CONTENTS

Chronological index (according to date, session and presentation)

TUESDAY JULY 5

Opening, 9.30h-10.00h

Keynote Speaker, 10.00h-11.00h

Missing Data, Multiple Imputation and Causal Inference

Joseph L. Schafer

In recent years, we have seen a growing consensus among statisticians with respect to causal inference. For the most part, causal effects are now defined in terms of potential outcomes, the hypothetical responses that would be observed for a study participant under different experimental treatments. Because only one potential outcome is observed for each participant, the task of causal inference becomes a missing-data problem. The most popular techniques for causal inference in the potential-outcomes framework are based on propensity scores, which come from a probability model describing how the experimental treatments were assigned. An alternative approach is to apply multiple imputation to the incomplete dataset, impute the missing potential outcomes under a plausible model and estimate the causal effects directly. In this talk, we will review the key ideas of causal inference with potential outcomes, contrasting this approach with the usual regression-based methods that have long dominated the social and behavioral sciences. Drawing upon the statistical literature from missing data, we will review a variety of adjustment techniques for estimating average causal effects in a population, including direct maximum-likelihood estimation, imputation and weighting. The motivating example involves estimating the causal effects of various weight-control strategies (dieting and exercise) from a longitudinal survey of American youth.

Symposia and sessions, 11.30h-13.10h

Divided by a Common Language

John C. Gower

The common language is that of matrix decompositions (spectral and singular value) and least-squares formulations. These divide statistical and data-analytical methods of various kinds that are the basis for analyzing arrays, two-way tables and data matrices. The substantive statistical differences, rather than computational similarities, will be emphasized between methods such as principal components analysis, biadditive models for two-way tables, multiple correspondence analysis and correspondence analysis. The important distinctions are between two-way tables classifying a single response variable (which may be quantitative or categorical), contingency tables and multivariate data matrices of categorical and/or quantitative variables.

Analysis of Heterogeneous Data through an Extension of Multiple Factor Analysis: Application to Survey Data with Quantitative, Qualitative and Free-text Answers

Mónica Bécue-Bertaut and Jérôme Pagès

In a socio-economic survey, it is useful to tackle a complex topic by using complementary closed and open-ended questions. In such situations we are faced with the simultaneous analysis of quantitative, qualitative and frequency variables, in particular when we aim at applying the classical strategy combining a principal axes method and a clustering step. Multiple factor analysis (MFA) tackles the analysis of tables in which a set of individuals is described by several groups of variables. This method includes a balancing of inertias between the different groups and, in this way, it is possible to take into account, as active elements, both quantitative and qualitative groups.
The introduction of contingency tables as groups of variables (where one table defines one group) is problematic in the correspondence analysis context: when the row margins are different, the weights of rows vary across the groups of variables. We propose here a solution to this problem through the method that we have called multiple factor analysis for contingency tables (MFACT).
By combining MFA and MFACT, we can deal with quantitative, indicator and contingency tables in a same principal axes method. Then, it is possible to operate a clustering of the individuals starting from their principal coordinates. Thus we can obtain a partition of the individuals based on their answers to the three types of questions.
An example, extracted from a large post-electoral survey carried out in Belgium, will illustrate the methodology.

Assessment Of Multiple Correspondence Analysis

Ludovic Lebart

Multiple correspondence analysis (MCA) provides useful visualisations (e.g. in socio-economic surveys, in marketing), highlighting associations and patterns between several categorical variables. However, the outputs (parameters, graphical displays) are often difficult to assess. In an era of computer intensive techniques, we cannot be satisfied just with the criterion of “interpretability” of the results that was frequently used during the first decades of data analysis techniques. In both cases of simple and multiple correspondence analysis, most of the theoretical results provided by mathematical statistics are not applicable.
We will focus then on the two following issues.
- External validation, involving external data or meta-data (usually considered as supplementary elements, also called passive or illustrative) and allowing for some classical statistical tests, including cross-validation procedures in the scope of supervised learning.
- Internal validation, based on re-sampling techniques such as the bootstrap and other Monte Carlo methods. The bootstrap replication scheme allows for drawing confidence ellipses or convex hulls for both active and supplementary categories, as well as supplementary continuous variables. The *partial bootstrap* that considers the replicates as supplementary elements, is easy to apply, but its validity is subject to some conditions. The *total bootstrap* is a much more conservative validation procedure, in which each replication leads to a separate MCA, but the absence of a common subspace of reference may induce a pessimistic view of the results, even if some specific corrections are carried out.

Scaling Monotone Latent Variable Models With Multiple Correspondence Analysis

Matthijs J. Warrens and Willem J. Heiser

In this paper we consider the application of multiple correspondence analysis as a method of scaling. For this type of application some monotone latent variable models are considered, all of which are characterized by item and person parameters. We shall point out what multidimensional information on the parameters can be obtained from applying multiple correspondence analysis when the monotone latent variable models are used as gauges. The term “gauge” was introduced by Gifi to denote benchmark data sets, or mechanisms to generate such data sets, for studying the behaviour of data analysis techniques such as simple and multiple correspondence analysis. Some old results on the one-dimensional case are reviewed, but many new insights and results on the multi-dimensional case will be presented as well.

Assessing the Quality of Ordered Categorical Data

Jörg Blasius

When dealing with social surveys, one of the main problems is the quality of the data. By “quality of data” I am referring to the degree to which respondents understand the content of the questions and the degree to which they were able to give reliable answers. These kinds of questions about the quality of data are seldom discussed, in most cases it is assumed that all respondents were fully aware of the content of the items. However, often respondents do not fully understand the content of the items, for example, they do not notice the direction of the questions. These misunderstandings can not be considered as distributed at random, they may be correlated with educational level or with other variables related to the substantive questions of interest. The problem to be discussed in the paper is to subdivide the sample into two (or more) groups: one group contains the respondents that understand the questions whereas the other group does not. A similar problem arises in international comparisons – in some countries the respondents are able to answer the questions in the intended way, in other countries the respondents are not able to do so.

To avoid misinterpretations based on “bad items”, it is necessary to screen the data for possible systematic misunderstandings before further investigation. Restricting the examples here to ordered categorical data, the paper will discuss strategies of evaluating the items to detect groups of respondents that are not able to handle the items in the intended way.

Extensions of Two-Way Imputation to Multidimensionality

Joost R. van Ginkel, L. Andries van der Ark, Klaas Sijtsma

Two-way imputation is a simple method for missing-data handling. Earlier we showed that this method recovered Cronbach's (1951) alpha, Loevinger's (1948) *H*, and the cluster solution from Mokken (1971) Scale Analysis well for unidimensional datasets and two-dimensional datasets with a positive correlation between the two dimensions. In the current study the aim was to generalize two-way imputation to multidimensional settings. Four-dimensional data were simulated with varying correlations between the latent traits, and item scores were deleted according to three different missingness mechanisms: missing completely at random, missing at random, and not missing at random. Four new versions of two-way imputation were studied, which all aimed to correct for multidimensionality. The performance of these methods was compared with the original two-way method. Listwise deletion and multivariate normal imputation were used as lower and upper benchmark, respectively. Cronbach's alpha, Loevinger's coefficient *H*, and the item cluster solutions from Mokken scale analysis and principal components analysis of the complete data were compared with the corresponding results based on the data including imputed scores.

Exploring the Heterogeneity of Adolescent Depression through Latent Class Analysis

Charity Morgan, Bruce Compas, Stuart Hauser, Jerome Kagan, Nancy Snidman and Donald Rubin

A latent class model postulates the existence of a missing indicator variable which denotes class membership. Latent class models are useful tools for data that contain categorical variables and arise from heterogeneous populations. Psychological data sets are often characterized by these features, and as a result, latent class analyses are enjoying increasing popularity in psychological applications.

One potential application is major depressive disorder (MDD). Our current research deals with a sample of depressed adolescents and hypothesizes that there are multiple distinct, biopsychosocial profiles underlying adolescent-onset depression. In this presentation, we describe the formulation of a latent class model to test this hypothesis. Using EM algorithms and MCMC routines, we impute the missing MDD index given the observed biological and psychosocial variables.

Working Memory Impairments in Schizophrenia Patients and their Relatives: A Bayesian Item Response Theory Analysis

Samantha Cook

With collaboration of: John Barnard, Yungtai Lo, Donald B. Rubin, Michael J. Coleman, Philip S. Holzman, Deborah J. Levy, and Steven Matthysse

Several studies have shown that spatial working memory is impaired in schizophrenia patients. In our study, schizophrenia patients and normal controls participated in a memory test designed to measure both spatial and object working memory. The test items were designed to have differing levels of difficulty, making standard analyses inappropriate. The data were analyzed using a bivariate Bayesian Item Response Theory (IRT) model.
Item response theory is a method for analyzing test scores in which the test items themselves are analyzed in addition to the test-takers' abilities. Analyzing the data in this way accounts for the fact that the questions were not all equally difficult, and also produces results which are more generalizable and less test-dependent. Earlier work in item response theory has focused on univariate abilities; here we extend these methods to include bivariate abilities. Using the Bayesian framework improves upon standard EM methods for IRT models by producing standard error estimates which more accurately represent uncertainty about the parameters. Using the Bayesian IRT model, we find that schizophrenia patients appear to be impaired on both object and spatial working memory, and that the degree of impairment is the same on average for both tasks.
We also consider models that include data from clinically unaffected first-degree relatives of schizophrenia patients.

Multilevel Latent Growth Modeling With Dropout Depending On The Latent Trait: A Cluster Randomized Sex Education Intervention

Anders Skrondal, and Sophia Rabe-Hesketh

One approach to dropout is to specify a joint model for longitudinal substantive and dropout processes where dropout may depend on the contemporaneous substantive response (observed or missing) and/or lagged responses (e.g. Hausman and Wise, 1979; Diggle and Kenward, 1994). We extend these models in several ways. First, we consider latent growth models with a fallibly and possibly coarsely measured latent trait as substantive response. Second, we develop multilevel versions to accommodate dependence induced by clustered data. Third, dropout may depend on the latent trait instead of a fallible measure and can also be intermittent. The proposed multilevel latent growth model with latent-trait-dependent dropout is applied to a sex education trial where the intervention was randomized at the school level. The latent trait of interest, contraceptive communication, was measured by students' answers to three questions rated on an ordinal scale and is modeled using a logistic item response model. The effect of the intervention on contraceptive communication is modeled using a three-level latent growth model with students at level 2 and schools at level 3. Dropout is modeled using logistic regression models with the contemporaneous value of contraceptive communication as well as lagged values as covariates. Estimated intervention effects 'corrected' for dropout rest on unverifiable assumptions and dropout modelling may best be viewed as a sensitivity analysis. Whatever approach taken to dropout, the challenges of measurement error and clustered data cannot be brushed under the carpet.
All models considered are special cases of the GLLAMM framework and can be estimated in gllamm by maximum likelihood using adaptive quadrature.

Statistical properties of Multivariate Imputation by Chained Equations (MICE)

Stef van Buuren

Multivariate Imputation by Chained Equations (MICE) is a general sampling method for creating multiple imputations in multivariate incomplete data. The starting point is the distribution of each variable given all other variables. See www.multiple-imputation.com for more details. The theoretical weakness of MICE is that the specified conditional densities can be incompatible, and therefore the stationary distribution to which the sampler attempts to converge may not exist. Van Buuren *et al*. investigated the consequences of this problem by means of simulation. Missing data were created under four different MAR missing data mechanisms, covering both compatible and incompatible models. The results indicate that MICE produces essentially unbiased estimates with appropriate coverage in the simple cases investigated, even for the incompatible models. It thus appears that, despite the theoretical weaknesses, the actual performance of MICE for multivariate imputation can be quite good.

