysis of Repeated Measurements Data: A Model Comparison Approach

In particular research domains, mixed-effects models are employed exclusively for analyzing repeated measures data in which the within-subject profile is decidedly nonlinear. The goal is to summarize the average change process in the population, while concurrently describing individual patterns of development. In other research settings, latent curve models are used to characterize the nonlinear relationship between individual responses and occasion of measurement. Unlike their mixed-model counterpart, the focus of these analyses is typically on the individual. The average response necessarily follows a specific nonlinear function, however individual profiles need not take the same functional form. This approach to latent curve analysis is close in spirit to a mixed-effects model analysis and suggests that they would yield comparable results when used on the same repeated measures data. Two competing methods, suggested by Sheiner and Beal (1980) and Lindstrom and Bates (1990), for estimating the nonlinear mixed-effects model will be reviewed and compared with a structured latent curve approach proposed by Browne and Du Toit (1991) and Browne (1993). Similarities of these three models will be assessed by comparing individuals' fitted curves as well as the fit of the mean vector.

---

Jennifer P. Hatfield  
University of Minnesota

## A Comparison of Dichotomous and Nominal Polytomous Item Response Theory Models as Applied to Multiple Choice Test Items

Relatively little research is available regarding the application of nominal response IRT models to educational data and comparisons are scarce among the several nominal models now available. This study compared item, test, and test score characteristics resulting from the application of dichotomous and nominal IRT models to data from a large-scale multiple-choice K-12 achievement test. Two dichotomous models—the two- and three-parameter logistic (2PL and 3PL respectively)—and three nominal models—the nominal models of Bock (BNR), Samejima (SNR), and Thissen (TNR)—were used to calibrate items and score examinees.

All models showed fairly strong agreement in terms of rank ordering of correct-option parameter estimates. There was fairly strong agreement among the 3PL, SNR, and TNR in terms of rank ordering of parameters associated with guessing. Per degree of freedom, nominal models had smaller item chi-square fit statistics than dichotomous models. The more complicated nominal models (SNR and TNR) did not appear to result in a marked improvement in model-data fit. Nominal models had greater information across the range of ability. Finally, rank order of ability estimates was very similar across the dichotomous and nominal models. However, nominal scoring returned higher, and somewhat less variable, ability estimates than dichotomous scoring.

---

Todd C. Headrick  
Southern Illinois University-Carbondale

## Distribution Theory for the Power Method

The power method is a moment-matching procedure (Fleishman, 1978, four moments; Headrick, 2002, six moments) that generates a class of non-normal distributions often used in Monte Carlo studies involving achievement and psychometric measures. The primary advantage of the power method is that it provides computationally efficient algorithms for simulating multivariate data with arbitrary correlation matrices (Vale & Maurelli, 1983; Headrick, 2002; Headrick & Sawilowsky, 1999). A problem with the power method transformation is that its exact distribution is unknown. Specifically, the transformation's probability density function (*pdf*) and distribution function have not been defined. Thus, in general, it is difficult to determine the (a) modality (uni- or multi-modal), (b) location of the mode(s), (c) tail densities, (d) peakedness, and (e) quantiles for distributions generated by the power method. This presents a problem in terms of the power method's general ability to simulate the shapes of non-normal distributions.

Therefore, the purposes of this presentation are to derive the power method's *pdf* and inverse distribution function. Thus, and in the context of the class of distributions associated with the power method, this will enable the research methodologist to (a) determine whether or not a distribution is also a valid *pdf*, (b) calculate probabilities associated with *pdfs*, (c) classify a distribution in terms of its modality, (d) locate the mode(s) of a distribution, and (e) plot the graph of a distribution. Numerical examples will also be provided to demonstrate the methodology.

Keywords: Moments, Monte Carlo, Multimodal distributions, Power Method

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Willem J. Heiser  
Leiden University

## Geometric Representation of Association Between Categories

Categories can be counted, rated, or ranked, but they cannot be measured. Likewise, persons or individuals can be counted, rated, or ranked, but they cannot be measured either. Nevertheless, psychology has realized early on that it can take an indirect road to measurement: what can be measured is the strength of association between categories in samples or populations, and what can be quantitatively compared are counts, ratings or rankings made under different circumstances, or originating from different persons. The strong demand for quantitative analysis of categorical data has thus created a variety of statistical methods, with substantial contributions from psychometrics and sociometrics.

What is the common basis of these methods dealing with categories? The basic element they share is that the sample space has a special geometry, in which categories (or persons) are point masses forming a simplex, while distributions of counts or profiles of ratings are centers of gravity, which are also point masses. Rankings form a discrete subset in the interior of the simplex, known as the permutation polytope, and paired comparisons form another subset on the edges of the simplex. There are a few ways to define distances between point masses, and these form the basic tool of analysis.

The paper gives some history of major concepts, which naturally leads to a new concept: the shadow point. It is then shown how loglinear models, Luce and Rasch models, unfolding models, correspondence analysis and homogeneity analysis, forced classification and classification trees, as well as other models and methods fit into this particular geometrical framework.

---

Cheryl D. Hill  
University of North Carolina at Chapel Hill

## Precision of Parameter Estimates for the Graded Item Response Model

The graded response model is used in the context of item response theory (IRT) analyses of data for Likert-type questionnaires and educational assessments scored using rating scales. This research examined the recovery of item parameters for the graded model using simulated data for several short test lengths (3-15 items) and a range of sample sizes (50-1000). Simulated data sets also varied with respect to the number of response categories for each item (3-9) and the distributions from which the true parameters were drawn. Maximum marginal likelihood item parameter estimation was used. Bias and root mean squared error (RMSE) were calculated for each item parameter in each of 1000 replications for every combination of test length, sample size, number of categories, and parameter distribution. Results describe the precision of parameter estimation as a function of test length, sample size, and number of response categories. The results of this study suggest that there is no reason to combine responses in categories with low endorsement, as long as the number of responses in each category is greater than zero.

---

Andrew Ho  
Stanford University

Diego Zapata  
Educational Testing Service

Jon Templin  
University of Illinois

## Fast Classification and other operational issues for large scale testing

MCMC is in general too slow for operational scoring, especially for operational testing in which the number of test takers is high and score turnaround period is short. Several approaches were investigated for the Next Generation TOEFL. This paper describes the methods and findings.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Dennis Hocevar  
University of Southern California

Susan Page Hocevar  
Naval Postgraduate School

What is Good for Science is not Good Enough for Public Policy

“What is good for science is not good enough for public policy” is the theme of this paper. Few psychometricians would argue with the proven utility of the basic assumptions that underlie the use of scaled measures in behavioral/social science research:

*Ability, achievement, values, dispositions, performance etc. can be scored in terms of degree or amount on a numeric continuum because:*

1. *additive scale scores are equal interval.*
2. *the distribution of additive scale scores is continuous/normal.*
3. *the standard error of measurement for additive scale scores is randomly distributed.*
- 4.

In an era of accountability, additive scaled scores have taken on new meaning. Not surprisingly, the *same scaling assumptions* that we use in the conduct of science have been uncritically accepted by the public. Scaling is now considered as an objective, reliable, valid and fair way to assess organizational productivity in both the private and public sectors. To illustrate, many US States now scale school performance on a numerical continuum and use some type of change in these same scale scores to indicate “adequate yearly progress” (AYP). Using California’s Academic Performance Index (API) school data as an example, the pitfalls of using psychometric reasoning to make high stakes policy decisions will be illustrated.

---

Paul Holland  
Educational Testing Service

Linking Tests

In 1999 the National Research Council panel produced the report Uncommon Measures: Equivalence and Linkage Among Educational Tests. Here we review some of the rationale for the panel’s conclusions and then report on research that evolved out of his work on that NRC panel. This includes work with Neil Dorans on measuring the sensitivity of test linking to the target population, and research with Machteld Hoskens on direct and indirect true score prediction—a form of test linking weaker than equating. We will summarize what is known about how test differences affect linking results, and urge measurement statisticians who have access to real testing data to amplify and fill in the gaps in our knowledge of this important psychometric problem.

---

Chun-Wei (Kevin) Huang  
CTB/McGraw-Hill

Robert J. Mislevy  
University of Maryland-College Park

An Application of the Andersen/Rasch Multivariate Measurement Model within the Framework of Evidence-Centered Design to Explore Students’ Problem-Solving in Physics

Most of the analyses of physics assessment tests have been done within the framework of classical test theory in which only the number of correct answers is considered in the scoring. More sophisticated analyses have been developed recently by physics researchers to further study students’ conceptions/misconceptions in physics learning (in particular, the Newtonian mechanics) to improve physics instruction. It was suggested that in most cases a student is in a mixed model state, indicating that he/she is simultaneously occupied by a number of distinct model states and which state would be invoked depends on the feature of the question. However, these methods are not connected with the well-developed psychometric machinery.

The little-known Andersen/Rasch (AR) multivariate model that deals with mixtures of strategies within individuals is then introduced within the framework of the evidence-centered design (ECD). The ECD aims to coordinate psychology (of human behaviors), task design, and psychometric analysis when performing assessments. A small physics data set, analyzed under *BUGS* computer program that carries out the Markov chain Monte Carlo (MCMC) estimation procedure, is used to demonstrate how the ECD works as well as the usefulness of AR model in terms of students’ problem-solving in physics.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Philippe Huber, Elvezio Ronchetti, and Maria-Pia Victoria-Feser  
University of Geneva

## Estimation of Generalized Linear Latent Variable Models

Generalized Linear Latent Variable Models (GLLVM) enable modeling of relationships between manifest and latent variables when the manifest variables. They extend structural equation modeling techniques, which are powerful tools in the social sciences. However, because of the complexity of the log-likelihood function of a GLLVM, an approximation such as numerical integration must be used for inference. This can limit drastically the number of variables in the model and lead to biased estimators. In this paper, we propose a new estimator for the parameters of a GLLVM, based on a Laplace approximation to the likelihood function and which can be computed even for models with a large number of variables. The new estimator can be viewed as a M-estimator, leading to readily available asymptotic properties and correct inference. A simulation study shows its excellent finite sample properties, in particular when compared with a well established approach such as LISREL. A real data example on the measurement of wealth for the computation of multidimensional inequality is analysed to highlight the importance of the methodology.

---

Heungsun Hwang  
HEC Montreal

William R. Dillon  
Southern Methodist University

Yoshio Takane  
McGill University

## An Extension of Multiple Correspondence Analysis for Capturing Unobserved Respondent Heterogeneity

We propose an extension of multiple correspondence analysis that takes into account unobserved heterogeneity in respondents' preferences/choices. The method involves the combination of multiple correspondence analysis and k-means in a unified framework. The former is used for uncovering a low-dimensional space of multivariate categorical variables while the latter is used for identifying clusters of homogeneous respondents. The proposed method offers an integrated graphical display that provides insightful information on cluster-based segmentation structures inherent in multivariate categorical data as well as the interdependencies among the data. An illustrative example is given to demonstrate the empirical usefulness of the proposed method.

Keywords: Multiple correspondence analysis, k-means, unobserved respondent heterogeneity, alternating least squares.

---

Heungsun Hwang  
HEC Montreal

Yoshio Takane  
McGill University

## Generalized Structured Component Analysis

We propose an alternative method to partial least squares for path analysis with linear components, called generalized structured component analysis. The proposed method replaces factors by exact linear combinations of observed variables. It employs a well-defined least squares criterion to estimate model parameters. As a result, the proposed method avoids the principal limitation of partial least squares (i.e., the lack of a global optimization procedure) while fully retaining all advantages of partial least squares (e.g., less restricted distributional assumptions and no improper solutions). The method is also versatile to capture complex relationships among variables, including higher-order components and multi-group comparisons. A straightforward estimation algorithm is developed to minimize the criterion.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Hidetoki Ishii  
National Center for University Entrance Examinations, Japan

Hiroshi Watanabe  
The University of Tokyo

**Bayesian Consideration of Test-Retest Reliability Coefficient**

The test reliability should be confirmed when a psychological test is developed. Test-retest method is one way to estimate the reliability coefficient of the test. Because the population correlation coefficient between two strong parallel tests is equal to the reliability coefficient of the tests, sample correlation coefficient between test and retest is regarded as an estimate of the reliability coefficient of the test. However, almost always, a sample scores do not satisfy the assumptions of strong parallel tests. Therefore, sample correlation coefficient could be suspected as an estimate of the reliability coefficient. On this talk, Bayesian estimation of test-retest reliability coefficient is considered. It is presented that 1) a sample correlation coefficient should over estimate the test-retest reliability whenever sample means are not the same or sample variances are not homogeneous, on the other hand, 2) Bayesian estimate can adjust the differences among sample statistics and will be more appropriate estimate.