Reference:
Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M., Rubin, D.B. Fully Condi-tional Specification in Multivariate Imputation. Submitted for publication.

Sensory Difference Tests

Per Brockhoff

Simple sensory difference tests such as e.g. triangle tests are still used extensively in the sensory field both in research and in the industry. Ideas of Thurstonian modelling and d-prime estimation are only slowly entering industrial practice. In the sensometric literature there has been a debate about the handling of replicated difference testing. One of the key suggestions is to use beta-binomial models, but also other approaches such as generalized linear mixed models have been discussed.

Zero or Nothing? Scales in Time-Intensity Studies

Paul Eilers

In time-intensity (TI) research, panel members indicate, with a computer mouse or slider, the strength of a sensation (taste, smell, substance firmness, ...) during a set interval. The results are time series of about one minute long, sampled every second (or more frequently). TI curves can be rather erratic and their analysis is a real challenge. There is a growing tendency towards parametric models, some based on differential equations. When developing such models, assumptions about the nature of intensity become crucial. Can we interpret it as a concentration, and does zero mean nothing, or can it better be interpreted on a logarithmic scale, on which zero means a lower detection threshold? I will discuss this subject and do suggestions for specific experiments. Part of this work was done jointly with Garmt Dijksterhuis and Per Moeller.

Sensory Profiling

Pascal Schlich

Sensory profiling, also known as descriptive analysis, is the process during which a panel of trained assessors score several sensory attributes on a number of products to be compared. The data results in a 3-way table crossing panellists, products and attributes. The data analysis could be directed to describe either the panellist performances or the product differences. This analysis can be done by attribute and multivariately, that is by sets of attributes defining different types of sensations, such as appearance, odour, texture, taste, flavour... Since the choice of the attributes is, to some extent, arbitrary, the choice of the multivariate approach is compulsory. However, since individual differences in perception and understanding of these attributes can never be completely removed by training, the attributes should be considered as subjective indicators of latent sensory variables. The paper will discuss the use of Canonical Variate Analysis and STATIS to take into account the specificity of profiling data and the needs of both panel leader and product developer.

Preference Testing And Mapping

Anne Hasted

This talk will concentrate on sequential monadic testing of more than two products using Central location testing. The need for warm up samples, designs balanced for carry over effects and care in asking questions will be outlined. Internal and external preference mapping techniques are widely used to assist in New Product development. Some useful extensions such as Partial Least Squares and Cluster based mapping will be discussed and the use of contour plotting to position products most effectively will be illustrated.

Capturing Heterogeneous Utility Curves: A Bayesian Spline Approach

Jin Gyo Kim, Ulrich Menzefricke, and Fred Feinberg

Decision-makers often weight units of a valued attribute unequally, resulting in non-linear or irregularly shaped utility functions. Although various functional specifications for utility can be imposed exogenously, this approach can be ad hoc, tedious, and require imposing various metrics to decide which specification is ‘best’. Using truncated power basis splines in a general additive regression framework, we show how to incorporate individual-level, flexibly-shaped utility functions in choice models. Heterogeneity in spline knot configuration across individuals is accomplished via birth-death processes and reversible jump methods (Denison et al. 1998; Green 1995). Our formulation is appropriate whenever individual-level montonicity is justified theoretically and empirically, but linearity is typically not (e.g., in the estimation of price response), and we show how exogenous constraints on utility shape can be accommodated in general. The method is illustrated in a conjoint application where all covariates are splined simultaneously, and in three panel data sets each of which has a single price spline. Results indicate that piecewise linear splines with a modest number of knots fit these data well, substantially better than heterogeneous linear and log-linear *a priori* specifications. In terms of price response specifically, we find that although aggregate market-level curves can be nearly linear or log-linear, individuals often deviate widely from either. Hold-out prediction improvement over standard models ranges from 5.9% to 14.3% in the scanner applications and exceeds 20% in the conjoint study. Moreover, ‘optimal’ profiles and policies can differ markedly when strict linearity is relaxed to monotonicity using splines.

An Integrated Model of Choice and Response Time with Applications to Conjoint Analysis

Thomas Otter, Greg M. Allenby, Trish van Zandt

Response times and choices are jointly modeled using a time varying Poisson process where choice alternatives generate signals in the consumer's mind. A choice is made when a sufficient number of signals are generated, and the response time corresponds to the time taken for the fastest process to reach a threshold. The integrated model nests the standard logit model, and is shown to improve part-worth estimates in conjoint analysis.

A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules

Timothy J. Gilbride and Gregory M. Allenby

Many theories of consumer behavior involve thresholds and discontinuities. In this paper, we investigate consumers' use of screening rules as part of a discrete choice model. Alternatives that pass the screen are evaluated in a manner consistent with random utility theory; alternatives that do not pass the screen have a zero probability of being chosen. The proposed model accommodates conjunctive, disjunctive, and compensatory screening rules. We estimate a model that reflects a discontinuous decision process by employing the Bayesian technique of data augmentation and using Markov chain Monte Carlo methods to integrate over the parameter space. The approach has minimal information requirements and can handle a large number of choice alternatives. The method is illustrated using a conjoint study of cameras. The results indicate that 92% of respondents screen alternatives on one or more attributes.

The Optimal Sequence of Questions for Aiding Consumer Decision

John Liechty

Typically empirical demand functions are used to aid managerial decisions (new product development, pricing, promotion, segmentation, …). With the advent of the Internet, it is has become possible to create and readily distribute decision aids that are targeted at helping individual consumers make decisions. Assuming that the goal of these systems is to help consumers make decisions that have a high or potentially even maximal utility, then the sequence of questions used to elicit and estimate an individual utility function should be as efficient as possible; this will result in questions which are different from the questions used to supply an empirical demand function for a managerial decision aid system. This presentation summarizes several ongoing research projects, which investigate methods for determining the ‘best’ sequence of questions, in the context of creating a consumer decision aid system. The first study (which is reported in the accompanying working paper) focuses on using demographic and product use questions in order to form Treed Regression models (CART like models) and related models, which are then used to determine the best sequence of demographic and product use questions. The second study ignores demographic and product use questions and focuses on the sequential design problem of identifying the best sequence of profiles for an individual to rate. This approach is based on a decision theoretic framework, where the loss function is the predictive error. In this setting we can formally talk about an optimal sequence of questions. Current efforts are underway to contrast this method with the recently proposed Fast Polyhedral method. In a related third study, we are extending the Fast Polyhedral approach by introducing a probabilistic framework for the Fast Polyhedral models and then generalizing this class of models so that the resulting class of new models nests both the Fast Polyhedral models and their Hierarchical Bayesian counterparts. This framework will not only offer a unified approach to estimation, but it opens the possibility of improving the Fast Polyhedral sequential design algorithm, by proposing profiles that are based on both the current constrained sub-space (the Polyhedral) and the underlying probabilistic framework introduced by this new class of models. For all of these studies we will discuss the theoretical framework and include some empirical results.

Component Analysis of Q-score profiles

Michael W. Browne and Longjuan Liang

Q-factor analysis originated with William Stephenson and involved the factor analysis of an interperson correlation matrix calculated over Q-scores, which are ratings with a forced symmetric distribution. Most applications have replaced the factor analysis with a principal component analysis. We avoid any conceptual involvement of interperson correlations by basing the analysis directly on the singular value decomposition of a matrix of raw Q-scores. Person Q-score profiles are then approximated by linear combinations of a small number of "Typical" profiles. Residuals from raw Q-scores instead of from interperson correlation coefficients are considered. Double-sided orthogonal or oblique rotations are carried out aiming at weights loading on a small number of profiles for each person and/or typical profiles with meaningful extreme scores.

Factor Prediction in Time Series Factor Analysis

Keiji Takai and Yutak Kano

The factor analysis model has an intrinsic problem, namely factor (score) indeterminacy, in which observed variables can not perfectly predict latent common factors. For the traditional factor analysis model, the problem has been reasonably resolved by considering the case where the number of observed variables tends to be infinity. The problem, however, has not been yet attacked in time series factor analysis models being recently developed. In this talk we address the problem of factor indeterminacy in a time series factor analysis model with a single common factor, and extend the traditional result to time series data. It is first shown that the mean squared error of the best linear predictor based on the present and past observed vectors can be expressed as the function of the following two sequences. One is the sequence with respect to the dimension p of the observed vector and the other is the one with respect to the number T of the past observed vectors. Using this result, we prove that the factor indeterminacy can be eliminated under the same condition as in the traditional factor analysis. It is also shown that the factor score can not be predicted perfectly even when using infinitely many past observations except for trivial cases, provided that p is held fixed.

The link between Minres and Minimum Rank Factor Analysis

Gregor Sočan and Jos M.F. ten Berge

Minres Factor Analysis and Minimum Rank Factor Analysis (MRFA) have been considered two quite different estimation approaches in Common Factors Analysis: Minres minimizes the sum of squared residuals while MRFA minimizes the sum of the smallest $p - r$ eigenvalues of the $p \times p$ reduced covariance matrix, both after extracting r common factors. A recent estimation method, called Quadratic Minimum Rank Factor Analysis (MRFA-Q), shall be discussed, which presents the link between MRFA and Minres. Its objective function is the same as the one of Minres (that is, sum of the squared smallest $p - r$ eigenvalues of the reduced covariance matrix), except for the constraint that the reduced covariance matrix must remain positive semidefinite. On the other hand, because both MRFA and MRFA-Q produce a positive semidefinite reduced covariance matrix, both methods make possible to estimate the amount of common variance left unexplained after a certain number of factors have been extracted. In the second part of the talk, we shall discuss the results of a comparative simulation study which included various levels of sample size, average communality and number of variables per factor. Finally, some guidelines on the choice between the three methods shall be given.

Empirical Verification Of Assignment Of Items To Subtests: Multiple Group Method Or Common Factor Confirmatory Factor Analysis

Ilse Stuive, Henk A.L. Kiers, Marieke E. Timmerman, Jos M.F. ten Berge

Confirmatory Factor Analysis is often used for empirical verification of a partitioning of items into subtests. The most often used method is Common Factor Confirmatory Factor Analysis, with loadings defined such that they correspond to the partitioning of items to subtests. Another method is the Multiple Group Method. In this method subtests are simply defined as unweighted sums of the scores on all items assigned to the subtest. If each item correlates highest with the subtest to which it is assigned, then this indicates that the assignment is supported by the data.
A simulation study is performed to investigate the degree to which both methods are capable of indicating whether or not the suggested assignment of items to subtests is supported by the data. The influence of several variables, such as level of communality and sample size, on the quality of the judgments will be discussed.

Some Results on Mean Square Error for Factor ScorePredition

Krijnen, W.

For the confirmatory factor model a series of inequalities is given with respect to the mean square error (MSE) of three main factor score predictors (Ten Berge, Krijnen, Wansbeek, & Shapiro, 1999). The eigenvalues of these MSE matrices are a monotonic function of the eigenvalues of the matrix $\Gamma_p = \Phi^{\frac{1}{2}}\Lambda_p'\Psi_p^{-1}\Lambda_p\Phi^{\frac{1}{2}}$, where Λ_p is the loadings matrix, Φ the factor correlations matrix, and Ψ_p the error variance matrix. The matrix Γ_p increases with the number of observable variables *p*. A necessary and sufficient condition for mean square convergence of predictors is divergence of the smallest eigenvalue of Γ_p , or equivalently, divergence of signal to-noise (Schneeweiss & Mathes, 1995). The same condition is necessary and sufficient for convergence to zero of the positive definite MSE differences of factor predictors, convergence to zero of the distance between factor predictors, and convergence to the unit value of the relative efficiencies of predictors. Various illustrations and examples of the convergence results are given as well as explicit recommendations on the problem of choosing between the three main factor score predictors.

References:
Schneeweiss, H. & Mathes, H. (1995). Factor analysis and principal components. Journal of Multivariate Analysis, 55, 105-124.
Ten Berge, J.M.F., Krijnen, W.P., Wansbeek, T.J., & Shapiro, A. (1999). Some new results on correlation preserving factor scores prediction methods. Linear Algebra and its Applications, 311-318.

Measurement And Prediction In Culturally Heterogeneous Populations

Annette M. Maij-de Meij, H. Kelderman, and H. van der Flier

Psychological measurement and prediction procedures usually assume that individual differences can be described by one standard measurement and prediction model. However, empirical psychometric research has shown that these models are not always invariant over (cultural) groups. Therefore, it is important to study the qualitative differences in measurement and prediction and to design culture-fair procedures that take these differences into account. For example, in selection psychology, different candidates from different cultural groups have to be compared with respect to their suitability for a certain job. In practice, information about group-membership is often used in the decision. This practice has been criticized, because it may result in different selection probabilities for equally fit individuals from different groups.
A new measurement and prediction method is proposed, that takes qualitative group differences in test performance into account, but does not rely on information about group membership. Selection decisions are solely based on item response patterns. Qualitative group differences are only taken into account to the extent that they may be inferred from the response patterns. The presentation will discuss the proposed mixture measurement model, which describes both within and between group heterogeneity. Furthermore, the model will be applied to the results of a sample of subjects for whom criterion data are available, to study the differences in accuracy of the prediction.

Selecting The Most Robust Unidimensional IRT Model When The Data Are Multidimensional: A New Approach To Model Fit

Shu-chuan Kao, Tony D. Thompson, and Won-Chan Lee

Given that most IRT model-fit indices focus on overall fit, this paper proposes an alternative method for inspecting data-model-fit and provides ways to show how the UIRT model fits data at different true score levels which can be shown graphically to form profiles.
Research data were simulated from 10-dimensional NOHARM calibrations to approximate the effect of content variety, speededness, and testlet effect (item dependency) in the real test data. To decide the most robust UIRT model for the multidimensional data, two kinds of simulation studies were conducted to evaluate the conditional fit for Rasch, 2PL, and 3PL models. The first study focused on how the UIRT model-based expected true scores perform conditioned on raw or scale scores from 5000 replications. The second study compared the model-based conditional standard error against the theoretical conditional standard error (Lee, Brennan, Kolen, 2000) over 1000 replications.
The results of the study indicate that 2PL and 3PL models have slightly better estimation than Rasch model, but no major differences were found among the three models. Additionally, the Rasch model tended to over predict how well it fits the data, especially for outperforming examinees.

Limited Information Goodness-Of-Fit Testing In Multidimensional Contingency Tables

Albert Maydeu-Olivares and Harry Joe

We introduce a family of goodness-of-fit statistics for testing composite null hypotheses in multidimensional contingency tables of arbitrary dimensions. These statistics are quadratic forms in marginal residuals up to order r. They are asymptotically chi-square under the null hypothesis when parameters are estimated using any consistent and asymptotically normal estimator. We perform a small simulation study to illustrate how when r is small (r = 2) the proposed statistics have accurate empirical Type I errors when a widely used item response model is estimated using only 300 observations in contingency tables of over 10^7 cells. Furthermore, the proposed statistics are more powerful than Pearson's X^2.
Also, we show that the proposed statistics (but not X^2 even for the maximum likelihood estimate) are asymptotically chi-squared under the null hypothesis when applied to subtables. Thus, the proposed statistics can be used for a piecewise goodness-of-fit assessment to determine the source of misfit in poorly fitting models.

Bayesian Modification Indices For Detecting IRT Model Misfit

Jean-Paul Fox

Bayesian modification indices are presented that provide information for the process of model evaluation and model modification. These indices can be used to investigate the improvement in a model if fixed parameters are re-specified as free parameters. The indices can be seen as a Bayesian analogue to the modification indices commonly used in a frequentist framework. The aim is providing diagnostic information for multiparameter models where the number of possible model violations and the related number of alternative models is too large to render estimation of each alternative practical. As an example, the method is applied to an item response theory (IRT) model, that is, to the two-parameter model. The method is used to investigate nonignorable missing data, differential item functioning, and violations of the assumption of local independence.

Characterization Of The Distribution Of The *Lz* Index Of Person Fit According To The Estimated Proficiency Level

Gilles Raîche and Jean-Guy Blais

The distribution of person fit indices is not easy to determine in tests where the item sample is too small to conform to a theoretical asymptotic statistical distribution, particularly the normal N(0,1). In practice, it is very often the case and, consequently, it is difficult to get the critical percentile value indicating person misfit. We investigated if we can obtain at least a $\mathrm{N}(\bar{\boldsymbol{L}}\boldsymbol{z} \mid \hat{\theta}, \boldsymbol{S}^2_{Lz} \mid \hat{\theta})$ distribution of Lz conditional on the estimated proficiency level. Then, we verified if it was possible to linearly predict the critical first and fifth percentiles of this same distribution by the estimated proficiency level and its' standard error. Results of 4000 simulations of a 85 fixed items test, each of 23 simulated proficiency levels, showed that the first and fifth percentiles of the Lz distribution generally can't be predicted by a $\mathrm{N}(\bar{\boldsymbol{L}}\boldsymbol{z} \mid \hat{\theta}, \boldsymbol{S}^2_{Lz} \mid \hat{\theta})$ distribution according to the related 23 estimated proficiency levels. The proficiency level was estimated with MLE, MAP, WLE, EAP and AEAP methods. More promising, the first and fifth percentiles of the Lz distribution were well predicted by a linear regression on the estimated proficiency level and its' standard error ($\boldsymbol{r} > 0.90$), particularly with the MLE, WLE and AEAP methods.

Invited speakers, 14.30h-15.10h

Phase Transitions In Psychological Processes

Han van der Maas

In non-linear dynamical systems, bifurcations demarcate parameter spaces in which behavior is relatively simple to describe. At a bifurcation point the number or type of equilibria change. Bifurcation analysis (numerical or analytical) requires a mathematical description of the system. Such a description usually is missing in the social sciences, yet the importance of dynamical system theory for this field is evident. Psychological processes are very complex, and best understood as nonlinear dynamical systems.
Psychological behavior can be described with basic quantitative relations within specific parameter ranges and changes qualitatively, often abruptly, between such ranges. Although we usually do not have a mathematical description of the system, we can search for qualitative changes in psychological behavior in order to give a first, general description of the system. This can be done by detecting phase transitions in behavior. We developed a methodology for testing phase transitions using experimental designs and statistical techniques. The latter include methods based on categorical latent variable models, finite mixtures, functional data analysis, stochastic catastrophe models, but also new non-parametric time series techniques. We have applied this methodology to various domains of cognitive development (stage transitions), multi-stable perception, attitudes, emotions, and speed-accuracy trade-off.

Time Series Factor Analysis

Erik Meijer and Paul D. Gilbert

Standard factor analysis is based on the assumption that the observations are identically and independently distributed. For time series data, dynamic factor analysis and more general state-space models are available. Unlike standard factor analysis, these are not ”pure” measurement models, but include a model for the time dependency of the observations as well. In some situations, however, it is desirable to study a measurement model without having to specify a substantive model for the dynamic process governing the factors. In such situations it seems attractive to apply standard factor analysis techniques to a time series of indicators. In this paper we study whether and how the estimators and statistical theory of standard factor analysis have to be adapted to accommodate this context and to what extent this gives statistically sound results.
We show that with a combination of a few standard results from factor analysis, time series analysis, and the generalized method of moments, asymptotically correct statistical inference is possible under very weak assumptions about the time dependency of the observations.
Our study is motivated by the macro-economic problem of measuring transactions money. The characteristics of this application imply some nonstandard characteristics of the factor analysis model, give rise to studying nonstandard methods for rotation and normalization of the factors, and require some small adaptations of standard factor score prediction methods.
A striking empirical result is that the sampling variability in the parameters accounts for most of the errors in the factor score predictions, and the fundamental inability to estimate the factors consistently has only a very small impact for this type of data.

Symposium and sessions, 15.10h-16.10h

Stability of Nonlinear Multiple Regression (CATREG) using the Lasso

Anita J. van der Kooij, and Jacqueline J. Meulman

The CATREG program performs nonlinear multiple regression using optimal scaling to (monotonically or nonmonotonically) transform categorical variables, incorporating (spline) nominal, (spline) ordinal and numerical scaling levels. The CATREG algorithm minimizes a least squares loss function (OLS) using alternating least squares to find optimal quantifications for categorical variables, while simultaneously optimizing the squared multiple regression coefficient. The quantifications are optimal for the particular set of predictor variables that is used in the analysis. Thus, CATREG will obtain different quantifications for the same predictor variables if a subset of the predictors is used.
The lasso was developed by Tibshirani to combine two standard techniques for improving OLS estimates: subset selection for improving interpretability and ridge regression for improving prediction accuracy. Subset selection can result in reduced prediction accuracy, while with ridge regression (shrinking the coefficients of all predictors) interpretation may not be easy. In contrast to ridge regression, the lasso shrinks some coefficients to zero, so it improves both interpretability and prediction accuracy.
In this paper we incorporate the Lasso in the CATREG method and study the effect on the stability of the optimal quantifications and the predictive accuracy for different (non)monotonic spline transformations.

Cronbach's Alpha Based On Pooled Within-Class Correlations As An Index For Detecting Single-Peakedness

Marike Polak, Mark de Rooij, and Willem J. Heiser

In psychometrics two types of attitude items can be distinguished, monotone and single-peaked items. The two types of items typically require different types of analyses. Hence, researchers should be able to examine what type of data they have, so that they can choose the appropriate analysis. In the case of optimal scaling techniques, monotone items should be analyzed by multiple correspondence analysis (homogeneity analysis), while single-peaked items should be analyzed by simple correspondence analysis.
In this paper we will explore a new statistic (T). T is defined as $T = \alpha_1 - \alpha_2$, where α_1 is the usual Cronbach's alpha, and α_2 is a similar measure based on within-class correlations. For monotone items $T > 0$, while for single-peaked items $T < 0$. The classes on which α_2 is based are homogeneous groups on a latent trait. Methodology will be presented for deriving the classes.
This method will be compared to the criterion of irrelevance of Thurstone and Chave (1929), which is a graphical method for detecting single-peakedness.

Consistency Of Strategy Use On A Set Of Arithmetic Items: An Application Of Multiple Correspondence Analysis Of Repeated Measurements

Cornelis M. van Putten, Gabriëlle Rademakers, Meindert Beishuizen, and Jan Janssen

In this study we categorized written strategies in solutions of ten division problems in a Dutch national mathematics assessment at the end of primary school. Three types of strategy were used most often: Traditional long division algorithms, so-called realistic strategies, and mental solutions (only an answer, no working). Most students were rather consistent in their strategy use over the set of problems: 46% could be characterized as using a traditional strategy either exclusively (T) or combined with mental solutions (TM); 29% as using a realistic strategy exclusively (R) or combined with mental solutions (RM); 19% as exclusive mental problem solvers (M).
Multiple correspondence analysis with the HOMALS program of the strategy item reflected these subgroups of consistent strategy users in a peculiar, triangular plot of student quantifications. However, this analysis did not indicate any special relations between items and strategies as could be expected for students using two types of strategies on the ten items, e.g. using a traditional written strategy for some problems and a mental strategy for other problems. The plot seems to be dominated by the students repeatedly using only one strategy. A diagnostic will be proposed to partition the objects into the five groups along two edges of the triangle.