---

Jennifer L. Ivie  
University of Kansas

**Using Structural Equation Modeling to Assess the Role of Rules Problem Solving as in the Raven's Matrices Test**

The Raven's Matrices Test (RMT) has been shown to measure general intelligence in young children. Carpenter, Just and Shell (1997) demonstrated that solutions for these problems can be found using particular rules defining patterns across rows and down columns in matrices of these types. The purpose of this study was to use structural equation modeling path analyses as outlined by Dimtrov and Reykov (2003) to assess the relationships between these rules in solving items of this type. A sample of 1,364 6th graders from a New England school district, having taken 3 speeded 12-question RMTs, was used for this study. The data supported the hypothesis that four of these rules used in this test did have increasing difficulty levels that were related to increasing difficulties of the items requiring the use of each of the rules. The data also supported the hypothesis that knowing an easier rule increased a child's probability of successfully completing a matrix requiring the use of a harder rule.

---

Margo G.H. Jansen  
University of Groningen

**A Comparison of Latent Trait Models for Speed Tests with Different Distributional Assumptions**

Response times are considered relevant in a wide variety of psychological and educational measurement situations. In particular the use of computers for test administration has opened the possibility to study observations that are not available in paper and pencil tests, such as the response times on individual items. Response times can be seen as achievement measures in their own right but also as explanatory variables for other types of test behavior. We consider models for item- or test response latencies. Both the lognormal and the gamma distribution, can be viewed as suitable distributions for modelling continuous positive random variates. The extended Rasch model assuming gamma distributed response times, is compared with a mixed regression model assuming lognormal distributed response variables. For fitting the lognormal mixed regression model standard multi-level modelling software can be used. A number of empirical examples are given.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Robert I. Jennrich  
University of California, Los Angeles

## Rotation Algorithms: From beginning to end (?)

The development of rotation algorithms from graphical to current analytic methods will be reviewed. At present there are simple very general and reliable algorithms for both orthogonal and oblique rotation.

More or less in historical order, beginning with quartimax and varimax, orthogonal pairwise methods for the orthomax family were developed. These were generalized to include orthogonal pairwise methods for arbitrary quartic criteria. Adding a line search later generalized these to arbitrary criteria.

Indirect oblique methods are based on reference structures rather than loading matrices. Beginning with the quartimin criterion, methods that modified one column of the reference structure at a time were given for the oblimin family.

Direct oblique methods are based on loading matrices. A one parameter representation for the result of rotating one factor in the plane of two was introduced. Starting with the quartimin criterion this was used to produce pairwise algorithms for the oblimin family. Adding a line search later gave a direct oblique algorithm for arbitrary criteria. An alternative approach used a gradient projection method and numerical gradients to give another direct oblique algorithm for arbitrary criteria.

The general orthogonal method and these last two methods come close to providing a complete solution to the orthogonal and direct oblique rotation problems. Computer code for these methods for use in a variety of computing environments may be downloaded and is free.

---

Xiaoying Jiang  
Harcourt Assessment, Inc.

Grace E. Kissling  
University of North Carolina, Greensboro

## Using Quadratic Discriminant Model to Predict Baccalaureate Nursing Students' Passing Rate on NCLEX-RN

This study focuses on the exploration of the useful variables for predicting NCLEX-NR passing rate on the first attempt. There are altogether 690 observations with no missing data on NCLEX-RN. Forty-seven variables were considered, including forty-four continuous variables (Nur360, MOSBY Assessment Score, etc) and three categorical variables (age, gender, and type of students). A quadratic discriminant model was developed based on the characteristics of the data. It was discovered that the best predictive variables are overall GPA for all nursing courses and MOSBY Assessment Test Scores, which jointly predicted correctly 78.26 percent of the students who failed, 82.93 percent of the students who passed NCLEX-RN on the first time. This finding is similar to the previous study done by Beeson and Kissling (2001) except that age was excluded in the current model. Implications of the study and further studies were recommended.

Key words: MOSBY Assessment Test, GPA, NCLEX-RN, Discriminant Analysis, and Predictor Variables

---

Timothy R. Johnson, Ph. D.  
University of Idaho

## Generalized Linear Models with Ordinally-Observed Covariates

An ordinally-observed variable is a continuous variable that is only partially observed through an ordinal proxy variable. Statistical models for ordinally-observed response variables are well known but relatively little attention has been given to the problem of ordinally-observed explanatory variables. Point estimates and standard errors for fully-observed explanatory variables are not necessarily consistent and may be significantly biased when ordinally-observed covariates are measured using ordinal proxy variables. Thus the ordinal nature of ordinally-observed covariates must be taken into account to obtain correct inferences. In this talk I will outline the problem of ordinally-observed covariates, propose a general modeling framework for generalized linear models (and extensions thereof) with ordinally-observed covariates, and discuss issues of model specification, identification, and estimation within this framework.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Yutaka Kano and Shohei Shimizu  
Osaka University

## Between ICA and SEM

SEM has used higher-order moments to create robust estimators and/or robust asymptotic covariance matrix of estimators in nonnormal populations. The usage of higher-order moments does not solve any inherent problems of SEM such as equivalent models and inability of model assessment of saturate models. Recent research by Kano, Shimizu and our collaborators has shown that more ACTIVE use of higher-order moments as in independent component analysis (ICA) can solve many of these problems. In this talk, we first present basic ideas of such use of nonnormality and show how SEM people can apply the fundamental theorem by Comon (1994) to structural equation modeling including statistical causal analysis. Some comprehensive examples are provided. Some coincidence between ICA and SEM about independence assumption is pointed out. More technical results and applications will be given in later talks of this session.

---

Tzur Karelitz  
University of Illinois, Urbana

## Ordered Category Attribute Coding Framework for Cognitive Assessment

Cognitively Diagnostic Assessment models define skills as binary. Examinees are described as either 'skill masters' or 'non-masters' and items as either requiring the skill or not. In an attempt to enhance the diagnostic information provided by such models, I propose an Ordered Category Attribute Coding (OCAC) framework. This approach defines skill  $k$  by the  $M_k$  steps taken to master it. Consequently, the entries of the categorical  $Q$  matrix represent mastery level skills required by test items. Examinees knowledge patterns represent their location on the learning path of each skill. The flexibility of the OCAC framework allows for a more informative, parsimonious and efficient representation of task requirements and examinee knowledge. In addition, levels of required skills can be estimated simultaneously with the examinees knowledge states as well as noise parameters, with high recovery rate of simulated and real data.

---

Henk Kelderman  
Vrije Universiteit

## Measurement Models Based on Item Exchangeability

To date, it is concluded that item responses measure the same attribute if they fit a 1-latent-variable model. However, because the latent variable is unobserved, the precise nature of its relations with the item responses and other variables cannot be determined on empirical grounds. Several psychometricians, such as Bartholomew, Ramsay, and Goldstein have noted this weakness.

Thus, one needs a theoretical justification. However, unlike Physics, where theory is almost always cast in mathematical form, the Behavioral and Social Sciences rarely have such theories. Early psychometricians such as Gulliksen and Lazarsfeld tried to formulate criteria for different test items to measure the same attribute. Gulliksen required that measurements of the same attribute should be interchangeable in the sense that 'it is indifferent as to which of the measures is used'. Lazarsfeld proposed a similar criterion.

In this paper we justify latent trait models on the more formal criterion of item exchangeability. The concept of item exchangeability has appeared earlier in the work of TenHave, Lauritzen, Huynh and myself and is more in accordance with original ideas. It is shown that latent trait models are observationally equivalent with exchangeability models. Exchangeability models have the charming property that, apart from some type of item exchangeability, only some basic assumptions about scale type and distribution of manifest item responses are needed.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Jee-Seon Kim  
University of Wisconsin-Madison

Jeroen K. Vermunt  
Tilburg University

## A Family of Longitudinal Association Models with Latent Variables

This talk presents a family of conditional RC(M) association models where the column corresponds to a latent class variable and the row corresponds to a vector of indicators. It is shown that characteristics of a large number of individuals and their responses can be effectively summarized using this method as dependency among the indicators are simplified by conditioning on the latent variable, and the row and column scale values define coordinate locations of latent classes and response categories in a low dimensional space. Moreover, the time-varying latent structure offers a flexible representation of change in longitudinal data. The dimensional reduction, classification, and modeling change can be performed even when only one indicator is available at each time point, although multiple indicators provide a richer framework such as allowing tests of measurement invariance over time. Both single and multiple indicator cases and their implications are discussed.

---

Julius M. M. Kitutu  
University of Pittsburgh School of Nursing

## Post High School Career Expectations: A comparative Study between Pittsburgh (USA) and Essen (Germany)

The study employed secondary data on school-to-work transitions of youth aged 18-24 years, from Pittsburgh (US) and Essen (Germany). Separate theoretical models were fitted to each sample based on the youth's educational level, age, and marital status, parental educational attainment, and current transitional status. Future career expectation was measured by career assurance; parents' expectations for their youth achieving career goals, and youths' perceived career goals. Parental educational level in the Essen sample related significantly with both high school and choice of post-secondary program. Job-search patterns were different for both samples. A majority (90%) of the Essen youth used their teachers, whereas for Pittsburgh sample, 82% directly submitted written applications to potential employers. The hypothesized model was supported in both samples: Pittsburgh ( $\chi^2_{(39)}=52.469$ ,  $p=.073$ , NFI=.993; CFI=.998; RMSEA=.034) and Essen ( $\chi^2_{(10)}=15.657$ ,  $p=.110$ ; GFI=.995; NFI=.994; CFI=.997; RMSEA=.027). For the Essen model, 37% of variance associated with current transitional status was accounted by its predictors, whereas for the Pittsburgh model, its predictors explained 27% of variance associated with future career expectation. The model shows that youth in Pittsburgh experience delayed career outcomes.

---

Andreas G. Klein  
University of Illinois at Urbana-Champaign

## Efficient Estimation of Nonlinear Effects in Both Cross-Sectional and Longitudinal SEM

This paper deals with an efficient methodology for the estimation of nonlinear SEM, which provides the analysis of nonlinear effects for a broad range of cross-sectional and longitudinal SEM. The proposed method provides an approximate ML estimator and outperforms the currently available methodology (covariance structure analysis, two-stage least squares) with respect to statistical power and flexibility. Parallel to the estimation method, a new general latent variable modeling framework is presented which incorporates the modeling of multiple nonlinear effects in both cross-sectional and longitudinal models. For the case of cross-sectional models, this framework includes models with multiple interaction effects among latent exogenous variables. For the case of longitudinal models, the applicability of the framework is illustrated by a heterogeneous growth curve model. Formally, heterogeneity of growth refers to the fact that some subgroups of individuals grow more consistently than others, and to model this heterogeneity statistically correct is essential for optimal prediction of individuals' development. The new methodology enables the researcher to analyze more efficiently what heterogeneity of growth could depend on and provides a tool for an improved prediction of individual growth, based on initial status or other covariate information.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Iosif Krass and Alan Nicewander  
DMDC

The Effects of Item Position on the Calibration of Items Seeded into Adaptive Tests

When pre-testing potentially new items for CAT pools, it is important that examinees respond in the same manner as if they were taking the item operationally; if so, then the item parameters will reflect the operational performance of the item. The use of try-out items seeded into adaptive tests is an efficient way for developing new items. It may be the case that the position in which an item is seeded will have an effect on the item parameter estimates due to context effects. This paper uses real data to look at the effects of seeded-item position on item parameter estimates. Try-out items were administered during operational CAT testing for ten fixed-length tests containing 10 or 15 items. For each examinee, a single try-out item was randomly administered in positions 1-11 for the 10-item tests and positions 1-16 for the 15-item tests. For each test, approximately 20 pretest items were administered in each possible position. Approximately 1200 responses were collected for each pretest item in each position. Item parameter estimates were compared across item positions, for each try-out item. Negligible effects of item-position on item parameter estimates were found.

---

Sik-Yum Lee and Xin-Yuan Song  
The Chinese University of Hong Kong

A Unified Maximum Likelihood Approach to Structural Equation Models with Missing Non-Standard Data

In this paper, we present a unified approach for maximum likelihood analysis of structural equation models that involve subtle model formulations and non-standard data structures. Based on the idea of data augmentation, we describe a generic Monte Carlo Expectation-Maximization algorithm for estimation. We propose path sampling for computing the observed data likelihood functions that usually involve complicated integrals, and show how to apply this method for computing the Bayesian information criterion (BIC) for model comparison. Some illustrations are given.

---

Jun Corser Li  
University of California, Berkeley

A Multilevel Covariance Structure Model for Causal Connection Research of Group Effectiveness

In group effectiveness research, and particularly in school effectiveness multilevel analysis, hierarchical linear models (Raudenbush, 1988; Goldstein, 1997) and multilevel covariance structure models (Longford and Muthén, 1992; Muthén, 1994) have been widely applied. This paper first discusses the problems of interpreting school effectiveness, the choice of performance indicators, and the distinction between effectiveness and residuals. It then proposes an improved multilevel covariance structure model (Li, 2002) that extends and unifies these approaches and explores a dialectic relation (a mutually exclusive causal connection) between group level performance indicators and their effectiveness. Bayesian inference for this model is implemented via a Markov Chain Monte Carlo procedure with the Gibbs sampling algorithm. The paper concludes with two simulation studies for causal connection.