Estimating Response Bias And Testing For Random Error In Randomized Response.

Maarten Cruyff, Ardo van den Hout and Peter van der Heijden

In randomized-response surveys the respondents have to answer sensitive questions according to a randomization scheme. It is assumed that the randomization scheme eliminates response biases that are due to the sensitive nature of the questions. However, recent randomized-response surveys on violations of social welfare rules and regulations in the Netherlands indicate that not all respondents follow these randomization instructions. The results show that some (affirmative) response patterns have lower prevalence than expected on the basis of the randomization scheme.
A model is presented that incorporates response bias. In this model, it is assumed that some of the respondents follow a self-protective answer strategy (SP) and deny every violation, regardless the outcome of the randomization scheme. Since the model is not identified, the violation probabilities are estimated by maximizing the profile log-likelihood with a fixed of SP-parameter. Given that with a fixed SP-parameter the model is saturated, we expect the best model to show a perfect fit. If it does not fit perfectly, either our assumption with regard to the SP-respondents is inadequate, or some response patterns are underrepresented due to random error in the randomization scheme. In the former case, we should reject the model, in the latter we can accept it. A test is presented to decide between these two options.

Bayesian Inference For Non-Compliance In The Randomized Response Model

Ardo van den Hout

The randomized response model is a misclassification design which is used to protect the privacy of respondents w.r.t sensitive questions. The misclassification is induced by the respondents themselves and it is to be expected that some of them do not comply with the design. These respondents cause perturbation which is not accounted for in the standard randomized response model. An extension of the standard model is presented that takes non-compliance into account. The extended model is a nested latent class model: At the first level, class A consists of respondents that comply and class B consists of respondents that do not comply. At the second level, class A consists of the classes that are formed by the latent status regarding the sensitive item. The extended model is not identified but can be specified using assumptions about non-compliance.
Bayesian inference is used to assess the extended model. The posterior is evaluated using a Metropolis algorithm. Goodness-of-fit tests and posterior predictive p-values are used to investigate model fit.
The research is illustrated by Dutch randomized response data concerning violations of regulations for social benefit.

Estimating Different Forms Of Cheating In The Randomized Response Design Using The Latent Class And The Rasch Model

Ulf Bockenholt and Peter G.M. van der Heijden

Self-reports about compliance with rules and regulations that govern public life are prone to response biases. Out of the currently available set of methods that are available to obtain more honest answers to personal and/or sensitive questions, the method of randomized responses has been shown to be most effective. So far, in applications of randomized-response (RR) methods, it has been assumed that they eliminate response biases. We investigate the validity of this assumption using RR data obtained in two recent surveys about compliance with regulations under the Dutch Disability Benefit Act.
We propose a class of multivariate RR models with two components. One component deals with different ways in which the randomised response design can be violated by the respondents. The other component deals with the way the randomised response items are related, among others by a latent class model, a semi-parametric and a parametric Rasch model.
We show that RR methods do not eliminate response biases but that these biases can be modelled in effective and informative ways. For the presented examples, we find that accounting for the response biases leads to compliance estimates that are at least twice the size of the estimates obtained when response biases are ignored.

Effect Of Computerized Adaptive Testing On Test-Taking Motivation

Andreas Frey, Johannes Hartig, and Helfried Moosbrugger

Since early work on Computerized Adaptive Testing (CAT) by Betz and Weiss (1976a, b) a positive effect of CAT on participants' test-taking motivation is widely accepted. However, in recent publications this assumption is criticized frequently (Bergstrom, Lunz & Gershon, 1992; Eggen, 2003; Eggen & Verschoor, 2004). To bring light into the controversy the effect of adaptive testing with the Frankfurt Adaptive Concentration Test (FACT, Moosbrugger and Heyden, 1997) on test-taking motivation (self-report) was examined experimentally (N=88). Within the FACT item difficulties are tailored to participants' proficiency levels by varying item presentation time. Two independent samples received the FACT either in an adaptive or a non-adaptive (non-CAT) version. Test-taking motivation was significantly lower in the CAT- condition than in the non-CAT-condition, $F(1, 86) = 5.35$, p=.023, η^2=.059. Further analyses showed that the differences in test-taking motivation between experimental conditions (CAT vs. non-CAT) are mainly due to the perceived probability of success and not to other motivational components like perceived probability of failure or perceived challenge. It is discussed whether the findings can be generalized to other adaptive tests and how adaptive tests can be designed to avoid negative effects on participants' test-taking motivation.

Computerized Adaptive Testing For Polytomous Motivation Items: Administration Mode Effects And A Comparison With Short Forms

A. Michiel Hol, Harrie C. M. Vorst, and Gideon J. Mellenbergh

Administration mode effects and efficiency of a computerized adaptive test were studied. The item pool consisted of 24 polytomous motivation items and was selected using Mokken and Molenaar's nonparametric counterpart of Samejima's graded response model (GRM). However, in the analyses the parametric GRM was used. Calibration data showed that the parametric GRM did not fit the data optimally; therefore a simulation study was done to assess possible consequences. Administration mode effects were tested in an experiment (n=515) comparing a computerized test (CT) and a CAT. CAT efficiency was studied by comparing CATs of different lengths with two different short forms. The first short form consisted of items most frequently administered in the CAT version; the second short form consisted of items that had the largest discrimination (a-) parameters. CATs and short forms of the same length were compared; lengths varied between 1 through 23 items. Results showed no administration mode effects for CATs consisting of at least 4 items. The efficiency analyses showed that CAT outperformed the two short forms in almost all aspects. Comparison of the simulation study and the real life CAT study showed that results were quite similar, thereby throwing a new light on model fit issues.

Scores And Response Times In The Short Version Of The NEO Test Battery

Tom Backer Johnsen

At the previous IMPS meeting I presented results from a several web-based data collections indicating a clear relation between the response to a item and the position of the item within the test. In a long series of similar items with rating scales in the same direction, the time needed to produce a response is systematically reduced towards the end, while the ratings became more positive. In a new study (not yet complete) I use the short version (Norwegian translation) of the NEO with the same type of data collection procedure. This test contains a number of items (26 of 60) which have to be reversed in order to form the standard five scales. Compared to the first set of studies, there is a systematic reduction in the item response times towards the end of the sequence, but not a very clear rise in the mean ratings. The latter result is perhaps influenced by the variation in polarity of the items. In addition, preliminary analysis of the data indicate that (a) responders require more time to respond to negative than positive items, while negative items have a systematic lower mean rating (after reversal) than the positive ones. There are also interesting results for different patterns in the sequence of items (e.g. positive items following negative, vs. positive items following positive etc.). If response or latency reflects processing time, this may mean that it is not a trivial matter to include both positive and negatively formulated items in the same test, since the two types may imply at least somewhat different processes.

An Investigation of Examinee Pacing on a Large-scale, High-stakes Examination

Kimberly A. Swygert, Eric S. Muller, Dave Swanson, Gerard Dillon, Brian Clauser, and Melissa Margolis

With computer administration of large-scale, high-stakes tests, it has become possible to capture specific examinee performance information such as the amount of time spent on an item. Much past research on examinee pacing has focused on a single timing condition. The present study extends research on examinee pacing by examining performance on a large-scale, high-stakes operational examination under two timing conditions, a pre-timing change condition and a post-timing change condition in which the time allotment per item was increased. Descriptive analyses reveal performance differences between pre- and post- timing change groups. Performance indicators included: 1) the relationship between total test score and the amount of time examinees used on the test; 2) the relationship between item position and probability of a correct response, and 3) the relationship between total test score and the discrepancy between an examinee’s performance on the initial and final items in a section. HLM analyses will be conducted to examine the relationship between timing and item characteristics such as item position, word count, inclusion of graphical or pictorial content, and item type within a three level model that includes a timing condition factor. Results highlight the complexity in interpreting data collected under a single timing condition.

Object Oriented Approach For Managing, Visualizing, Linking, And Analyzing Large-Scale Dynamic Databases In Behavioral, Educational, And Psychological Sciences

Chris Chiu, Ivy Li, Marilyn Seastroms

Accurately depicting massive amounts of data in psychometrics can be a real challenge because the displaying media (e.g., computer screens or paper) are frequently limited in size and resolution. A common approach to overcome such a challenge is to utilize rendering techniques such as zooming, panning, lighting, and layering. In order to utilize these techniques, it is necessary to structure and systematically index the data. However, human behaviors and cognitive processes are dynamic and complex to structure and index. In addition, unlike physical objects, human behaviors and cognitive processes do not always have a tangible shape for display. The proposed study introduces a framework and a set of computational tools that are suitable for modeling and displaying dynamic data collected in the behavioral, educational, and psychological sciences. The tools include linked lists, multidimensional arrays, mesh grids, multi-way data structure, and systems of structured matrices. A successful example in applying these tools to model binary data is discussed in the context of large-scale surveys and psychometric tests.

Explanatory IRT-models and Figural Matrices Items

Alexander Freund & Heinz Holling

Considered "classical" representatives of Spearman's *g*, figural matrices tasks can be generated using rational item construction principles. This approach allows for the systematic analysis of explanatory IRT-models such as the LLTM, which takes into account item properties to explain item difficulty. However, not only item properties, but also different person properties can be incorporated into the model, which then can be called a latent regression LLTM (De Boeck & Wilson, 2004). In this talk, the method and results of a study carried out to investigate difficulty-inducing factors of matrices items are presented. LLTM and latent regression LLTM analyses show the contributions of different item properties and moderating person effects.

References:
P. De Boeck & M. Wilson (Eds.) (2004). *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach*. New York: Springer.

The Finite Sample Behavior of Satorra-Bentler and Bartlett-Corrected Fit Statistics in Large Models

Walter Herzog, Anne Boomsma, and Sven Reinecke

Designed to be robust against nonnormal distributed variables, Satorra and Bentler (1988, 1994) developed a *scaled* (mean-corrected) chi-square statistic, $\hat{\chi}^2_{\mathbf{SC}}$, and an *adjusted* (mean- and variance-corrected) chi-square statistic, $\hat{\chi}^2_{\mathbf{AD}}$, for evaluating overall fit of structural equation models. Fouladi (2000) has shown that the distribution of $\hat{\chi}^2_{\mathbf{AD}}$ converges fast to a chi-square distribution and can therefore be used with small sample sizes as well.
Satorra and Bentler (2001) claimed that both corrections are also reliable for large models, i.e., models with many degrees of freedom. The aim of the present investigation is to test this claim in a comparative simulation study of fit statistics for confirmatory factor models, both under variation of model and sample size. Multivariate normal variables are generated to isolate the model size effect. Findings indicate that $\hat{\chi}^2_{\mathbf{AD}}$ produces adequate Type-I error rates even for models with about 1000 degrees of freedom. In contrast, both the traditional likelihood-ratio statistic, $\hat{\chi}^2_{\mathbf{ML}}$, and $\hat{\chi}^2_{\mathbf{SC}}$ overestimate nominal Type-I error rates up to 80.7%. The study also includes an assessment of Bartlett-corrected statistics (cf., Nevitt & Hancock, 2004).
The utility of the findings for a contemporary evaluation of structural equation models is discussed. The results suggest to use $\hat{\chi}^2_{\mathbf{AD}}$ *in general* – even under conditions of normality.

Approximations To The Distributions Of Fit Indexes Under Fixed Alternatives In Nonnormal Samples

Haruhiko Ogasawara

Approximations to the distributions of fit indexes for misspecified structural models under fixed alternative hypotheses are obtained in nonnormal samples as well as normal ones. The fit indexes include the normal theory likelihood ratio chi-square statistic for a posited model, the corresponding statistic for the baseline model of uncorrelated observed variables and various fit indexes as functions of these two statistics. The approximations are given by the Edgeworth expansions for the distributions of the fit indexes under arbitrary distributions. Numerical examples in nonnormal samples with the asymptotic and simulated distributions of the fit indexes show the inappropriateness of the normal theory approximation using noncentral chi-square distributions.

Corrected Version of AIC for Selecting Covariance Structures in a General Nonnormal Case

Hirokazu Yanagihara

This paper deals with a bias correction of Akaike's Information Criterion (AIC) for a selection of covariance structure models under nonnormality when we assume that an observation is distributed according to the normal distribution. The AIC under the normal assumption has a constant bias that depends on a kurtosis of the true distribution. We correct this bias by partly using the cross-validation method. We verify that our criterion is better than the AIC and TIC by calculating a theoretical bias and conducting a numerical experiment.

Symposia and sessions, 16.40h-18.00h

Analysis of a Set of Brand Switching Matrices by Using Asymmetric Multidimensional Scaling

Akinori Okada, Tadashi Imaizumi, and Hiroyuki Aoshima

Brand switching data among canned coffee brands are analyzed by asymmetric multidimensional scaling. A brand switching matrix is calculated for each consumer from her/his purchase data. The brand switching matrix has a hypothetical brand to deal with the difference between inflow and outflow of the two consecutive months for each consumer. The brand switching matrices over all consumers are added to derive a mean brand switching matrix. A set of mean brand switching matrices, each comes from different time period, is derived. Each matrix constitutes one-mode two-way asymmetric proximities at a period, and is analyzed by one-mode two-way asymmetric multidimensional scaling to derive a multidimensional configuration of the brands, where each brand is represented as a point and a circle (sphere, hypershpere). The hypothetical brand in the previous period can be changed to the other brands in the present period, and the hypothetical brand in the present period can come from the other brands in the previous period. By comparing the location and the radius of the circle (sphere, hypershpere) representing the hypothetical brand for several periods, it is possible to differentiate the brands to be purchased more and those to be purchased less in the next period.

Feature selection in Feature Network Models for proximity data

Laurence E. Frank, and Willem J. Heiser

A typical psychological approach to learn more about how human cognition processes stimuli consists of analyzing the ratings of perceived similarity of these stimuli. In certain experimental settings, it is useful to characterize the objects of the experimental conditions as sets of binary variables, or features.

If N respondents evaluate the dissimilarity of m stimuli and F binary features characterize the stimuli, the number of features in which two stimuli are distinct yields a dissimilarity coefficient that can be used as a structural model to be fitted to the data. Feature Network Models 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. The additivity properties of networks make it possible to consider the model as an univariate multiple linear regression problem with positivity restrictions on the parameters.

In some research settings, the features that characterize the objects are not known a priori. The total number of binary features for m objects is equal to 2^m, which can be efficiently enumerated using Gray codes. If m is not larger than 18 it is feasible to find an adequate subset of features by enumerating the total number of features first and then perform a subset selection using the Lasso algorithm (Efron, Hastie, Johnstone & Tibshirani, 2004) that was modified to meet the positivity constraints of our model. If m exceeds 18, filter techniques have to be applied to reduce the number of features before using subset selection.

Reference:

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, *32*, 407-499.

Incomplete Unfolding

Frank M.T.A. Busing, Maaike J. Velderman, and Mark de Rooij

Collecting paired comparison data for multidimensional scaling (MDS) can be a burden,
especially for a large number of objects (n) to be compared. The number of comparisons $n(n-1)/2$ increases n with each additional object. Instead of collecting and using all paired comparisons, a subset of the data is sufficient to obtain a reliable MDS representation. Spence and Domoney (1974) investigated several deletion patterns for the comparisons and calculated a minimum fraction, which could safely be deleted. Connectedness of the objects was essential, without which separate MDS problems would emerge.

Marketeers, exploring the preferences of consumers with, for example, multidimensional unfolding (MDU), are sometimes confronted with the problem that subjects can only judge or taste so many objects (imagine soups, ice-creams, whiskeys, or other digestives). Instead of collecting and using all preferences of all subjects, a subset might be sufficient to obtain a reliable MDU representation. The results from earlier incomplete MDS research, however, do not apply to incomplete MDU: the deletion patterns can not be transferred from square symmetrical to rectangular matrices and connectedness is more difficult to determine in the incomplete MDU case (Velderman, 2005).

In this study, four MDU deletion patterns were compared and minimum fractions were determined for each of the patterns.

References:

Spence, I. and Domoney, D.W. (1974). Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika, 39*, 469-490.

Velderman, M.J. (2005). Unfolding with incomplete data. Unpublished master’s thesis, Leiden University.

Locations And Scalings In Distance Association Models For The Analysis Of Change

Mark de Rooij

Distance association models as proposed by De Rooij and Heiser (2005) are reparametrizations of *RC(M)*-association models. Recently two papers (Gabriel, 2002; Gower , 2004) studied scalings in biplot models, of which the *RC(M)*-association model is a special case, and concluded that these have little impact on the interpretation. In this paper again scalings, but also locations and non-singular transformations, are studied for these models but now from a distance perspective. It is shown that different locations, scalings, and non-singular transformations have a major impact on the interpretation. So, contrary to the earlier findings, we conclude that locations, scalings, and non-singular transformations are in fact very important for interpretation. If one is willing to go along with the assumption of the author that diagrams are -in practice- often interpreted by a distance rule, the findings in this paper influence all biplot models.

References:

De Rooij, M. and Heiser, W.J. (2005). Graphical representations and odds ratios in a distance association model for the analysis of cross-classified data. *Psychometrika*, *70*, 1-24

Gabriel, K.R. (2002). Goodness of fit of biplots and correspondence analysis. *Biometrika, 89*, 423-436.

Gower, J.C. (2004). The geometry of biplot scaling. *Biometrika, 91*, 705-714.

Probabilistic Networks For The Course Of Emotions

Frank Rijmen

The experience sampling methodology (ESM) is a relatively new data collection technique that results in multiple assessments of the same person. This way, it becomes possible to model within-person processes over time. The data I will discuss stem from an ESM study in which participants were repeatedly questioned about the occurrence of positive and negative feelings in their daily lives. The starting point is a latent class model that allows for transitions between classes over time. This basic model is closely related to the latent transition (Collins & Wugalter, 1992) and latent markov model (Langeheine & van de Pol, 1994). However, the latter models currently allow for a few timepoints only. It is shown how probabilistic network theory can be used to estimate models that involve much more measurement occasions. The success of probabilistic networks stems from the fact that probabilistic networks integrate aspects from both probability and graph theory. Further advantages of linking up with probabilistic network theory include that missing data are easily dealt with if the missingness mechanism is ignorable, and that the basic model is easily extended. Some extensions will be discussed.

Scale Invariant Stochastic Catastrophe Theory.

Eric-Jan Wagenmakers, Han L. J. van der Maas, Pascal A. I. Hartelman, and Peter C. M. Molenaar

Catastrophe theory provides a general mathematical description of how continuous changes in control variables lead to sudden jumps or discontinuities in observed behavioral variables. Fitting a catastrophe model to a noisy system constitutes a serious problem, however, because catastrophe theory was formulated specifically for deterministic systems. Cobb attempted to solve this problem by developing a stochastic counterpart of catastrophe theory (SCT) based on Itô stochastic differential equations. In SCT, the stable and unstable equilibrium states of the system correspond to the modes and the antimodes of the empirical probability density function, respectively. Unfortunately, SCT is not invariant to smooth and reversible transformations of the dependent variable this is an important limitation, since invariance to transformations is a defining feature of deterministic catastrophe theory. We derive a generalized version of SCT that does remain invariant under transformation and can include Cobb¹s SCT as a special case. The generalized SCT can be instantiated via a simple time series techniques. We illustrate the invariance problem and its solution with practical applications.

Testing For Hysteresis

Han van der Maas and Brenda Jansen

Hysteresis occurs when the timing of transitions depend on the history of the system. Hysteresis is a strong indication of a real phase transition as opposed to a continuous acceleration. However, testing hysteresis is difficult. We will present designs and statistical methods to test for hysteresis. We use examples from the domains of conservation and proportional reasoning.

Switching Processes In The Speed-Accuracy Trade-Off

Ingmar Visser and Han van der Maas

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). It is more general than those models because it is also suitable long (single subject) time series. We discuss an application about investigating hysteris in switching between fast-guessing and accurate responding in a reaction time task. Some extensions of the model are discussed such as the inclusion of (time-dependent) covariate. These can be used to partially overcome the restrictive Markov assumption.

Creating and Evaluating a Third Grade Science Benchmark Test using a Cognitive Diagnosis Model

Terry Ackerman, Deb Bartz, Bob Henson, and Jon Templin

The evaluation team for the North Carolina Partnership for Improving Math and Science was charged with assessing the science impact of its five year K-12 program on students. Currently there are no end-of-grade tests in North Carolina for the elementary grades. Thus, the evaluation team created a third grade benchmark test that was aligned with the state standards. Six underlying skills that form the foundation of the K-3 standards were identified and a pool of items that measured various combinations of the skills was created. Items were piloted using two parallel forms. After analyzing the two forms, a single final form was created using a target Q-matrix and pretest item characteristics. This matrix identifies which combination of underlying skills each item is measuring. Response data from the final administration were analyzed using the reparameterized unified model (Hartz, 2002; DiBello & Stout, 2003) using the computer software Arpeggio. For each student a profile of skills that were mastered or not mastered was created. These data were also aggregated at the classroom level. This presentation will detail the test development and administration process, modeling of the data, and how results were reported to teachers and administrators.

Item Response Theory In Medical Measurement: The AMC Linear Disability Score Project Item Bank

Rebecca Holman and Cees A.W. Glas

The AMC Linear Disability Score (ALDS) project item bank aims to quantify the ability of respondents with a range of chronic conditions to perform activities of daily life. Examples are washing and dressing oneself, mobility, food preparation and housework. The ability to perform activities of daily life is seen as a good measure of the impact of disease and treatment on the patient as a whole. Using item response theory based techniques to measure the ability to perform activities of daily life presents a number of challenges. This paper concentrates on the challenges encountered when calibrating the ALDS item bank. These include adapting methodology to medical research ethics, dealing with responses in a 'don't know, not applicable' category, differential item functioning between age and gender based groups and the possibility that the main 'ability' represents more than one dimension. In addition, a plan for and the results of the calibration analysis for the item bank in two populations will be presented.