Keywords: multilevel covariance structure model, group effectiveness, causal connection, Gibbs sampling.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Mei Liu  
Law School Admission Council

Paul Holland  
Educational Testing Service

Population Invariance of Non-Linear Equating Using LSAT Test Data

A fundamental requirement of an adequate equating is that equating functions should be population invariant. Since population invariance can never be satisfied completely, practitioners should routinely assess whether population invariance holds sufficiently for equating. Research on the Dorans and Holland measures of population invariance has shown that such indices are sensitive to the linking of tests that measure different constructs or differ in reliability. Results from these studies also call for more research examining invariance for other sub-populations, with other exams and using other equating/linking methods (Von Davier, Holland & Thayer, 2002, Dorans et al, 2002, Tateneni & Dorans, 2002, Yang et al, 2002).

Using Law School Admission Test (LSAT) data, Liu and Holland (2004) examined population invariance of linear equating functions for male and female examinees and examinees with different ethnic backgrounds, as well as across sub-populations defined by examinees' geographic regions, whether they applied to law school(s) and their law school admission status.

The current research expands the Liu and Holland study by first replicating it on data from different LSAT administrations and then extending it to explore the sensitivity of non-linear equating and linking functions across the aforementioned sub-populations.

Keywords: population invariance; equating; linking

---

R. Duncan Luce  
University of California, Irvine

Measurement Analogies: Comparisons of Behavioral and Physical Measure

Physical measurement involves attributes, variables, constants, tradeoffs of factors, interlocking structures and distribution laws, and numerical measures of varying types: ordinal, interval, and ratio. Consider the following questions. Are behavioral measures in some way similar to physical measures? Are, for example, measures of intelligence more like measures of mass or more like hardness, the former forming a ratio scale and the latter an ordinal one? What does it take to answer that question? Can a person be viewed as, in some sense, a medium in which concepts such as loudness and brightness act like attributes that are affected by factors that can be manipulated? If so, what measurement representations result? How does hunger differ from loudness and do we know how to measure it in a principled way? In what sense is an ordinal scale weaker than a ratio scale or, could it be, exactly the opposite? The talk addresses such issues.

---

Robert C. MacCallum  
University of North Carolina

Michael W. Browne  
The Ohio State University

Li Cai  
University of North Carolina

Testing Differences between Models: Power Analyses and Null Hypotheses

In applications of structural equation modeling and related methods it is common for researchers to seek to evaluate competing alternative models and to identify which among two or more models is optimal in some sense. For comparing nested models, the standard procedure is to employ the likelihood ratio test of the difference in model fit, where the null hypothesis is that the two models fit equally well in the population. A procedure for determining the statistical power of this test will be presented, and factors that affect statistical power will be delineated. In addition, a modification of the standard null hypothesis of zero difference in fit will be proposed. It can be argued that in practice this null hypothesis is virtually never true and is empirically uninteresting. A modified testing procedure will be presented that allows for a null hypothesis of a specified small difference in fit, versus an alternative hypothesis of a larger difference, along with corresponding power analysis procedures.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Randall MacIntosh  
California State University, Sacramento

Elizabeth Cauffman  
University of Pittsburgh School of Medicine

An Assessment of DIF on the Basis of Race/Ethnicity in The Massachusetts Youth Screening Instrument (Maysi~2) Among a Sample of Incarcerated Adolescent Offenders

The juvenile justice system needs a tool that can identify and assess mental health problems among youths quickly, with reliability and validity. The goal of this paper was to evaluate the racial/ethnic item bias of the Massachusetts Youth Screening Instrument-Second Version (MAYSI~2) using the Rasch Model. Data are presented from 3,906 assessments of male and female juvenile offenders between 13 and 17 years of age who are incarcerated in the California Youth Authority (CYA). DIF is identified in some scales, raising concerns about the suicide ideations scale as well as the traumatic experiences scale that may require some further examination and revision.

---

Louis T. Mariano and Maria Orlando  
Rand Corporation

A Bayesian IRT Model for Comparative Item Performance under Dual Administration Modes

We investigate the effect of survey interview mode, phone versus a self-administered questionnaire, on the characteristics of survey items. Responses to the Center for Epidemiological Studies Depression scale were collected from each of 246 subjects using a random mode choice. The scale was then re-administered roughly 30 days later in the alternative mode. A traditional Differential Item Functioning (DIF) analysis would allow for a comparison of the item characteristics in each mode. However, the DIF modeling assumptions do not allow for the use of repeated measures, and the parameters of the traditional procedure do not facilitate interpretation of patterns of differences due to group membership. As an alternative, we develop a Bayesian Item Response Theory model that explicitly quantifies the survey mode effect at the item and category levels and utilizes the full information contained in repeated measures. When no mode effects are present, this model reduces to a Bayesian Graded Response Model. We demonstrate mode effects within items, quantifying the effect (i.e. DIF) present in each item, and also mode effects across items, indicating a fundamental difference in the performance of the survey scale across the two modes. Potential broader applications to problems without repeated measures are discussed.

Keywords: Item Response Theory, Differential Item Functioning, Bayesian models, repeated measures.

---

Gunter Maris  
Cito, The Netherlands

The Power of Posterior Predictive Checks

Posterior predictive checks offer a powerful and flexible tool for the evaluation of model fit. However if we want to test a classical null hypothesis against a well defined alternative we need to consider the issue of statistical power: Can we correctly reject the null hypothesis? It turns out that posterior predictive checks frequently fail to do so. We offer an explanation and a solution for the lack of power of posterior predictive checks.

To illustrate our solution we consider the three parameter logistic model. In particular, we show that the parameters of the three parameter logistic model are in a nontrivial way not identifiable if the values of the discrimination parameters are all the same. For this reason we propose a test for the null hypothesis that the values of discrimination parameters are all the same against the alternative that they are not. That is, we test the hypothesis that the parameters of the three parameter logistic model are not identifiable. We show that a classical posterior predictive check has virtually no power whereas our alternative procedure does have power.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Catherine McClellan  
Educational Testing Service

## Basics of Large-Scale Educational Surveys

This presentation will provide some basic context setting on what large-scale educational surveys are, what the data structures are, and where the major challenges in analysis of such data lie.

---

Lori D. McLeod and Sheri E. Fehnel  
RTI Health Solutions

## Defining Minimal Clinically Important Differences: Does Item Response Theory Have the Answer?

Historically, comparisons of medical treatments have focused on clinical measures such as blood pressure or tumor size. However, because treatments can have a negative impact on patients' health-related quality of life (e.g., toxicity associated with chemotherapy) clinicians are starting to also pay attention to patient-reported outcomes (PROs) as important factors when evaluating treatment options. Although most physicians see the value of including PRO results, they seek guidance for interpreting scores on these instruments so that they can better understand score significance and how to apply PRO findings to their practice. One attempt at providing this guidance is to define change scores in terms of a minimal clinically important difference (MCID). In general, a MCID is the smallest score difference that can be judged as worthwhile. There has been no consensus on how to define MCID, although various methods have been proposed. First, this study utilized item response theory (IRT) to define MCID on an acne-specific quality of life instrument. The IRT-approach was intuitively appealing because it offers a metric that can be used to compare scores across samples and items. Second, this study compared the results from IRT to those from three other methods.

---

Geoff McLachlan  
University of Queensland, Brisbane

## Mixture Model-Based Clustering of High-Dimensional Data

In the cluster analysis of high-dimensional data, k-means and hierarchical agglomerative methods are convenient off-the-shelf choices of the scientist. However, there are advantages to be had by adopting a normal mixture model-based approach to clustering. But given the high-dimensionality of the data, there is a need to first reduce the dimension of the feature space and/or to adopt parsimonious models for the component-covariance matrices. An obvious way to reduce the dimension of the feature space in an unsupervised context is to perform a principal component analysis. Unfortunately, there is a potential problem with the determination of an appropriate number of components useful for clustering. We describe an alternative approach that clusters the data on the basis of a mixture of factor models. There is the option of first reducing the number of variables by eliminating those considered not to have a useful individual role in clustering the data and then clustering the retained variables into groups. The cluster analysis of the data is then effected on the basis of representatives of the groups of variables.

---

Roger E. Millsap and Oi-Man Kwok  
Arizona State University

## Partial Invariance and Selection Accuracy in Two Populations

Studies of factorial invariance seek to determine whether a given common factor model holds across multiple populations with identical parameter values (e.g., factor pattern, latent intercepts, unique variances). Partial factorial invariance is said to exist when some, but not all, parameters have the same values across populations. The existing literature on factorial invariance is unclear about what action, if any, should be taken in response to findings of partial invariance. For example, should measures that lack invariance be dropped? Should such measures be retained if the variation across populations is "small?" Here we present one approach to answering these questions by evaluating the impact of partial invariance on accuracy of selection based on a composite of the measures whose factor structure is under study. We assume that a single-factor model holds in two populations for the measures under study. Accuracy of selection based on the composite of these measures can then be evaluated analytically under varying degrees of partial invariance and bivariate normality for the factor and composite scores. Some examples are presented, with discussion of extensions and limitations.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Masashi Miyamura and Yutaka Kano  
Osaka University

## Robustified Covariance Selection

Covariance selection is an exploratory procedure to find causal relations among many variables. It intends to assess a direct relation between two variables after eliminating any effects of all the other two variables in observational studies. It is based on normality assumption and uses a partial correlation.

Outlying observations are observations that come from a population other than a target population. If a data set to be analyzed contains outlying observations, any statistical analysis can cause serious biases and can be misleading. Observational studies will be difficult in avoiding outlying observations. It is known, however, that correlation coefficients and partial correlation coefficients are very sensitive to existence of outliers, and thus ordinal covariance selection is also very sensitive. We develop a new robustified maximum likelihood method for the covariance selection, where the likelihood function is weighted according to how the observation is deviated. Test statistics associated with the robustified estimators are developed, which include statistics for goodness-of-fit of a model. In addition, an outlying score, similar to the Mahalanobis distance but more robust, is proposed, which makes it easier to identify outlying observations.

Real data sets are analyzed to illustrate usefulness of our procedure. It is turned out that our procedure leads to more reasonable results than does the ordinal covariance selection.

---

Ab Mooijaart  
University of Leiden

## Model Selection in Latent Variable Models when Common Statistical Assumptions are Violated

Model selection is one of the most important issues in Structural Equation Models (SEM). This selection is mostly based on goodness-of-fit measures. Computer packages like, LISREL and EQS, print out a lot of these measures. A few of these measures do have a known statistical distribution if some statistical assumptions hold. The most common assumptions are: The variables are normally distributed and the sample size is large. However, in most practical situations one or both assumptions do not hold. In this paper it will be shown that by using some resampling technique, some variant of the bootstrap technique, may lead to a procedure by which it is possible to decide whether a model holds or not. The basic feature of this procedure is that data are resampled under the condition that the model holds. It will be shown by some examples that for some well-known SEM models with latent variables the procedure leads to a test with a proper type I error and, under which conditions, the test yields a good statistical power.

---

Hiroto Murohashi and Hideki Toyoda  
Waseda University

## Model Specification Searches Using Genetic Algorithms for Factor Analysis Model

Model specification searches in structural equation modeling is thought to be a task that is to find the most appropriate combination of free and constrained parameters from the aspect of fitting functions and capability of interpretation. So, we can treat model specification searches as one kind of combinational optimization problem. But it is known that finding an optimum solution for combinational optimization problem is very hard because of the explosive increase of combinations. To deal with this problem, many kinds of meta strategies are proposed. Meta strategy is a searching method that combines a number of approximate means to derive solutions that has enough accuracy for combinational optimization problem. In case of structural equation modeling, the number of estimable models rapidly increases with increase of the number of observed variables. So, using meta strategies for model specification searches in structural equation modeling is supposed to have a certain efficacy. But there are not so much researches to examine that. In this research, we applied genetic algorithms to the model specification searches in factor analysis models. Simulation study to validate accuracy and application to real data are shown.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Ratna Nandakumar and Lawrence Hotchkiss  
University of Delaware

## Diagnostic Tools for Modeling Attitudinal Data

Attitudinal data is typically different from achievement data. Most achievement test data is modeled using monotone models, where probability of correct response to item increases as the ability increases. However, attitudinal survey item data, can be modeled, depending upon the type of response elicited by the statement, either by monotone models or non-monotone, single-peaked models known as unfolding models. It is known that items resulting in single-peaked item response functions have certain characteristics. For example, a principal components analysis of such data results in 2 linear factors. The plot of factor pattern results in a simplex graph, and the factor scores have a quadratic relationship. It has also been shown that the distribution of item locations on the latent continuum can determine if data follows unfolding models. For example, if a survey instrument contains only items with strong positions on the latent continuum without any ambivalent statements, then monotone models can describe such statements. Using simulated and real data, this study develops diagnostic tools to determine when unfolding models fit data.

---

Alan Nicewander, Rebecca Hetter, Gary Thomasson, Iosif Krass, Mary Pommerich, Daniel Segall, and Kathleen Moreno  
DMDC

## A Simulation Study of Parametric and Non-Parametric Algorithms for Calibrating 3-PL and Non-3PL Items Seeded into Adaptive Tests

The maintenance of stable item parameters and a consistent measurement scale are fundamental requirements for an operational CAT program. When items are used operationally in a CAT setting, their parameters may drift over time, as examinee populations change and respond differently to the items. This paper presents the results of a high-fidelity simulation in which five rounds of new item calibration and pool development are simulated, for changing examinee populations. The try-out items simulated were both 3-PL and non-3PL--of the non-3PL items, many had non-monotonic IRCs. Non-3PL IRCs were generated using a cognitive model that simulated two, commonly-used strategies of item writers; namely, 1) Make the foils similar in difficulty; and 2) Use a foil that is attractive to persons with partial knowledge of the item content (a strategy that often produced non-monotonic IRCs). Within each round, try-out items were calibrated to be on the scale of the operational pools, and new pools are constructed. The new pools are then administered operationally in the next round, and new items are pre-tested. In the last round, the items that comprised the initial operational pools are pre-tested, calibrated, and scaled, enabling a complete evaluation of item parameter and scale drift over time. Drift is also assessed after each round. The performance of several calibration/scaling methods is evaluated by comparing true and estimated parameters. The calibration methods compared were: 1) BILOG-MG; 2) IFACT (Segall, 2002), a fully Bayesian procedure utilizing MCMC estimation; and 3) ForScore (Levine, 1998), a non-parametric procedure.

---

Shizuhiko Nishisato  
University of Toronto

## Dual Scaling Approach to Correlation between Categorical Variables

When a variable has more than two categories, it can typically be represented in multidimensional space. Then, correlation between categorical variables should therefore be defined in multidimensional space. Strangely enough, however, the general practice is to use correlation defined by a single dimension, for instance, see the Kendall-Stuart type canonical correlation as used in SEM and the use of the correlation matrix associated with the first dual scaling solution. The current study is based on the dual scaling framework, and the procedure starts with obtaining multidimensional solutions (components), and then defines a measure of correlation between categorical variables in terms of all components, thus using the entire information associated with the two variables. The proposed correlation, indicated by  $\nu$  (Greek letter 'nu'), becomes identical to Cramér's  $V$  in total space. The main advantage of  $\nu$  lies in the fact that it can be expressed as comprising associations over multiple dimensions, an aspect which is lacking in  $V$ .

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Haruhiko Ogasawara  
Otaru University of Commerce

Asymptotic Robustness of the Normal Theory Asymptotic Biases Under Nonnormality in Structural Equation Modeling

The asymptotic robustness of the normal theory asymptotic biases of the least squares estimators of the parameters in covariance structures against the violation of normality is shown, which is obtained under the conditions required for the asymptotic robustness for the normal theory standard errors and the usual chi-square statistic. The asymptotic robustness holds not only for the estimators of the parameters whose normal theory asymptotic standard errors are asymptotically robust, but also for the non-robust ones. The Wishart maximum likelihood estimators are also shown to have the asymptotic robustness. A numerical illustration for the factor analysis model shows that the empirical biases of robust estimators under nonnormality are close to their corresponding normal theory asymptotic biases.

---

Akinori Okada  
Rikkyo (St. Paul's) University

Tadashi Imaizumi  
Tama University

A Joint Space Model of Asymmetric Multidimensional Scaling

A joint space model and an associated algorithm for two-mode three-way asymmetric multidimensional scaling (object x object x source) were extended from Okada and Imaizumi (2003a, b). In the model, objects and sources are represented as points respectively in a same multidimensional space. The model consists of the common joint configuration showing the relationships among objects, among sources, and between objects and sources, and the asymmetry weight showing the salience of asymmetric relationships among objects for each source. In the configuration of objects for a source, each object is represented as a point and a circle (sphere, hypersphere). An object does not have its own radius, but the radius of the circle representing an object for a source, showing the asymmetry of the object for the source, is determined by the distance between two points representing the object and the source.

---

Joseph A. Olsen  
Brigham Young University

Estimating and Comparing Classical Test Theory Measurement Models

Four measurement models from classical test theory (congeneric, tau-equivalent, parallel, and strictly parallel) are commonly estimated and tested for goodness of fit, usually using confirmatory factor analysis. The latter three are typically defined by cumulatively imposing equality restrictions on the factor loadings, residual variances, and item intercepts of a basic congeneric structure. Other combinations of these restrictions give rise to four additional models which are seldom discussed or have yet to be formally addressed in the literature. This paper considers all eight of these models, outlining procedures for model estimation using both structural equation modeling and linear mixed models. The paper also examines the 19 nested and 9 non-nested comparisons among the models in terms of conditional and unconditional tests of equality restrictions on sets of model parameters, as well as general goodness-of-fit evaluation. The paper broadens the range of potential options for researchers seeking parsimonious and well-fitting measurement models, and encourages the treatment of measurement problems within the context of more general statistical modeling strategies.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Thomas R. O'Neill  
National Council of State Boards of Nursing (NCSBN)

## Using Paired Comparisons and a One-Faceted Rasch Model to Create the Semantic Construct of Frequency

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. 396 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. Furthermore, the methodology can easily be applied to other similar semantic continua. This study does have a small sample size and the pool of raters was not very diverse, therefore future studies address those issues.

---

Maria Orlando and Kitty S. Chan  
RAND Corporation

## Evaluating DIF in Psychological Scales: Is Statistical Significance Enough?

IRT methodology is increasingly employed for scale development and refinement outside the education field. However, these non-educational IRT applications often involve established scales composed of carefully selected items, raising unique, and as yet unresolved challenges. For example, in DIF applications involving scales of this type, it is usually not practical to delete or replace DIF items. Thus it is not always clear how to evaluate the importance of observed significant DIF, and what, if anything, to do about it. This talk will present methodological issues that emerged from an NIMH study of DIF in two adolescent depression measures. Using the likelihood ratio method of detection, several significant DIF items were identified according to gender, age, health status and substance abuse risk using data from two large public-use datasets. To evaluate the importance of the significant DIF, the size of the DIF, calculated as the average difference between reference and focal group location parameters, and the area under the curve between the reference and focal group category response functions were examined, in conjunction with ICCs and consideration of item content. Integrating information from these different perspectives may be useful to the decision-making process when significant DIF is encountered in non-educational applications.

---

Tatsuo Otsu  
National Center for University Entrance Examinations

## Linking Tests By Nonlinear Factor Analysis For Continuous And Binary Variables

Comparisons and linking of achievement tests often require analysis of nonlinear relationship between scores. Here, a nonlinear factor analysis model that uses spline transformation of latent variables is proposed. Continuous and binary variables are admitted by the model. For binary valued variables, logits of the binomial mean parameters are expressed as piecewise polynomials of the latent variables. Using weak prior distributions for unique variances, stable estimation of nonlinear transformation parameters was achieved. Discrete approximation of the latent variables with assumption of conditional independence enables easy adaptation for the missing values of a MAR (missing at random) condition.

Examples of artificial data and achievement test data with variables of uncommon subjects are shown.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Mark Otten  
UCLA

A Longitudinal Examination of the Dimensionality and Predictors of Sport Confidence

The purpose of the study was to test the dimensionality of the Sport Confidence Inventory and examine how sport confidence changes in relation to motivational climate over the course of a competitive volleyball season. 163 female volleyball players aged 11 to 18 years were administered a modified version of the Sport Confidence Inventory and the Perceived Motivational Climate in Sport Questionnaire at the beginning (T1), middle (T2), and end (T3) of the season. An exploratory factor analysis using minimum residual factor extraction and Tandem Criteria rotation revealed a general sport confidence factor, with each of the 22 items loading at greater than .40 at each time point. Change over time was tested with a latent growth model and indicated sport confidence did not significantly increase or decrease, while performance climate increased and mastery climate decreased across time. Regression analyses revealed that performance climate negatively predicted and mastery climate positively predicted sport confidence at each time point. Cross-lag regression analyses revealed that performance climate at T1 influenced sport confidence at T2, and T2 mastery climate predicted T3 sport confidence. Applied implications associated with the stability of sport confidence and the differential influence of motivational climate on sport confidence will be discussed.

---

Koken Ozaki and Hideki Toyoda  
Waseda University

Paired Comparison IRT model by 3-value judgment: estimation of item parameters prior to the administration of the test

Currently, test operation using IRT (Item Response Theory) requires test items to be undergo parameter estimation by the examinees. Furthermore, after equating, the items may be included in an item pool used for several tests. However, this test operation method contains the probability of item content leakage. Thus, estimating item parameters while keeping item contents secret will become useful. In this study, to make such a situation possible, a model in which item parameters were estimated using paired comparison from the perspective of the difficulty of items by a rater familiar with the field is proposed. And, item parameter estimation using information function is also proposed, to reduce the number of ratings while keeping sufficient estimation accuracy. The estimation accuracy of this model was confirmed in a simulation study, and the feasibility of its use in practical settings is demonstrated using actual data.

---

Insu Paek  
Harcourt Assessment, USA

Mark Wilson,  
University of California, Berkeley, USA

Type I error and Power of Multidimensional and Unidimensional DIF Methods in a Multidimensional Test: MRCML DIF model, RCML DIF model, and SIBTEST

Based on the multidimensional random coefficient multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997), a multidimensional DIF model (MRCML DIF model) is formulated and its performance is investigated using simulated dichotomous response data. The foci of the investigation were the recovery of the DIF parameter in the MRCML DIF model, its Type I error rates, and its statistical power.

The unidimensional RCML DIF model (Paek & Wilson in preparation) and SIBTEST (Shealy & Stout 1993a) were also applied to see the effect of using a unidimensional DIF investigation for a multidimensional test. The MRCML DIF model performed very well in terms of DIF parameter recovery, Type I error rates and power. Interestingly, the applications of the unidimensional DIF models to these between-item multidimensional data also resulted in a very good performance in their DIF parameter recoveries, Type I error rates and power.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Richard J. Patz  
Stanford University

Lihua Yao  
CTB/McGraw-Hill

## Hierarchical and Multidimensional Models for Measuring Developmental Growth in Educational Achievement

This paper examines psychometric modeling approaches for measuring growth in educational achievement across years of schooling. Focusing on item response theory (IRT) scaling procedures, we examine model assumptions used in the "vertical scaling" of test forms for different grade levels, and we compare "divide and conquer" and unified approaches to models and data analyses in this context. We introduce and explore two unified approaches: 1) a hierarchical multi-group IRT model that allows explicit estimation of the functional form of the grade-to-grade growth patterns, and 2) a multidimensional, multi-group IRT model that captures differences in dimensionality and scale definition across grade levels. We explore properties of these models using simulated data and data from a writing assessment that spans grades 3-7.

---

Ling Peng and Adam Finn  
University of Alberta

## What Exactly Do Consumers Respond to in a Concept Test?

Concept testing plays a pivotal role in resource allocation decisions in the new product development process. This study uses generalizability theory as a conceptual framework for identifying what exactly consumers respond to in a concept test and quantifying the factors that contribute to its outcome. The observed measurement outcomes are determined not only by concept factors, but also by response task factors, respondent characteristics and situational factors and their interactions, reflecting response to both the concept and the testing context. An illustrative G-study varies the structure of a concept test by randomly sampling some factors (e.g., concepts, scales and subjects), manipulating the fixed levels of other factors (e.g., occasions) while keeping other factors constant. Estimates of the variance components from different sources of variation in the test are used to show managers how to design cost- and time- efficient concept tests for different decision making purposes. This study provides a better understanding of what exactly consumers respond to in a concept test, and as a result can suggest more appropriate concept testing practices.

---

Marika Polak, Willem J. Heiser, and Mark de Rooij  
Leiden University

## Correspondence Analysis as an alternative to Principal Component Analysis for single-peaked data

Single-peaked (unimodal) data naturally arise in a variety of research settings, e.g. marketing research, ecological research and in archeology. In psychology single-peaked response curves can be found, for instance, in attitude measurement: people with moderate tolerance towards euthanasia will less likely agree with statements that are either very much in favor of euthanasia or very much against euthanasia. As a consequence, only agreement and not disagreement, is an indicator of similarity among respondents in terms of the underlying trait being measured.

In many cases principal component analysis (PCA) is used to evaluate the psychometric properties of a measurement instrument. However, in the case of single-peaked data this technique is not optimally suited because it is based on a linear or monotonic model.

In this paper we will explore the surplus value of correspondence analysis (CA) for evaluating a measurement instrument for personality development, which is expected to produce single-peaked data. CA is known to correctly represent perfect single-peaked data (e.g. Heiser, 1981), however the quality of representation is unknown in the case of error or approximate single-peakedness.