Employing Item Response Theory to Validate Probability Judgements

Niels Smits, and Gideon J. Mellenbergh

This paper presents Item Response Theory (IRT) as a framework to study probability judgments. Dutch high school students estimated the probabilities that they produced correct answers to the items of an English writing skills test. On the basis of item- and person parameters of an IRT model (the One Parameter Logistic Model [OPLM]), for each student-item combination a probability of producing a correct answer was derived. This normative subjective probability was compared with the corresponding probability judgment of the students. The results showed overconfidence, and a modest correlation between probability judgements and IRT probabilities. In addition, students receiving a very rough indication of item difficulty showed less discrepancy between both types of probabilities than students without such information. More proficient students gave higher probability judgements, and for easier items higher probability judgments were given. For the most difficult items, students were somewhat underconfident. The paper discusses the implications of the outcomes for both confidence testing and probability judgement research.

Definition and Estimation of Skills Classification Reliability

Lou DiBello, Bob Henson, Jon Templin, and Russell Almond

An essential aspect of developing and implementing large scale tests is the ability to conceptualize reliability and make reasonable estimates from data from a single test form. The idea of reliability is just as important in profile scoring applications in which the reported scores are profiles of discrete classification inferences. This paper reviews classical definitions and standard estimates of test score reliability and relates them to methods for defining and estimating reliability of skills classifications. Connections are drawn to Shannon information, Kullback-Liebler information divergence, and weight of evidence. Practical estimation methods are described and are applied to specific testing programs. Estimates from both real and simulated data are examined and interpreted in terms of test design and test quality.

Latent Class Analysis With Several Latent Variables And Its Application To POS Data

Naoko Sakurai and Kazunori Yamaguchi

New trial of drawing market segmentation by latent class model with consumer’s purchasing behavior, who has been mature within computer network, is proposed. Mixture latent class model with two kinds of latent variables is introduced for the analysis and interpretation of multivariate count data in heterogeneous market population. This model provides a representation of individual behaving differences in purchasing at the retail fields. Latent class analysis is employed to describe variation in the market segmentation based on individual recurrent choice data in POS focusing on a specified product. Application to real POS at a convenience store, which is one of typical typed in Japan, is presented for certification of the model. An EM algorithm is employed to estimate each parameter. Appropriate number of class should be selected being agreeable to practical use. Study on the propriety of class number and model identification will be added based on AIC or other information criterion.

Mixtures of latent Markov models for simple discrimination learning

Verena D. Schmittmann Ingmar Visser Maartje E. J. Raijmakers

Behavioral and Neuropsychological data suggest that multiple systems are involved in category learning. Through the application of mixtures of latent Markov process models, the existence of multiple modes of behavior in learning a simple discrimination is examined. The data stem from a cross-sectional study by Raijmakers, Dolan and Molenaar (Memory & Cognition, 2001), in which subjects aged 4 to 19 years carried out a simple discrimination learning task. Models with two learning modes provide a better fit to the sequential data than models with a smaller or higher number of learning modes. The development of the learning modes in the given age range is investigated with multi-group mixtures of latent Markov process models. All models were fitted by maximum likelihood estimation in the R-package *depmix* (Visser, 2005). The results support the hypothesis of distinct learning modes instead of a single general mode of learning with a continuum of appearances.

Latent Class Analysis of the Paired Comparison Data

Okubo Tomoya and Mayekawa Shin-ichi

In the paired comparison situation, where a person is asked to choose one of the two stimuli presented, he/she may not make his/her decision consistently through the whole session.
In this research, two latent class models for the analysis of the paired comparison data were presented.
Firstly, MXPREF-1, the latent class extension of the MDPREF model in which the subjects are treated as the observations to be classified, was developed.
Compared to the original MDPREF, this model gives us the consistent estimates of the parameters since the number of parameters to be estimated does not increase according to the number of subjects.
Secondly, MXPREF-2, the latent class extension of the MDPREF model in which the paired comparison judgments of each subject are treated as the observations to be classified, was developed.
Under this model, each paired comparison is classified into the latent classes which represent the different scales.

A Matching Method Based On Latent Class Models

Kazunori Yamaguchi, Fumitake Sakaori, Msao Kuji, and Michiko Watanabe

The propensity score matching methods are very often used for causal inference and sample selection bias in non-experimental settings in which (i) few units in the non-experimental comparison group are comparable to the treatment units, and (ii) selecting a subset of comparison units similar to the treatment units is difficult because units must be compared across a high-dimensional set of pretreatment characteristics.
We propose a new matching method based on latent class models, and implement it using data from the Internet survey. The proposed method can be used for the data with missing values. The method would be easier to handle interaction effects of covariates than the propensity score matching methods.

A Note On The Likelihood Ratio Test In Structural Equation Modeling

Reinoud D. Stoel, Francisca Galindo Garre, and Conor Dolan

The aim of the present paper is to show how the use of inequality constraints on parameters in structural equation models may affect the asymptotic distribution of the likelihood ratio test. Inequality constraints are implicitly used in the testing of commonly applied structural equation models, like the common factor model, the autoregressive model, and the latent growth curve model. Such constraints are the result of the null hypothesis placing the parameter value(s) on the boundary of the parameter space. It will be shown that the limiting distribution of the likelihood ratio cannot be treated as that of a χ^2-random variable with degrees of freedom equal to the number of constraints. Based on the work of Self and Liang (1987) and Stram and Lee (1994) the correct distribution for testing one, or a few parameters at a time will be inferred. An important message of this paper is that the traditional distribution of the test statistic (with df equal to the number of constraints) will have too heavy a tail.

References:
Self, S. & Liang, K.Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical* Association, *82*, 605-610.
Stram, D.O. & Lee, J.W. (1994). Variance component testing in the longitudinal mixed effects model. *Biometrics, 50,* 1171-1177.

Some Properties Of Likelihood Ratio And Nested Tetrad Tests Comparing Factor Models With Different Numbers Of Factors

Todd E. Bodner, and Timothy R. Johnson

The likelihood ratio test (LRT) and nested tetrad test (NTT) are two statistical tests for deciding between nested common factor models that differ by the number of factors. Application of the LRT in these cases is questionable given the violation of the regularity condition that parameters not lie on the boundary of the parameter space. Empirical and analytical analyses of the LRT sampling distribution under the null hypothesis for variance component and latent class models show that it can be far from chi-square. The NTT does not assume this regularity condition and thus may be a viable alternative when the LRT is not reliable. In this talk we discuss the results of Monte Carlo analyses of the sampling distribution of the LRT and NTT for nested common factor models -- in particular the Type I error rates for the tests. Results show that both tests tend to be conservative in smaller samples but that NTT slowly approaches the nominal error rate as the sample size increases. We discuss the implications of these results for the practitioner and recommend that the limitations of the LRT be more widely acknowledged in textbooks and didactic articles on factor analysis and structural equation modeling.

Evaluation of the Bayesian and Maximum Likelihood Approaches in Analyzing Structural Equation Models with Small Sample Sizes

Sik-Yum Lee and Xin-Yuan Song

The main objective of this paper is to investigate the empirical performances of the Bayesian approach in analyzing structural equation models with small sample sizes. The traditional maximum likelihood (ML) is also included for comparison. In the context of a confirmatory factor analysis model and a structural equation model, simulation studies are conducted with the different magnitudes of parameters and sample sizes n=da, where d=2, 3, 4 and 5, and a is the number of unknown parameters. The performances are evaluated in terms of the goodness-of-fit statistics, and various measures on the accuracy of the estimates. The conclusion is: for data that are normally distributed, the Bayesian approach can be used with small sample sizes, whilst ML cannot.

Normal Theory Likelihood Ratio Statistic for Mean and Covariance Structure Analysis Under Alternative Hypotheses

Ke-Hai Yuan, Kentaro Hayashi, and Peter M. Bentler

The normal theory likelihood ratio (LR) statistic is widely used in structural equation modeling. Under a sequence of local alternative hypotheses, the distribution of this statistic has been shown to approach a noncentral chi-square variate. In practice, the population mean vector and covariance matrix as well as the model and sample size are always fixed. It is hard to justify the validity of the noncentral chi-square distribution for the resulting LR statistic even when data are normally distributed and sample size is large. By extending results in the literature, this paper develops several normal distributions to describe the behavior of the LR statistic. A sequence of local alternative hypotheses is not necessary for the proposed distributions to be valid. Empirical results indicate that some of the distributions match that of the LR statistic very well even when sample size is relatively small, implying that the commonly used noncentral chi-square distribution should be replaced by the proposed normal distributions in power evaluation and model inference for mean and covariance structure models.

Imputation Of High-Dimensional Categorical Data Using Loglinear Models With Regard To Structural Zeroes

Frank van den Eijkhof

Official agencies for statistics often have information regarding forbidden combinations of values in a social dataset, like ‘people can not be married and less than 16 years old’. Such forbidden combinations are known as edits, or structural zeroes. These datasets mostly contain ca. 30 up to several hundreds of variables per person, for each person in a household. We propose a method for estimation of the parameters of loglinear models for such high-dimensional data with regard to a set of structural zeroes. The models under consideration contain main effects and small interaction effects. The proposed method is aimed at usage of data without missings, also known as complete data. It can be used in a simple imputation algorithm where parameter estimation is only based on the records without missings, or in an EM-algorithm where both records with and without missing data are used for parameter estimation. An experiment shows the feasibility of this method.

Modeling Nonignorable Missing Data in Speeded Tests

Jonald L. Pimentel, Cees A.W. Glas and A.A. Béguin

If a test is administered under a limited-time condition, items at the end of the test are often not endorsed. In most instances, the pattern of missing responses depends on the ability that is measured and, therefore, the missing data are not ignorable in statistical inference. In the present paper, the data are modeled using a combination of two item response theory models: one IRT model for the observed response data and one IRT model for the missing data indicator. The missing data indicator is modeled using the sequential model by Tutz (1990, also see, Verhelst, Glas & de Vries, 1997). The two IRT models are connected by invoking with the assumption that their latent person parameters have a joint multivariate normal distribution. The model parameters are estimated using maximum marginal likelihood.
Simulation studies showed that treating the missing data as ignorable leads to considerable bias in the parameter estimates. Further, it was found that including an IRT model for the missing data removes this bias in the parameter estimates.
A real data example is presented to illustrate the impact of the model is practical situations.

References:
Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39-55.
Verhelst, N.D., Glas, C.A.W., & de Vries, H.H. (1997). A steps model to analyze partial credit. In: W.J. van der Linden and R.K. Hambleton (Eds.), *Handbook of modern item response theory*. (pp. 123-138). New York, NJ: Springer.

Multiple Imputation Of Missing Data In A Multilevel Setting

G.W. Jacobusse, S. van Buuren, and C.G.M. Groothuis-Oudshoorn

Several approaches for dealing with missing data are used in practice. Multiple imputation (Rubin, 1987, 1996; Schafer, 1999) is a method to draw valid inferences from incomplete data. Draws from the multivariate distribution can be made from the incomplete data in a Gibbs sampling process (Van Buuren and Oudshoorn, 2000).
For clustered data, multiple imputation poses new challenges. Standard imputation models for exchangeable observations cannot be used in a multilevel setting since these would destroy the multilevel structure and lead to incorrect statistical inferences. The presentation will include a simulation study to demonstrate this.
We created a stand-alone software tool to generate multiple imputations for incomplete level-1 variables under the linear mixed model. The program is called MAMMI, and runs on the Windows platform. It is freely available. MAMMI creates multiple imputed datasets that can be analyzed using standard statistical packages. The current version is limited to linear models with a maximum of two levels.