Keywords: unipolar data, ANACOR, validity research, developmental psychology

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Mary Pommerich, Daniel O. Segall and Iosif A. Krass  
DMDC

## Evaluating the Scale of Items within CAT Pools Using Adaptive Data

Items within a CAT pool may often come from different pretest administrations. Item parameters from different pretest administrations may be calibrated separately and placed onto the same scale, typically using a common item linkage. The scaling of a CAT pool (i.e., whether items from separate sources are truly on the same scale) is only as good as the quality of the linkage conducted to place the parameters on the same scale. After pools have been used operationally, it may be useful to evaluate whether items from different calibration samples are indeed on the same scale. This paper compares two procedures for computing transformation constants that place within-pool item parameters from one source onto the scale of item parameters from another source. The methods are applied to a single group design, where some common items are shared across examinees. The single-group design is inherent in the administration of a CAT pool, and is in contrast with the design used with traditional procedures for computing scale transformation constants (i.e., examinees across two groups take different sets of items that each contain a fixed set of common items). Hence, the traditional procedures cannot be used in the case of CAT data with no fixed set of common items. The first procedure that is studied computes marginal maximum likelihood estimates of the transformation constants, where the maximization is conducted with respect to the transformation constants and the population mean and variance. The second procedure finds the transformation constants that minimize a quadratic loss function. The accuracy of the procedures in recovering the true transformation constants is assessed using a realistic CAT simulation.

---

Wai-Yin Poon and Sik-Yum Lee  
The Chinese University of Hong Kong

## Structural Equation Model Analysis of Missing and Heterogeneous Data

Structural equation models are widely used to model relationships among latent unobservable constructs and observable variables. Two problems in practical applications are the presence of missing observations and the heterogeneity of the data set. We develop a two-level model to account for heterogeneity, and use an EM type algorithm to handle missing observations. The resultant procedure can be applied to analyze a wide range of models, can handle data sets with unbalanced design and small level-one units, and allows constraints to be imposed on parameters that are associated with different levels. The work in this presentation was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (RGC Ref No. CUHK4347/01H).

---

Gilles Raïche  
Université du Québec à Montréal

Jean-Guy Blais  
Université de Montréal

## Comparison of the marginal bias and standard error of the proficiency level according to the true and the estimated proficiency levels

According to the usual practice, when the bias and the standard error of the proficiency level are computed in simulated IRT studies, it is in reference to selected marginal generating true proficiency levels. But, considering the fact that future practitioners can't have access to these selected generating proficiency levels, the need would be to obtain the bias and the standard error according to marginal selected estimated proficiency levels, the only values of interest to him. In this study, the bias and the standard error of the proficiency level were compared between two marginalization strategies, estimated and true proficiency levels, and between five different estimation methods, MLE, BME, WLE, EAP and AEAP based on the 3PL model. Results of 2000 simulations of 85 fixed items tests at each of 15 simulated proficiency levels showed the advantage to marginalize according to the estimated proficiency level. Results also indicated that when marginalization is applied with the estimated proficiency level in place of the true one, different conclusions about the superiority of the different estimation methods according to diminishing bias and standard error are drawn. Even the WLE estimation method lost his superiority, being supplanted by the AEAP estimation method.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Nilam Ram, Sy-Miin Chow, Kevin J. Grimm  
University of Virginia

Frank Fujita  
Indiana University South Bend

John R. Nesselroede  
University of Virginia

## Examining the Dynamics of Pleasant and Unpleasant Emotions Using Spectral Analysis and the Rating Scale Model

Capitalizing on trait (or ability) estimates obtained from the Rating Scale Model (Andrich, 1978), a Rasch-family model for polytomous items, we use these estimates to run a series of spectral and cross-spectral analyses. Starting with a complete data set in which daily self-reports of love and anger from 178 college students were collected over 52 days, selected items were dropped from the data set such that for each of the pre-defined measurement periods, only data from a subset of items were retained. The Rating Scale model was used to provide linkages among these items to yield daily estimates of each individual's latent levels of love and anger. Comparing spectral analysis results based on the complete and incomplete data revealed that the Rating Scale model provides an effective and accurate way of linking polytomous items that measure the same latent trait and can be of great utilities to researchers interested in implementing longitudinal studies with planned incomplete design.

---

Tenko Raykov  
Fordham University

John Tisak  
Bowling Green State University

## Studying Stability of Reliability in Repeated Measure Models Eliminating Variable Specificity

A covariance structure analysis method for testing stability of reliability in multiple wave, multiple indicator models is outlined. The approach accounts for observed variable specificity and permits in addition point and interval estimation of reliability in terms of "pure" measurement error variance. The proposed procedure is developed within a confirmatory factor analysis framework and is illustrated with data from a cognitive intervention study.

---

Mark D. Reckase  
Michigan State University

## Using IRT to Design a Fixed Length Test

Test design has been considered more of an artistic endeavor than a scientific one. This paper will show how to design the statistical characteristics of a fixed length test within an item response theory framework. Two different form construction philosophies are described and used to create statistical test specifications: testing examinees randomly selected from a specified population to a desired level of precision; and testing over a range of the trait scale to a desired level of precision. In both cases, the ideal difficulty distribution is broader than is typically seen in professionally developed tests. For the test to measure over a range of the trait scale, the ideal test has counter intuitive characteristics that have very interesting properties. Several examples of ideal test specifications will be provided.

---

Frank Rijmen  
HCIV

## An IRT model with a parameter-driven process for change

An IRT model for binary longitudinal data is presented. The heterogeneity between persons is taken into account by a continuous latent variable, as in common IRT models. Autodependencies are accounted for by assuming within-subject variability with respect to the parameters of the IRT model. More in particular, the parameters of the IRT model are governed by an unobserved or "hidden" homogeneous Markov process. The model includes the mixture linear logistic test model (Mislevy & Verhelst, 1990), the mixture Rasch model (Rost, 1990), and the Saltus model (Wilson, 1989) as specific instantiations. The model is applied to a longitudinal experiment on discontinuity in conservation acquisition (van der Maas, 1993).

---

## IMPS 2004 Abstracts

(Organized by surname of first author)

---

Fumiko Samejima  
University of Tennessee

LPE Graded Response Model, a Natural Expansion of the Logistic Positive Exponent Family of Models for Dichotomous Responses

Samejima (2000) has proposed the logistic positive exponent family of models (LPEF) for dichotomous responses which provides: 1) consistent ordering of the maximum likelihood estimates of ability  $\theta$  obtained from the individual response patterns, and 2) ways of dealing with two opposing principles depending on the psychological reality because of the third parameter, called acceleration parameter. This family of models includes the logistic model (Birnbaum, 1968) as a transition from one principle to the other.

The LPEF can be expanded naturally to a graded response model. In the present paper, it will be shown how it can be done.

Samejima (1969, 1972) has expanded dichotomous response models such as the normal ogive and logistic models (Lord & Novick, 1968) to graded response models, which belong to the homogeneous case (Samejima, 1972). A big difference of the present model, called LPE graded response model, is that it belongs to the heterogeneous case as opposed to the homogeneous case. The resultant model is observed and discussed with respect to its operating characteristics, basic functions, item response information functions, item information function, etc.

Samejima (1973, 1974) has also shown that those graded response models in the homogeneous case can be expanded to models for continuous responses. It will be demonstrated that this can be done with the LPE graded response model, in somewhat different way since it belongs to the heterogeneous case, but just as naturally as those models in the homogeneous case.

---

Albert Satorra  
Universitat Pompeu Fabra

Structural equation models for complementary data

---

Victoria Savalei  
UCLA

A Statistically Justified Pairwise ML Method for Incomplete Nonnormal Data: A Comparison with Direct ML and Pairwise ADF

This paper proposes a new approach to the statistical analysis of pairwise-present covariance structure data. The estimator is based on maximizing the complete data likelihood, while the associated test statistics and standard errors are corrected for misspecification. These corrections (known as Satorra-Bentler corrections) adjust for misspecification due to the use of the incorrect weight matrix resulting from pairwise computed covariances, as well as for possible nonnormality of the data. A Monte Carlo study is used to evaluate this methodology. The new pairwise method is compared to direct ML with the Yuan-Bentler corrections and to pairwise ADF. Data were generated from a four-factor model; sample size, missing mechanism, and proportion of missing data were varied. Model fit, relative accuracy of estimates, and their relative efficiency were assessed. The results generally favored direct ML over either of the pairwise methods, while ADF performed the worst. The inferior performance of the two pairwise methods relative to direct ML was primarily due to inflated test statistics. Among unusual findings was that MCAR data presented more problems for all methods than did MAR data; it is suggested that number of missing patterns might be more important than missing data mechanism, especially when data are nonnormal.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Kathleen Scalise  
UC Berkeley

A New Approach to Computer Adaptive Assessment with IRT Construct-Modeled Item Bundles (Testlets): An Application of the BEAR Assessment System

This paper will describe a new computer adaptive assessment approach being used in the UC Berkeley "Smart Homework" implementation of ChemQuery, an NSF-funded project. The approach uses testlet theory and IRT (facets model) to calibrate a series of testlets against a three-dimensional construct framework, generating item difficulties used to adaptively deliver customized homework sets to each student according to a new construct-based non-hierarchical delivery algorithm. Data has currently been collected from about 500 students on one dimension and analysis on the validity, reliability and fit characteristics of the approach is now underway. The proposed paper will discuss the approach and first results. Note that models and approaches used are an adaptation of the BEAR (Berkeley Evaluation and Assessment Research) Assessment System. The modeled framework is the Perspectives of Chemists, developed as a multi-dimensional performance construct for chemistry by the UC Berkeley chemistry department and the Mark Wilson group in quantitative measurement in education. The framework is embedded in a new chemistry curriculum, Living by Chemistry, currently being deployed to school districts nationally.

---

Daniel O. Segall  
DMDC

Modeling and Detecting Collaboration: A Multidimensional Item Response Theory Approach

This paper presents a new method for assessing consistency of test performance across two occasions, where on one occasion the level of performance may have resulted from a collaboration of two or more test-takers, and on the second occasion it is not. The new procedure is based on the application of Bayesian model assessment methodology to multidimensional item response theory. A simulation study based on a high-stakes multiple-aptitude test-battery was conducted to evaluate the proposed method. The new procedure was found to be superior to one based only on a discriminant analysis of final test-scores.

---

Eisuke Segawa  
University of Illinois, Chicago

A growth model for multilevel ordinal data

In order to analyze growth of a trait latent variable measured by ordinal items, we formulated multi-indicator growth models as special three-level hierarchical generalized linear models. Items are nested within a time-point and time-points are nested within a subject. They are special because they include factor analytic structure. Our models can analyze, not only data with item- and time-level missing observations, but also data whose time points are freely specified over subjects. Further, we implemented features useful for longitudinal analyses, "autoregressive error degree one" structure for the trait residuals and estimated time-scores. Our approach is Bayesian with Markov Chain and Monte Carlo and our models are implemented in BUGS. They are illustrated with two simulated data sets and one real data set with planned missing items that are missing within a scale.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Shohei Shimizu and Aapo Hyvarinen  
University of Helsinki

Yutaka Kano  
Osaka University

Exploratory causal inference using nonnormality

Path analysis is often applied to observational data to study causal structures.

The path analysis is an extension of regression analysis where many endogenous and exogenous variables can be analyzed simultaneously. Now path analysis is incorporated with factor analysis and allows latent variables in the model. The new framework is called structural equation modeling (SEM) and is a powerful tool of causal analysis. However, SEM is of confirmatory nature and researchers have to model true causal relationships based on background knowledge.

Lack of background knowledge often involves many difficulties such as inability of determining direction of a path and a serious bias of an estimate caused by unnoticed confounding variables.

These limitations mainly come from normal assumption in the SEM. Shimizu and Kano (2003) and Kano and Shimizu (2003) relaxed the restriction and showed that use of nonnormality extends conventional SEM and makes it possible to examine the following three models i) the determination of a direction of a path; ii) detection and adjustment of unobserved confounding variables; iii) a variant of bi-directed causal model. In this paper, we combine these three models and develop a new statistical method for exploratory causal inference using nonnormality of observed variables.

---

Klaas Sijtsma and Andries van der Ark,  
Tilburg University

Outlier detection in test and questionnaire data

Outliers are often identified as observations in the tails of the distribution of an interval variable. They are seen as a nuisance because they may exercise an extraordinary and undesirable influence on the outcome of the statistical analysis of one's data. For continuous variables and variables with many discrete categories the definition of outliers poses no real problems. This is different when a variable has only few, say, no more than five, categories. Such variables, or items, are typical of psychological tests and questionnaires. This paper proposes several simple and easy-to-use methods to identify and study outliers in test and questionnaire data based on N respondents who responded to J items, and compares these methods using simulated and real data sets with respect to their usefulness in data analysis. The methods are simple so as to be easily applicable in the research of colleagues who were not extensively trained in applied statistics but are motivated to solve problems such as those caused by unidentified outliers.

Recommendations are given with respect to the use of outlier detection methods in the analysis of real test and questionnaire data.

---

Mary Ann Simpson and Terry A. Ackerman  
University of North Carolina at Greensboro

Parameter recovery in Markov Monte Carlo Chain (MCMC) estimation of a generalized MIRT model

MCMC techniques have made it possible to estimate the parameters of complex multidimensional item response models. The current study is the first of a series of studies examining parameter recovery in MCMC estimation of the generalized multidimensional item response (GMIRT) model, a model in which the compensation between the abilities required by a task is assessed on a continuum from 0 to 1. The GMIRT model should prove useful in examining the relative importance of the cognitive components of a task-- for instance, to what degree does a mathematics problem require *problem solving skills*, and to what degree can a student "get by" with superior content knowledge. Item parameter recovery for the GMIRT model under varying conditions of sample size (N= 3000, 6000), item-pool size (n = 25, 50) and correlation among abilities ( $\Delta(2, 2) = 0, .30, .60$ ) will be assessed. A two-dimensional Rasch GMIRT model will be studied, and WINBUGS version 1.4 will be used for the MCMC simulations. Pilot results involving the Rasch GMIRT model with 30 items, 4000 examinees and  $\Delta(2, 2) = 0$  have been very promising. In these trials, the RMSE for the **b** parameter was .42 and that for the compensation parameter was .009.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Xin-Yuan Song and Sik-Yum Lee  
The Chinese University of Hong Kong

ML analysis of a multi-sample nonlinear structural equation model with fixed covariates and ordinal variables.

This paper discusses the ML estimation and model selection in the context of a multi-sample nonlinear structural equation model with fixed covariates and ordinal variables. To avoid computation of the complicated multiple integrals involved in the conditional expectations, a Monte Carlo EM (MCEM) algorithm is implemented to obtain the ML estimates, in which the E-step is completed with the help of a hybrid algorithm that combines the Gibbs sampler and the Metropolis-Hastings algorithm, and the M-step is completed by conditional maximization. Path sampling procedure for computing the observed data likelihood, which is involved in the computation of Bayesian information criterion (BIC) for model comparison, is discussed. The methodologies are illustrated with a real application to quality of life research.

---

Leonardo S. Sotaridona and Seung W. Choi  
CTB/McGraw-Hill

Rob R. Meijer  
University of Twente

The Effect of Misfitting Response Vectors on Item Calibration and Test Equating

Person fit statistics are very helpful in identifying examinees whose response behavior are not in accordance with the hypothesized item response theory model. Examinees having such behavior are said to have misfitting response vectors (*mr<sub>v</sub>*) or simply misfitting examinees. We hypothesized that the presence of *mr<sub>v</sub>* has an unpleasant effect of reducing the accuracy of the estimates of the item parameters and would yield misleading result on test equating. For example, the more response vectors in the dataset that are misfitting, the lesser precise the parameter estimates will be and the resulting equated parameters will not yield comparable results. An approach which is an application of person fit analysis is proposed that aimed to enhance both the accuracy of the estimated item parameters in an IRT framework and the quality of test equating results. We will show through empirical and simulation studies the practical usefulness of this approach.

---

Leonardo Sotaridona, Launa Hodgson, and Erica Connelly  
CTB/McGraw-Hill

Identifying Test Form

Standardized tests oftentimes consist of parallel forms. Cases occur, however, where examinees fail to reflect the test form taken in the answer sheet. This creates a problem for scoring, as different forms have different answer key configurations. The typical solution done in practice is to assign to the unknown answer sheet the test form that gives the highest score. However, no previous studies are available which explore the merit(s) and/or negative consequence(s) of this approach. This paper investigates the positive and negative consequences of the present approach. Additionally, two methods are proposed: one based on the likelihood of the observed response pattern given the test form; and the other one based on the response similarity between the observed response pattern and the answer keys. The new methods take into account the estimated ability of the examinee and the item parameters. Details of the methods will be illustrated and their usefulness will be investigated in empirical and simulation studies.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Douglas Steinley  
University of Illinois Urbana-Champaign

## Profiling Local Optima in K-means Clustering: Developing a Diagnostic Technique

Using the cluster generation procedure proposed by Steinley and Henson (2004), the performance of K-means clustering is investigated under the following scenarios: (a) different probabilities of cluster overlap, (b) different types of cluster overlap, (c) varying samples sizes, clusters, and dimensions, (d) different multivariate distributions of clusters, and (e) various multidimensional data structures. The results are evaluated in terms of the Hubert-Arabie adjusted Rand index and several observations concerning the performance of K-means clustering are made. Finally, the paper concludes with a the proposal of a diagnostic technique indicating when the partitioning given by a K-means cluster analysis can be trusted. By combining the information from several observable characteristics of the data (number of clusters, number of variables, sample size, etc.) with the prevalence of *unique* local optima in several thousand implementations of the K-means algorithm, a method capable of guiding key data analysis decisions is provided.

---

Yoshio Takane  
McGill University

Michael A. Hunter  
University of Victoria

Heungsun Hwang  
HEC Montreal

## An Improved Method for Generalized Structured Component Analysis

Generalized structured component analysis (GSCA) has been developed for path analysis with latent variables defined as exact linear combinations of observed variables. Unlike the partial least squares (PLS) approach, GSCA systematically minimizes a single global optimization criterion by an alternating least squares algorithm. In this paper we present an improved method for GSCA both conceptually and algorithmically. This improvement is afforded by rewriting all conceivable models to be fitted in the form of  $ZA = E$ , where  $Z$  is the data matrix (including both exogenous and endogenous variables),  $A$  is the matrix of path coefficients (structured in a variety of ways), and  $E$  is the matrix of residuals, and by minimizing  $SS(E)$  under a variety of constraints (zero constraints, linear equality and possibly inequality constraints, normalization constraints). Examples are given to illustrate the method.

---

Janneke te Marvelde  
Tilburg University

## A Comparison of Methods to Investigate an Invariant Ordering of Polytomous Items

An invariant item ordering (IIO) is an identical ordering of the items according to difficulty (ability tests) or popularity (personality or attitude tests) for every individual in the population, with the exception of possible ties. An IIO facilitates the interpretation of test results. In addition, several applications may require an IIO, such as the investigation of developmental theories, differential item functioning, aberrant response patterns, and the use of starting and stopping rules in test administration.

Previous research showed that only very restrictive parametric item response theory (IRT) models for polytomous item scores imply an IIO, which complicates the research in the context of the existing parametric IRT. In this presentation, I will compare three methods from a nonparametric IRT framework to investigate an IIO for polytomous items. First, two nonparametric models (the isotonic ordinal probability model and the strong double monotonicity model) that imply an IIO will be discussed. Second, the existing methods to investigate IIO are explained and extensions of these methods are discussed. Finally, sensitivity and specificity of these methods are presented.

Keywords: invariant item ordering, polytomous items, item response theory, nonparametric item response theory, strong double monotonicity model, isotonic ordinal probability model

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

David Thissen and Cheryl D. Hill  
Department of Psychology, UNC-CH

## Infinite Slope Estimates in Item Response Theory

It is well-known that the maximum likelihood estimates of the discrimination parameters for IRT models are infinite for data that lie on a boundary of the data space: For a set of items that appear to be "a perfect Guttman scale" given the data, the maximum likelihood estimates of the slope parameters are infinite and the trace lines are step functions. Discussion in this presentation focuses on data for which the maximum likelihood estimate of the discrimination parameter for a single item is infinite (given data that do not correspond to a Guttman scale pattern), while the discrimination parameter for another item has a (reasonable) finite estimate. Such data appear to be surprisingly likely when two items with three response categories each are fitted with Samejima's graded item response model. While less likely, it appears that it is possible to observe such results while fitting the graded model to pairs of items with more than three response categories. These observations lead to considerations of the relation between the parameter space and the data space for IRT models, and an analogy with the storied Heywood case of factor analysis.

---

Hideki Toyoda and Kentaro Nakamura  
Waseda University

## The reliability of students' evaluating university teaching: an analysis of four-facet data by generalizability model and structural equation modeling.

In our study, we examined reliability of students' evaluating university teaching. Through analyzing four-facet data (teaching, rater, viewpoint, occasion) by statistical model underlying generalizability theory, we identified and estimated the magnitude of each source of variation. We discussed the reliability from the point of view of generalizability coefficient and index of dependability, changing the number of raters and viewpoints. Conditions to keep sufficient accuracy of the evaluation were shown at each actual situation. We also showed a new method to evaluate the reliability at each level of facets by using structural equation modeling. It is tacitly assumed to be tau-equivalent tests in generalizability theory. But each level of facets may differ in terms of its mean or variance. By our new method, one can take into account these differences and examine the reliability of students' evaluating university teaching more accurately.

Keywords: evaluating university teaching, generalizability theory, reliability, structural equation modeling, rating data.

---

Hideki Toyoda and Akihiro Saito  
Waseda University

## Exploratory positioning analysis: Multi-mode multivariate analysis for semantic differential data

Semantic differential data (SD data) is a three-mode data consisting of "respondent" (e.g. John and Jane), "scale" (e.g. short-long and superior-inferior) and "target" (e.g. Harvard University and Columbia University). If the "target" consists of a combination of superordinate concept, such as university and school (e.g. Harvard Law School, Harvard Business School, Columbia Law School and Columbia Business School), the data is four-mode, or greater. If the "respondent" is multigroup (e.g. Harvard student and Columbia student), this is a three-mode multigroup data.

The purpose of this presentation is to propose an exploratory positioning analysis method for analyzing SD data as in the examples above, based on multi-mode multivariate statistical models. This new method can be easily implemented by using structural equation modeling (SEM) programs such as LISREL and CALIS, because the mean structure and covariance structure of this method can be expressed as submodels of SEM.

In this presentation, we will explain the process of the model establishment and positioning. Furthermore, actual application of this method to three-mode, three-mode multigroup and four-mode data will be shown.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Tsung-Hsun Tsai, Robert Sykes and Matthew Gordon  
CTB/McGraw-Hill

## Significant Characteristics of Anchor Items in the Common-Item, Nonequivalent Groups Design

The purpose of this study is to examine the effects on equating of varying significant aspects of anchor item sets (e.g., the number of multiple-choice and constructed-response items) of mixed-format examinations. Specifically, this study seeks to determine which anchor item set produces the best equating when the test forms contain mixed item types.

The anchor item set having similar content proportionality to the total test form is defined as the baseline anchor set. Stocking-Lord equating of two other anchor item sets is to be compared to the baseline anchor item set. The study utilizes a Reading state assessment. The old form and the three new forms were designed to meet the same content and statistical specifications. Each form was anchored to the preceding form. Each item is classified into one of six content standards. The reporting score is based on a raw score to scale score conversion table.

Results of the Poly-DIMTEST dimensionality analyses revealed that one of the new forms was not strictly unidimensional. The preliminary analyses on the data indicated that the difference between the parameters and the transformed estimates became smaller and the correlation increased, the more similar anchor item set content proportionality was to the entire test form.

---

Francis Tuerlinckx  
University of Leuven

## Estimating acquaintance volume with a hierarchical IRT model

Knowledge about the average acquaintance volume and the interindividual variation of the acquaintance volumes in the population is important in the study of social networks. In this paper, we tackle the estimation of both quantities by framing the research question as a capture-recapture problem. Second, a newly developed measurement instrument will be described. Next, we will propose a hierarchical IRT model (with acquaintances nested within respondents) to estimate the number of acquaintances for each respondent and the average number in the population. Finally, some results will be presented making use of Bayesian inferential techniques.

---

Rien van der Leeden, Marike Polak and Renske Doorenspleet  
Leiden University

## Scaling of Democracy: exploring changes over time

In political science, especially in comparative politics and international relations, the measurement and classification of democracy is a key topic. The most widely used democracy scale, positioning independent states relative to each other, has been developed within the Polity Project (Gurr, 1974 [Polity I]; Gur, Jagers and Moore, 1990 [Polity II]; Jagers and Gurr, 1995 [Polity III]; Marshall and Jagers, 2002 [Polity IV]). Gurr's scale relies on five indicator variables and uses an a priori coding and weighting scheme. Van der Leeden, Polak and Doorenspleet (2004) discuss a number of problems and ambiguities concerning Gurr's approach, and present an alternative scale of democracy based on a data analysis approach of the Polity IV data set.

In this paper we take the data analysis approach to the scaling of democracy one step further. In addition to the cross-sectional treatment of the problem of scaling democracy, we will explore the longitudinal nature of the Polity IV data set which contains data for most independent countries on annual basis since 1800. CATPCA was applied for scale construction in a longitudinal set-up, and results are compared with Gurr's approach regarding the ability to display the development of democracy over time.

Keywords: CATPCA, MRESPONDENCE, scale construction, longitudinal approach, democracy

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Wim J. van der Linden  
University of Twente

## Evaluating Equating Error in Observed-Score Equating

Traditionally, error in equating observed scores on two versions of a test is defined as the difference between the transformations that equates their quantiles in the sample and in the population of examinees. For example, this definition underlies the well-known approximation to the standard error of equating by Lord (1982). But it is argued that if the goal of equating is to map the observed-score distributions of examinees from one version of the test onto the distributions they would have had if they had taken another version, the criterion for evaluating equating error should be based on the difference between these distributions.