Shorter Survey Questionnaires Using A Dynamic Form Of Planned Missing Data

Andreas Wolf and Ulf Kroehne

Many large-scale surveys suffer from problems like non-compliance due to high respondent's burden, ongoing rigidity in answering behaviour, lack of concentration or respondent's displeasure. Own empirical results support this theory. Even worse, the selectivity of the sample increases with longer tests as a – presumably non-random – share of subjects is not willing to complete lengthy questionnaires.
The proposed solution is a further development of "Multiple Matrix Sampling" and Raghunathan's survey questionnaire design. It is a flexible tool for test administration with planned missing data – call it "individualized" Multiple Matrix Sampling – where a dynamically selected subset of items is given to each test-taker without knowledge of item interrelationships prior to the survey (pre-study).
The stepwise item selection is based on a probabilistic function accounting for empirical relationships of those items not yet administered to the actual test-taker and all previously administered items. The resulting structure of missing values is optimal for subsequent multiple imputation.
Extensive results with respect to the resulting bias of estimates and the statistical power are presented for simulated data. The proposed test administration model is compared to other sampling techniques. The solution offers a data-based optimization of the sampling problem for a matrix-sampling approach on the individual level.

WEDNESDAY JULY 6

Invited speakers, 9.00h-9.40h

Structural Equation Modeling and Nonnormality

Yutaka Kano

This talk addresses nonnormality in structural equation modeling (SEM) and factor analysis (FA). The research on nonnormality in SEM and FA used to imply studies on methodology that can properly analyze possibly nonnormal data, except for few works, where second-order moment structures are analyzed. Recent studies, influenced by independent component analysis, have developed a new framework in SEM and FA that makes something impossible to be possible. In this talk, we first review recent results on identification, estimation, and testing hypotheses in several SEM models using nonnormal information. In some situations, use of nonnormal information solves the problems of equivalent models and saturated models.

Assessing fit to IRT models

Cees A. W. Glas

In the last ten years, the family of item response theory (IRT) models has expanded tremendously (see, for instance, van der Linden & Hambleton, 1997; de Boeck & Wilson, 2004; Skrondal & Rabe-Hesketh, 2004). However useful these models are, the applications are only valid if the models fit the data. Unfortunately, interest in model fit is often very limited, and the tools for evaluation of model fit are often poorly rooted in statistical theory. In this presentation, it will be shown that the fit of a broad class of IRT models can be evaluated using modification indices. Modification indices are computed using the estimates of a null-model (the IRT model of interest) and gauge the effects of adding parameters that represent violations of the null-model. The obvious advantage is that many model violations can be evaluated using estimates of the null-model only. Modification indices for IRT models can both be defined in a likelihood-based framework and Bayesian framework. In the first case, they take the form of Lagrange multiplier tests (Glas, 1999). The tests can be defined in such a way that they are based on residuals (differences between predictions from the estimated model and observations) that support evaluation of the severity of the model violation. In the second case, they take the form of Bayesian modification indices (Fox & Glas, 2005) that can be computed in an MCMC framework as an alternative to the well-known posterior predictive checks.

It is shown that the framework supports evaluation of model fit from two perspectives: the items and the respondents. In the first case, for every item, residuals and item fit statistics are computed to assess whether the item violates the model. In the second case, residuals and person fit statistics are computed for every person to assess whether the responses to the items follow the model. Examples of assumptions of IRT models that can be evaluated are subpopulation invariance (differential item functioning), the form of the item response function, and local stochastic independence.

The presentation will end with an example of evaluation of assumptions about the factor structure of a multidimensional IRT model in a Bayesian framework (Béguin, & Glas, 2001).

References:

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multi-dimensional IRT models. Psychometrika, 66, 471-488.

De Boeck, P., & Wilson, M. (Eds.) (2004). Explanatory Item Response Models: A generalized linear and nonlinear approach. New York NJ: Springer.

Fox, J.P., & Glas, C.A.W. (2005). Bayesian modification indices for IRT models. Statistica Neerlandica, 59, 95-106.

Glas, C.A.W. (1999). Modification indices for the 2-pl and the nominal response model. Psychometrika, 64, 273-294.

Skrondal, A., & Rabe-Hesketh, S. (2004). Generalized latent variable modeling. London: Chapman & Hall.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). Handbook of Modern Item Response Theory. New York, NJ: Springer Verlag.

Symposia and sessions, 9.40h-11.00h

Interaction Detection and Variable Selection through Regression Trunk Approach: Some Applications

Claudio Conversano

In multiple regression analysis, interactions among variables are modeled as cross-products between covariates. Usually the analyst specifies interactions terms in a model based on different principles, like *a-priori* hypotheses concerning the distribution of the data as well as theoretical considerations among the relationships existing between covariates. When dealing with many predictors, both categorical and numeric, the search of the most significant interactions is not straightforward but it could be a very demanding task since the analyst should test all of them and retain the most important.
The Regression Trunk Approach is a flexible tool for modeling effects in regression analysis. This method allows to assess automatically interactions between a categorical variable and covariates (e.g., the so called aptitude-treatment interactions in psychological studies). The final result is a multiple regression model representing interactions as thresholds instead of cross-products.
Aim of this paper is to discuss some theoretical properties of the Regression Trunk Approach, as well as to present some real data applications (using economic and sociologic data). The applications highlight the properties of this method as well as its effectiveness with respect to other parametric and nonparametric alternatives.

Learning from Non Standard Tree-based Models: a Customer Satisfaction Study in Transports

Roberta Siciliano, Massimo Aria, and Antonio Mango

In the framework of supervised classification and regression, tree-based methodologies can be understood as data-driven or learning procedures of data mining enabling the user for an handy interpretation - through a tree oriented graph - of the dependence relations of a response variable (i.e., the output) on a set of predictors (i.e., the input) measured on a large set of objects. Ensembles of decision trees can be considered either for assigning a new case to a response class or for forecasting the response value of a new case for that only the input is known. The overall interpretation can be greatly improved using an interactive software.
This paper provides non standard tree based models for classification and regression when dealing with special structures of data, namely multi-class responses, multi-block (internally correlated) predictors, ordinal responses and predictors, missing data etc. An interactive software is also presented that includes standard and non standard procedures. A case study consisting in the analysis of Customer Satisfaction in Transports is finally discussed.

Globally Optimal Tree Models

Bart Jan van Os

Since their introduction, classification and regression trees have acquired popularity in many scientific fields. Such trees are usually constructed by growing the tree from the top, while optimizing each new split locally given previously acquired splits. As a result, these trees may not be the best trees possible with respect to the misclassification rate or the residual sum of squared error in the terminal nodes.
Here, the optimization of a classification tree with discrete predictors will be pursued through a global objective function defined on the classes formed by the terminal nodes, with size constraints on the tree, such as restricting the maximum depth of the tree or the maximum number of leaves. In practice, trees for different size restrictions are constructed in the same process and a cross-validation procedure is used to choose the best sized tree.
The size constrained tree problems can be phrased as particular constrained partitioning problems defined on the measurement space. An exact exponential time Dynamic Programming algorithm for solving this problem will be given, related to an algorithm for globally optimal divisive hierarchical clustering. For large problems, exact solutions to approximate problems can be found.
Examples will be given, including a globally optimal regression tree problem with eleven predictors for a sample of size 102 and an heuristic application to a large database (78000 cases).

Type I error and Power of the Regression Trunk Approach

Elise Dusseldorp, Claudio Conversano, and Bart Jan van Os

The regression trunk approach (RTA) integrates regression trees (also known as CART) into multiple regression analysis in order to assess interaction effects between predictor variables in a regression context. A core element of RTA is a different representation of interaction, that is, as a *threshold* interaction, as opposed to the commonly used *cross-product* representation in multiple regression analysis. A threshold interaction effect occurs, when the effect of a predictor variable on a response is different for those persons who score above and those who score below a certain threshold value on another predictor variable. We focus in this paper on interaction effects between one nominal scale predictor variable and one or more continuous predictor variables. Initial results reveal that the Type I error and power of RTA varies across different decision rules that determine the best size of the regression trunk. The decision rule that outperforms the others is the following: choose the smallest regression trunk with a cross-validated prediction error that is lower than or equal to the lowest error plus 0.67 times the standard error.

An Approximation of Cronbach's α and its Use in Test Assembly

Angela Verschoor

In this paper a new approximation of Cronbach's α is presented. It is especially suited in the context of test assembly. Using this approximation, two test assembly models are introduced. Being non-linear models, they are solved by Genetic Algorithms as the commonly used Linear Programming methods cannot be used here. A comparison is made with existing test assembly models.

More powerful multiple-choice questions

Huub Verstralen, Gunter Maris, and Timo Bechger

Traditionally, a student responds to a multiple-choice question by selection exactly one option. Students have more freedom to respond if we allow them to express their opinion by giving a weight to each of the options. In this way, they may also provide us with more information on their ability. A well-known disadvantage is that the student will not necessarily optimized his score by answering truthfully. We present a number of IRT-models for subjective weights and explain how one might ensure that truthful responses optimize estimated ability.

Application of Classical test theory Using IRT

Timo Bechger, Gunter Maris, and Huub Verstralen

Notwithstanding the many developments in item response theory (IRT), classical test theory (CTT) continues to provide the main framework for test construction. It is therefore useful to have a clear idea about the relations between IRT and CTT. This should improve the appreciation of both theories and facilitate communication to researchers and item writers who are frequently more familiar with CTT than with IRT.
Here, we survey relations between CTT and IRT but focus on applications of CTT which become feasible when an appropriate IRT model is available. It is possible, for instance, to estimate CTT reliability and its confidence interval, even if the intended population never took the test.

Multidimensionality in Practice

Ron Engelen

In an actual test-administration it is often the case that the test measures several abilities simultaneously. This can be intended and even necessary (e.g. in an exam), or this can happen just accidentally. Moreover, these abilities are usually not independent. So, although a test can be multidimensional, it is common practice to report only a single number to represent (or summarize) the examinees test-results. The major reason for this is that it becomes much easier to compare examinees. In practice there are two methods available to compute this summarizing number. The first and most appropriate method starts with estimating a multidimensional model. Then, given the model, the scores on the different dimension are summarized in one number. The second method one ignores the multidimensionality and analyses the data with a one-dimensional model. In this paper it is investigated when the second method gives reasonable results.

Plotting Supplementary Variables In Correspondence Analysis Of Rating Data

Michel van de Velden and Anna Torres-Lacomba

In this paper we consider the use of correspondence analysis (CA) of rating data. CA of rating data allows a joint representation of the rated items (e.g. attributes or products) and individuals. However, as the number of individuals increases, the interpretation of the CA map becomes difficult. To overcome this problem, we propose a method that allows the depiction of additional variables, for example, background characteristics that may be of interest in identifying consumer segments, in the CA map. The idea we use is based on the representation of supplementary variables in ordinary CA. However, as the format of the additional variables is typically different from the rating data, a recoding is required. We illustrate our new method by means of an application to data of a product perception study for five cream soups.

Specific MCA With Special Reference To Questionnaires

Brigitte Le Roux

The motivation for devising Specific Multiple Correspondence Analysis (Specific MCA) was the need to analyze questionnaires with nonresponses, getting rid of the constraints of complete disjunctive coding while preserving the structural properties of MCA. In the paper, the method is presented within the general framework of Geometric Data Analysis, then we deal with the special case of questionnaires, following the line of the paper by Le Roux (1999) and the book by Le Roux & Rouanet (2004). We first recall the properties of principal directions and variables of a Euclidean cloud. Then we apply these properties to a protocol for which individual profiles are measures, and we derive the properties of biweighted Principal Component Analysis (PCA). Then we recall the formulas of MCA viewed as a particular biweighted PCA.
We then compare Specific MCA to standard MCA, writing inequalities between eigenvalues and studying the rotation of principal subspaces when one goes from the global analysis to the specific one. Last we outline two applications of specific MCA: Bourdieu (1999), on the publisher space in France, and Chiche & al (2000) on the political space of French electors.