Two equivalent definitions of equating error based on this criterion are formulated. One definition is based on the difference between the transformation actually used and one that gives the correct equating; the other on the difference between the distribution functions for the equated scores resulting from these two transformations. It is shown how these definitions can be used to evaluate any method of observed score equating if the two tests fit an IRT model.

In an extensive empirical study, we evaluated the bias and MSE in the classical equipercentile method and two new conditional methods for the equating of two linear test as well as for these methods and a test-characteristic function (TCF) method for the equating of an adaptive test to a linear test. The results showed serious bias and MSE for the equipercentile and TCF method, whereas the conditional methods were virtually error free. It is argued that this result is due to the fact that the former are based on a single transformation that has to compromise between the observed-score distributions of different examinees, while the latter account for the differences between these distributions because they are based on a family of conditional transformations.

Key words: bias in equating; conditional equating; equating error; IRT observed-score equating; marginal equating; mean-squared error of equating; observed-score equating.

---

Wim J. van der Linden  
University of Twente

## Multilevel Modeling of Speed and Accuracy on Test Items

The literature on response times on tests shows various attempts to model response times in an IRT framework. The reason is an assumed interaction between the parameters that govern a person's response times and his/her response variable for the items. For one thing, it has been felt that response-time modeling should be based on the speed-accuracy tradeoff that has been the focus of much of the more substantive literature on response times (e.g. Luce, 1983). This assumption seems to entail the necessity of an accuracy/ability parameter in the response-time model. Besides, it is often assumed that more difficult items require more time. This assumption seems to entail the necessity of an item difficulty parameter in the response-time model. It is argued that these assumptions are wrong and that the only way to model such interactions is as a multilevel structure with higher-level components for both the person and item parameters.

In this presentation we focus particularly on the response-time aspect of the problem. We will discuss two versions of a normal and lognormal model for the response times with a parameter structure analogous to the 2PL response models in IRT. It is shown how the parameters can be estimated by a Markov chain Monte Carlo (MCMC) method (Gibbs sampler). We tested the validity of the model for a dataset from the adaptive version of a test from the Armed Services Vocational Aptitude Battery (ASVAB). Though the fit of the normal model was already satisfactory, the lognormal model showed improved fit in its upper tail, revealing a characteristic right skewness of the response time distributions. An equality constraint on the discrimination parameters of all items led only to negligible loss of validity.

In the final part of the paper we will discuss how to model response times on items with different response formats and indicate how our models can be used to improve the current practice of item calibration and ability estimation.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Joost R. van Ginkel  
Tilburg University

Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results

Missing data can be a serious problem in psychological research. Listwise deletion is the most commonly used method for dealing with missing data, although more sophisticated methods are available, based on multiple imputation. In general these methods work well, but are often too complicated to understand for social scientists that have no statistical background. In the context of test and questionnaire data several easier imputation methods have been developed as well.

The aim of this study was to compare the performance of the easier imputation methods to a more complicated method under several conditions. In a simulation study we compared the results of Mokken scale analysis and reliability analysis for complete test data with the results for imputed test data. Seven multiple imputation methods were studied: random imputation, two-way, two-way with normally distributed errors, response function with completely observed cases, response function with incompletely observed cases, CIMS with normally distributed errors and multivariate normal imputation.

The simple methods two-way with normally distributed errors and CIMS with normally distributed errors performed better than the more complicated method multivariate normal imputation. Furthermore, the bias of the results due to imputation were generally very small.

---

Cornelis M. van Putten, Arianne Smits, and Mark de Rooij  
Leiden University

Stability of category quantifications in multiple correspondence analysis under two ways to remove outlier dominance: A case study

HOMALS, a technique for multiple correspondence analysis of categorical data, generates dimensions dominated by outlying objects and categories when one or more objects have scored in combinations of categories not shared by the other objects. Two ways to remove outlier dominance were compared for a data set with outlier dominance on the first HOMALS dimension: disregarding the dimension dominated by outliers versus deleting the outlying objects. For each way the stability of the category quantifications for the resulting solution was investigated by bootstrap resampling with replacement. Confidence regions for these quantifications generated from the 1000 pseudo-samples of the bootstraps turned out to be comparable for both ways to remove outlier dominance.

Keywords: Categorical data, multiple correspondence analysis, optimal scaling, HOMALS, bootstrap resampling, confidence region, robustness

---

Peter van Rijn  
University of Amsterdam

Dynamic item response models

The modeling of individual change has been a topic of dispute in the literature on psychological measurement of change. Several models that allow for change over time have been developed in the framework of item response theory. Most of these models only contain parameters that specify group change. In this paper, a dynamic item response model for the analysis of multivariate dichotomous time series obtained from a single subject is presented. The model can be seen as a dynamic generalization of the Rasch model. The key difference between the common Rasch model and the dynamic version lies in the replacement of the latent person variable by a latent person process. With the current generalization, different types of dynamic structures can be modelled of which two are highlighted, namely the random walk and the auto-regressive model. Methods to obtain parameter estimates of these dynamic Rasch models are discussed and illustrated by a real data example.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Nathan A. Vandergrift  
University of North Carolina Chapel Hill

## Residuals based examination of fit for non-dynamically consistent models

Many longitudinal processes, particularly in psychology, evolve nonlinearly. Nonlinear longitudinal mixed effects models can be estimated as latent curve models (LCM) within the SEM framework. Unfortunately, LCM methodology cannot directly accommodate functions that are not dynamically consistent (not linear in the parameters). Therefore, LCMs of such functions are known, *a priori*, to be misspecified. Not merely in the sense that all models are approximations of reality, but even if the underlying longitudinal process is known, the LCM will still be wrong. Given this misspecification and that all chi-square based fit indices involve an assumption about the model being true in the population these type indices will often fail to provide good guidance for researchers. This paper will propose two related residual based metrics to evaluate model fit. One is based upon individual residuals and the other will aggregate over individuals, but within occasion of measurement. These residual metrics will provide an absolute account of how well the model is fitting the individual data through comparing model implied and individual trajectories. Potentially these metrics will be able to aid researchers in model selection; furthermore, they will provide a method for assessing model fit for individual curves rather than aggregate information.

---

Choulakian Vartan, Allard Jacques and Almhana Jalal  
Université de Moncton

## The Robustified Centroid Method.

This work addresses several aspects of Burt and Thurstone's centroid method in multivariate analysis. First, we describe how the centroid method and classical principal component analysis can be understood under a single framework. Then, we generalize the centroid method into an infinite family of discrete structure models that converge to classical principal component analysis. We also present robustified version of the discrete structure models. Finally, we present a genetic local algorithm to implement the robustified discrete structure models. Simulations are performed to assess the performance of the genetic local search algorithm and the robustness of the discrete structure method. Finally, a real dataset with missing values and multiple outliers is analyzed.

---

Ingmar Visser  
University of Amsterdam

## Multivariate latent Markov models for arbitrary length time series: An implementation and application

In this paper I consider latent Markov models with multiple indicators (MLMM). The MLMM can be seen as a generalization of 1) the standard latent or hidden Markov model (which in most applications has a single indicator) or 2) the latent class model (which has  $T=1$ , i.e., there are no repeated measurements). For long time series, say  $T>50$ , the usual estimation procedures for these models break down due to underflow issues. Therefore, procedures from the hidden Markov model literature are used. Analytical expressions for the score and information are derived for dependent (multivariate) and independent (multigroup) time series. When using gradients and Hessian, it is possible to use general purpose optimization routines to arrive at maximum likelihood estimates of the parameters. The derivations are quite general, and in particular, the model is not limited to categorical data but is also applicable to mixtures of continuous and categorical data. The aim of the model and program is to model individual multivariate time-series data in such a way as to be able to compare individuals and test for heterogeneity/homogeneity. That is, the aim is to investigate differences in intra-individual model structure.

---

Matthias von Davier  
Educational Testing Service

## Current Developments in Estimating Latent Distributions

This talk presents selected research and development efforts to extend the capabilities of methods for marginal estimation of latent distributions. The talk will discuss current examples of research on extending operational models, adapting the model to accommodate differences between subgroup response sets or individual response patterns, and on improving parameter estimation by eliminating technical approximations of previous approaches.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Matthias von Davier and Sandip Sinharay  
Educational Testing Service

## Application of Stochastic EM Methods to Latent Distribution Models

The reporting methods in large scale assessments like NAEP, IALS, TIMSS and PIRLS rely on virtually identical statistical models for generating multivariate posterior distributions of latent traits. Mislevy refers to these model as conditioning models while this approach may also be viewed as latent regression model, or, as Adams and coworkers view it, a multilevel IRT model. The first level of the model is a  $p$ -scale IRT measurement model defining the response probabilities on a set of items depending on a  $p$ -dimensional latent trait variable  $\theta=(\theta_1, \theta_2, \dots, \theta_p)$ . The second level models the conditional distribution of the latent trait  $\theta$  by a multivariate, multiple regression on a set of predictor variables, which consist of student, school and teacher variables in assessments like the ones mentioned above.

Maximum likelihood estimation of the parameters requires multivariate integrals to be evaluated, which can be done either by numerical integration or by some approximate solution of the integral. An algorithm for estimating this model is implemented in the MGROUP programs developed at the ETS. Mislevy (1984, 1985) introduced the EM algorithm for estimating this latent distribution model. The E-step is not straightforward, though, especially if multivariate latent distribution are involved. The methods implemented in the Bgroup (Mislevy & Sheehan, 1989) and Cgroup (Thomas, 1993) software are not entirely satisfactory and there is scope to improve. They either use numerical integration and are feasible up to only two dimensions, or they handle multivariate cases using Laplace approximations and rely on assumptions that may not always be met.

This paper presents a comparison of the operational methods, used in NAEP and other assessments since 1993, with an implementation stochastic EM methods utilizing importance sampling. Stochastic EM carries out the E-step by drawing a sample from an appropriately chosen distribution. The paper presents a study whether stochastic EM methods provide more accurate results than existing methods of estimation. We will outline the stochastic EM method and compare its results with currently available algorithms. Results based on simulated data modeled after a national 5-dimensional assessment (comparison to multivariate Cgroup) and data from a national large scale assessment (comparison to bivariate Bgroup and Cgroup) will be presented.

---

Matthias von Davier and Alina A. von Davier  
Educational Testing Service

## A Unified Framework for IRT Scale Linking and Scale Transformations

This presentation examines IRT scale transformations and IRT scale linking methods used in the Non-Equivalent Groups with Anchor Test design (NEAT) to link two tests, X and Y. More exactly, we propose a unifying approach to the commonly used IRT linking methods: "mean-mean", "mean-var linking", "concurrent calibration," Lord and Stocking, and Haebara characteristic curves approaches, and fixed-item parameters scale linkage. The main idea is to view any linking procedure as a restriction on the item parameter space. Once this is understood, a rewriting of the log-likelihood function, and an accordingly implemented maximization procedure of the log-likelihood function under linear (or non-linear restrictions) will accomplish the linking. The proposed method uses Lagrange multipliers and is general enough to cover the usual Item Response Models, the one parameter logistic (1PL) model, 2PL, and 3PL models as well as polytomous unidimensional IRT models like the generalized partial credit model.

Keywords: Item Response Models, Scale Transformation, Test Linking, Non-Equivalent Groups with Anchor Test Design, Nonlinear restrictions, Maximization procedures, Lagrange Multipliers.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Xiang Bo Wang  
The College Board

Louis Mi Wang  
Pennsbury High School

Investigating the Properties of the Reliability Estimation Formula by Gulliksen (1950)

In his 1950 book *Theory of Mental Tests*, Gulliksen devised the formula to estimate test reliability using item  $P+$  and point-biserial correlations.

This formula is further explicated by Crocker and Algina (1986) and Allen and Yen (1979). The advantage of this formula is that unlike KR-20 or KR-21, it does not require examinee score variance to estimate test reliability. This strength is very useful because it helps test makers predict reliability of new parallel test forms without actually administering it.

The object of this paper is to explicate some of the less known mathematical properties of this formula, including the conditions in which this formula produces negative reliability coefficients and  $P+$  has little effect on reliability. A combination of detailed mathematical proofs and systematic simulations will be presented.

---

Xiang Bo Wang, Wayne Camara, Jennifer Kobrin, and Ying Zhou  
The College Board

Accounting for Factors Affecting the SAT Performance of Asian American and Pacific Islander Students

Recently, there has been increased interest in the SAT performance of Asian American and Pacific Islander (AAPI) students as well as the factors that affect their test performance. Based on the historical data from the College Board over the past 11 years, the object of this paper is two-fold: first to describe the overall SAT performance trends of AAPI students as well as the those trends conditioned on a number of demographic variables, including students' social economic status, parental education, school type, first language and best language use, gender and so on. Secondly, this paper will explore the intrinsic relationships among SAT performance and the demographic variables using various data analysis and model methodologies including both regression and structural equation modeling. In addition to presenting empirical results, discussion will also be devoted to several methodological issues, such as theory formulation and model fit and selection.

---

Matthijs J. Warrens  
Leiden University

On ordering properties of classical optimal scaling

Guttman (1941) proposed a generalization of principal component analysis to multivariate categorical data. The method is known under many names but will be referred to, following McDonald (1983), as classical optimal scaling (abbreviated as COS). The method has some interesting properties when applied to categorical scores with an assumed latent dominance structure. Such data are usually analyzed by unidimensional cumulative item response models, either parametric or nonparametric (Mokken, 1971).

The work of Schriever (1985) on the dichotomous case is extended to the polytomous case. Conditions are specified under which the category scores of the unidimensional COS are ordered. The results are demonstrated on generated data, conforming to the polytomous perfect scale and the graded response model.

Keywords: Classical optimal scaling, ordered category scores

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Greg Welch and Kevin H. Kim  
University of Pittsburgh

An Evaluation of the Fleishman Transformation for Simulating Non-Normal Data in Structural Equation Modeling

In structural equation modeling, the Fleishman (1978) and Vale & Marelli (1983) transformations are often used to generate data with specified distributional parameters.

Maximum likelihood (ML) estimation is known to be robust to types of violations of normality. Unfortunately, the robustness of ML estimation to data obtained from the aforementioned transformations is unknown. To examine this issue, this study utilized 15 conditions falling into classes defined by transformations of the: (1) errors and/or factors; (2) variables; and (3) errors in a misspecified model. Within each of the classes, various levels of non-normality were imposed on a 3-factor confirmatory factor analysis model. Model test statistics and standard errors of the normal theory estimator (ML) and the rescaled version (ML-SB) (Satorra & Bentler, 1988, 1994) were compared. Class 1 conditions resulted in higher model rejection rates for the normal theory ML than for ML-SB. Rejection rates were very similar in class 2 and 3 conditions. Standard error estimates of factor loadings were different for each method within each class of conditions. The robustness of normal theory ML estimation to class 1 transformations can be overcome by applying alternative transformation methods, such as that put forth by Hu, Bentler, & Kano (1992).

---

Victor L. Willson and Zhongmiao Wang  
Texas A&M University

Indifference Regions for Goodness of Fit Indices in SEM

While maximum likelihood solutions provide optimal solutions under the assumptions of the method, point estimates do not necessarily provide sufficient information to evaluate either model adequacy or parameter adequacy. This paper details the results of a simulation study of the distributional characteristics of various goodness of fit indices such as the comparative fit index, normed fit index, and Bayesian Information Criterion among others with respect to parameter spaces for several different structural equation models. Sample size and parameter size and redundancy were investigated in an iterative graphically-based procedure using SAS macros. Initial results are shown that illustrate indifference regions (using values for high and adequate model fit according to the various fit indices). These regions vary greatly from index to index for the same data sets, as shown by various graphs.

---

Carol M. Woods & David M. Thissen  
University of North Carolina at Chapel Hill

Item Response Theory with Estimation of the Latent Population Distribution Using Spline-Based Densities

This research introduces a new method for fitting item response theory models with the latent population distribution estimated from the data using splines. Previous item parameter estimation systems have most often assumed a normal population distribution, which may be a mis-specification. Alternatives that use different functional forms, or a histogram, to represent the population distribution have not been unambiguously successful. A spline-based density estimation system provides a flexible alternative. A simulation study shows that the procedure is feasible in practice, and that when the latent distribution is not well approximated as normal, 2PL item parameter estimates can be improved over what they would be from the normal model. An example using responses to items of the Maudsley Obsessional Compulsive Inventory illustrates results that may be obtained with this method.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Werner Wothke, George Burket, LiSue Chen, Furong Gao, Lianghua Shu, and  
Mike Chia  
CTB/McGraw-Hill

Multimodal Likelihoods in IRT-Based Response-Pattern Scoring. Will The Real Maximum Likelihood Score Please Stand Up?

In psychological and educational testing with item response theory models, it has been known for some time that the likelihood function of a respondent's ability may have multiple modes, flat maxima, or both. These conditions, often associated with guessing of multiple-choice questions, can introduce uncertainty and bias to ability estimation by maximum likelihood when standard Newton solutions are employed. This paper evaluates the performance of several maximization methods, including initial (grid) searches probing the function slopes, simulated annealing, exhaustive likelihood evaluations, and the standard Newton algorithm. In extensive studies, involving several hundred thousand records of both generated and real data, the algorithms were evaluated with respect to precision and speed. Three methods, exhaustive search, simulated annealing, and grid search followed by Newton steps, all yielded maximum likelihood estimates at the required precision. At today's computer speeds, either of these algorithms is fast enough for high-volume response-pattern scoring.

---

Margaret L Wu and Raymond J Adams  
University of Melbourne

User-defined Fit Statistics for the RCML models

Wu (1997) derived a fit statistics for the random coefficients multinomial logit model (RCML; Adams & Wilson, 1996), and this was implemented in the ConQuest software (Wu, Adams & Wilson, 1998). This fit statistics is based on those presented by Wright and Masters (1982). The Wright and Masters statistics were extended in two ways. First they were extended for application to a more generalised model, providing fit at the level of the parameter rather than at the level of the 'item'. Second, the Wright and Masters statistics were developed for use with unconditional maximum likelihood estimates, and so they had to be extended for use with marginal maximum likelihood estimates.

In this paper, the Wu fit statistics are further generalised to provide user-defined fit tests. For example, one can devise a fit statistic for scores on groups of items instead of for individual items. Similarly, rater groups can be assessed in terms of the fit to the model. Fit statistics can be defined through the specification of a "design matrix", allowing for flexibility in terms of how items are combined. The software ConQuest now incorporates this feature of fit testing for variables constructed by the user.

---

Xueli Xu  
University of Illinois, Urbana-Champaign

Young-Sun Lee  
Teachers College, Columbia University

Jeff Douglas  
University of Illinois, Urbana-Champaign.

Nonparametric IRT Equating

IRT-based equating is developed for nonparametric models. Under the IRT framework, the items from two different test forms are first calibrated on the same scale through possibly different transformation functions, and then the IRT-based equating is to find the relationship between the transformations. For a 3 parameter logistic model, equating utilizes the observation that  $P(\theta; a, b, c) = P(A\theta + B; a/A, Ab + B, c)$ . Thus, it is natural to use a simple linear function to find the equating relationship. However, this may not be appropriate if the parametric form of ICCs is not correct. To address this, a nonparametric IRT equating method is proposed. In this research, many methods could be used to estimate nonparametric ICCs, a monotone spline function is used to estimate the equating function. This method is studied with simulation and real data. This method can also be considered as a general framework for IRT-based equating with the parametric linear transformation as a special case.

---

**IMPS 2004 Abstracts**  
(Organized by surname of first author)

---

Hirokazu Yanagihara  
University of Tsukuba

Corrected Version of AIC for Selecting Multivariate Normal Linear Regression Models in a General Nonnormal Case

This paper deals with the bias correction of AIC for selecting variables in multivariate normal linear regression models when the true distribution of observation is an unknown nonnormal distribution. We propose a corrected version of AIC which is partially constructed by predicted residuals and adjusted to the exact unbiased estimator of the risk when the candidate model includes the true model. It is pointed out that the influences of nonnormality in the bias of our criterion are smaller than the ones in AIC and TIC. We verify that our criterion is better than the AIC, TIC and EIC by conducting numerical experiments.

---

Lihua Yao and Richard D. Schwarz  
CTB/McGraw-Hill

Multidimensional Models for Tests Consisting of Mixed Item Types

Multidimensional IRT models have been proposed for better understanding the dimensional structure of data or to define diagnostic profiles of student learning. A multidimensional partial-credit model is presented that is a generalization of those proposed to date along with a noncompensatory multidimensional model for multiple-choice data. Estimation of these models using MCMC methods is discussed. Many assessment programs have a mixture of item types in which multiple-choice and constructed-response are administered together. A substantive example is presented in which the dimensional structure of tests containing mixed item types is examined. Goodness-of-fit testing under various model formulations is discussed.

---

Hsiu-Ting Yu and Carolyn J. Anderson  
University of Illinois

Empirical Comparisons of Estimates of Item Response Theory Models and Log-Multiplicative Association Models

The theoretical connections and justifications linking Item Responses Theory (IRT) models and log-multiplicative association models (LMA) are given in Holland (1990) and Anderson and Yu (2003). In this talk, we present the results of a study investigating the extent to which the theoretical connections hold empirically under different assumptions of regarding the distribution for the latent trait. In the standard IRT models, the underlying latent traits are assumed to be normally distributed, but in the LMA models, the conditional distribution of the latent variable given a response pattern is assumed to be normal. The estimated parameters obtained under different assumptions for the latent distributions will be compared using simulated data. Data are simulated according to standard IRT models and parameters are estimated by fitting LMA and IRT models to the simulated data. The accuracy of the estimates will be compared in terms of bias and error variances.

---

Ke-Hai Yuan  
University of Notre Dame

Peter M. Bentler  
University of California, Los Angeles

Standard Errors and Asymptotic Robustness in Multilevel Models with Distributional Violations

Data in social and behavioral sciences are often hierarchically organized. Multilevel statistical methodology was developed to analyze such data. Most of the procedures for analyzing multilevel data are derived from maximum likelihood based on the normal distribution assumption. Standard errors for parameter estimates in these procedures are obtained from the corresponding information matrix. Because practical data typically contain heterogeneous marginal skewnesses and kurtoses, it is interesting to know how nonnormally distributed data affect the standard errors of parameter estimates in a two-level structural equation model. Specifically, we study how skewness and kurtosis in one level affect standard errors of parameter estimates within its level and outside its level. We also show that, parallel to asymptotic robustness theory in conventional factor analysis, conditions exist for asymptotic robustness of standard errors in a multilevel factor analysis model.

---

# IMPS 2004 Abstracts

(Organized by surname of first author)

---

Duan Zhang and Victor L. Willson  
Texas A&M University

## Empirical Power and Type I Error Rates for Cross-Level Interactions in Multilevel Analysis

Multilevel analysis using hierarchical linear modeling (HLM) has been promoted in recent years as the preferred method for nested designs such as students in classes in schools. It can be theoretically specified as a special case of the more general structural equation model (SEM) even though its representation in SEM might be very complex.

This simulation study investigated the power differences among HLM, a hybrid approach of HLM and SEM, and a residual SEM technique from the perspective of the second level regression weight. We explored how estimation variation influenced the power of the models. All data were generated in SAS-PC and analyzed by PROC MIXED and PROC CALIS.

It appeared that for moderate first-level sample size the power of all three models increased with the increasing second level regression weight. HLM was affected more by the second level regression weight, unlike the other two methods. This trend was not found for the bigger first-level sample size given the same number of second level units. The discussion aimed at the possible explanation of the situation and its interpretation for the real world research.

---

Guangjian Zhang  
The Ohio State University

## Bootstrapping Dynamic Factor Analysis

Dynamic factor analysis is a useful tool for investigating intra-individual change over time. Evaluating dynamic factor analysis models, however, is much more difficult than evaluating the usual between-person factor analysis models. This is due to the difficulty of deriving the correct likelihood function for dynamic factor analysis models. Consequently, standard errors and fit measures obtained from normal theory maximum likelihood estimation may be wrong if the likelihood functions are intractable. We propose two bootstrap methods to obtain standard errors and fit measures for dynamic factor analysis models. The parametric bootstrap is like a Monte Carlo study in which population parameters are estimates obtained from the data. The moving block bootstrap forms exchangeable blocks from the data, and these blocks are then resampled with replacement. As an illustration, the proposed methods will be used to analyze mood data.

---

Wen Zhang and James O. Ramsay  
McGill University

## Bivariate Functional Regularization for the Detection of Cortical Region Transition

Detection of transition regions of adjacent areas and their subdivisions within the human cerebral cortex always poses difficulties in the area of neuroimaging due to the inhomogeneities in the patterns of cell arrangements.

In this paper we propose a bivariate functional regularization of both test lines (traverses) along the cortical surface and the pixel intensities (profiles) along one traverse. We expand the restrictions on covariate functions  $z$  (the structural information of profiles) and allow  $z$  takes values  $z(t; x)$  varying over both arguments, where  $t$  and  $x$  represent profiles and traverses respectively. We approximate the bivariate density functions with basis expansion and regularizing the fit with roughness penalties. What's new here is we restrict bases toward both arguments simultaneously to regularize the curvature of the surface.

As an alternative to the roughness penalty, we've also tried to apply the total variation which controls the size of jumps as well as the boundaries of different cortical areas. This approach provided evidence for a certain number of peaks in the cell density despite cortical convexities, concavities, and other staining inhomogeneities.

Keywords: Functional data analysis, bivariate functional regularization, cerebral cortex, cytoarchitecture, transition regions, traverses, profiles.

---