References:

Bourdieu P. (2000). Une révolution conservatrice dans l’édition, Actes de la recherche en Sciences Sociales, 127.
Le Roux B. (1999). Analyse spécifique d’un nuage euclidien, Mathématiques, Informatique et Sciences Humaines, 37, 65-83.
Chiche J., Le Roux B., Perrineau P., Rouanet H. (2000) . L’espace politique des électeurs français à la fion des années 1990. Revue française de Science politique, 50, 463-487.
Le Roux B., Rouanet H. Geometric Data Analysis (2004). Dordrecht, Kluwer.

Interpretation Of Data In Multidimensional Space

Shizuhiko Nishisato

When data span multidimensional space, the distance between any two points viewed in two-dimensional space, for example, cannot become smaller when looked at in three-dimensional space. Thus, two points lying close to each other in two-dimensional space may be far apart in three-dimensional space. In contrast, it is guaranteed that two points lying far apart in two-dimensional space are far apart in higher-dimensional space. Yet, the general practice is to seek only close points in reduced space with the risk of misinterpreting data. Why? The correlation between two variables, viewed in two-dimensional space, cannot become larger when viewed in three-dimensional space. Thus, the correlation between two variables in multidimensional space is generally much smaller than the same correlation viewed in reduced space. Yet the general practice is to look at this over-valued correlation in reduced space for an interpretation of data. Why? These fundamental aspects of multidimensional space are often overlooked in the general practice of interpreting data in reduced space. The focus of this paper will be on possible consequences of interpreting multidimensional data in reduced space. Although this paper is confined within quantification of categorical variables, the discussion equally applies to analysis of continuous data.

Thirteen Different Ways To Define Correspondence Analysis

Michael Greenacre

This review paper is dedicated to Shizuhiko Nishisato in his 70th birthday year, as well as to others who – like Nishi – have played a foundational role in the early development of correspondence analysis. I take a look at thirteen different ways to define this method of data analysis, which already has a 70-year history itself! This lighthearted talk will touch on such topics as correlating, averaging, scaling, weighting, regressing, canonizing, homogenizing, biplotting, unfolding, SVDecomposing, RC-modelling, Burting, and finally transmogrifying categorical variables.

Nonparametric Estimation of Standard Errors in Covariance Analysis using Infinitesimal Jackknife Methods

Robert I. Jennrich

In 1980 Jennrich and Clarckson proposed a simple nonparametric method for estimating standard errors in covariance analysis. It is shown here the standard error estimates produce are asymptotically correct when sampling from any population with finite fourth moments and that their method is essentially an infinitesimal jackknife method. The only application considered by Jennrich and Clarckson was to estimating standard errors for orthomax rotated loadings in maximum likelihood factor analysis. It is shown their method may be used to estimate standard errors for any form of covariance analysis including factor analysis, covariance structure analysis, and structural equation modeling. It is also shown how to do everything for correlation analysis as well. These methods can have distinct computational advantages over ordinary jackknife and bootstrap methods.

A General Approach to Construct Confidence Intervals for Many Multivariate Techniques: A Structural Equation Modeling Approach

Mike W. L. Cheung

Null hypothesis significance testing (NHST) has been dominated in behavioral and social sciences for decades though it has been severely criticized as non-constructive to the advances of scientific knowledge. Using confidence intervals (CIs) on the parameter estimates is suggested to be a better alternative to the NHST. Instead of deciding whether the parameter estimates are statistically significant or not, CIs provide more information on the precision of the parameter estimates. Recommendations on using CIs are generally observed in many disciplines. One of difficulties in hindering researchers applying CIs is that it is necessary to derive CIs for specific models being used by large sample approximation using the delta method or by the computational intensive approach such as the bootstrap method. By formulating many multivariate techniques within the framework of structural equation modeling (SEM), a general approach to construct CIs is proposed. It will be shown in this presentation that it is easy to construct CIs for a variety of multivariate techniques such as differences on correlation coefficients, indirect effects and reliability estimates. This approach is readily generalized to nearly all multivariate techniques in psychological research.

Fitting Nonlinear Random Coefficient Regression Models With Nonstochastic Nonlinear Parameters Using LISREL

Jeffrey Harring

LISREL is well-suited to handle the general class of latent coefficient models in which the repeated measures are effectively captured by regression functions linear in their parameters. We propose to extend this framework to include nonlinear random coefficient models, provided that the design matrix $\mathbf{X}$ can be expressed as $\mathbf{X}(\boldsymbol{\gamma})$ where $\boldsymbol{\gamma}$ represents a vector of nonstochastic nonlinear parameters. The proposed model is ideal for many types of behavioral data. As an illustration, repeated measures data from a learning experiment is analyzed using both a modified exponential curve as well as a segmented polynomial with an explicit point of transition. LISREL syntax for all fitted models is provided.

Comparing Indirect Effects in Structural Equation Modeling: A Sequential Model Fitting Method

Wai Chan

In social science research, an indirect effect occurs when the influence of an independent variable on the dependent variable is mediated by an intervening variable. Various methods have been proposed in the literature for testing the significance of an indirect effect (e.g., MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). However, relatively little is known regarding the comparison of two or more separate indirect effects within a sample or across different samples in the area of structural equation modeling (SEM). By using some computer programs like LISREL, the equality of two indirect effects can in fact be tested easily through imposing a nonlinear constraint on the corresponding model parameters, followed by standard model comparison techniques such as the likelihood ratio test. Some other programs, however, do not support analysis with nonlinear constraint and as a result, the test will not be as trivial as it originally appears to be. In the present talk, a sequential model fitting method is proposed for the comparison of indirect effects in SEM. Various examples will be given to illustrate how the proposed method is implemented in EQS without triggering the command of nonlinear constraints. Results would be compared to the LISREL solutions derived from models with nonlinear constraints.

Dynamic Analysis of Multivariate Panel Data with Nonlinear transformations

Kees van Montfort

Many models for multivariate data analyses can be seen as special cases of the linear dynamic model, also referred to as the state space model or linear dynamic system. Our orientation will be on data reduction, on simplification and description, so that we will strive to find a low-dimensional set of latent state variables. To distinguish our technique from the more classical approach to system analysis, we have labelled it linear dynamic analysis. A second distinction to the usual application of state space analysis, is that we will develop our model specifically for situations where not one, but several observation units have been measured. This is a necessary reflection on common research interest, which should reflect stability not only over timepoints, but over subjects as well. A third distinction of our model is that it can handle non-numerical and non-normal measurements, which are a common occurrence in day-to-day research practice.
Building from the idea of mixture distributions, we propose to view the state space model as a multilayered regression model. The output variables are then viewed as an imprecise manifestation of an unobserved continuous process, i.e. the latent state variables. The EM-algorithm will be used to derive estimators of the unknown parameters. The obtained estimation procedure will be applied to an empirical example.

Methodological Issues In Fitting Structural Equation Models And State-Space Models: Examples From Dynamic Systems Models

Sy-Miin Chow, Ellen L. Hamaker, and Conor V. Dolan

State–space models and the Kalman filter have been compared to structural equation models (SEMs) and their associated estimation algorithms in various contexts. Over the last few decades, tremendous progress has been observed in representing state–space models and structural models. We compare the two modeling approaches in light of these new methodological advances, particularly in representing dynamic models with intermediate and large numbers of measurement occasions. The differences between state–space models and structural models are highlighted analytically and conceptually. We then provide three simulation examples to demonstrate the parallels and divergence between SEM and state–space modeling, while outlining certain modeling aspects that are difficult to implement using one approach but not the other. Possible ways in which SEM and state–space modeling can supplement their own respective strengths in representing human developmental and other related processes are discussed

Traits and States in Structural Equation Modeling

Ellen Hamaker

Several structural equation models have been developed to decompose observed variance into trait and state variance. Among these models are: the stability model (by Hertzog and Nesselroade), the latent state-trait model (by Steyer and Schmitt), the hierarchical LTS model (by Schermelleh-Engel et al.), the latent curve LTS model (by Tisak and Tisak), the higher-order item model (by Marsh and Grayson), and the stable trait, autoregressive trait and state model (by Kenny and Zautra).
We discuss how these models differentiate between the four sources of variance identified by Cronbach: (a) lasting-general, which we refer to as common traits; (b) lasting-specific, which we refer to as unique traits; (c) temporary-general, which we refer to as common states; and (d) temporary-specific, which we refer to as unique states. Based on this we discuss the underlying assumptions about traits and states associated with these different models. Also, we show how these models are related to (multiple indicator) latent growth curve modeling. In addition, we present a new structural equation model to disentangle traits and states which allows us to test some of the assumptions underlying the existing models. We illustrate this with empirical examples.

Joint Modeling of Repeated Measures and Survival data

Francisca Galindo-Garre and Aeilko H. Zwinderman

In many clinical experiments both repeated measures data and survival data are collected. In kidney graft failure studies, for example, measurements of serum creatinine concentrations are used as a marker to study renal disease progression. The relationship between repeatedly measured covariates and survival time is commonly assessed using the Cox proportional hazards regression model. However, this approach is not accurate when the repeated measurements are collected irregularly, as usually occurs in clinical practice. Recently, models to estimate the joint distribution of longitudinal covariates and survival time have become popular in the statistical literature. It is commonly assumed that the repeated measured data follow a linear mixed effects model and that the survival data follow a Cox model in which the previously estimated random intercepts and slopes are included as time-dependent covariates. However, the linear mixed effects model may not fit the longitudinal covariates very well. In this paper, we investigate the performance of non-parametric models for the modeling of longitudinal covariates when the number of observations per patient is very large and the observations are irregular (unbalanced data). A Bayesian approach is used to estimate the parameters. Concretely, posterior distributions are obtained using the WinBUGS program. Finally, Deviance Information Criterion (DIC) is used for model selection.

Assessment of Differential Item Functioning in Testlet-based Items using the Mantel-Haenszel Procedure and Logistic Regression

Ya-Hui Su and Wen-Chung Wang

This study attempts to assess the performances of the Mantel-Haenszel procedure (Holland & Thayer, 1988) and logistic regression (Swaminathan & Rogers, 1990) in detecting differential item functioning (DIF) for dichotomous items in testlet-based tests. Iterative test purifications of these two procedures were considered. Five independent variables were manipulated: (a) DIF detection procedure, (b) percentage of DIF items in a test, (c) testlet response model, (d) mean ability difference between groups, and (e) variance of the random effects for testlets. Through simulations, it was found that: (a) These two methods yielded very similar Type I error, even under two different testlet response models; and (b) As the variance of the random effects or the mean ability difference between groups was increased, the power of detecting DIF was decreased. We thus proposed a revised logistic regression method, to mitigate the effect of the variances of the random effects so that the power of detecting DIF can be increased.

Modeling Differential Item Functioning (DIF) Using Multilevel Logistic Regression Models: A Bayesian Perspective

Saengla Chaimongkol, Fred W. Huffer, and Akihito Kamata

A multilevel logistic regression approach provides an attractive and practical alternative for the study of Differential Item Functioning (DIF). It is useful for both identifying items with DIF and for explaining the presence of the DIF. Kamata and Binici (2003) first attempted to identify group unit characteristic variables explaining the variation of DIF by using the hierarchical generalized linear models. Their models were implemented by the HLM-5 software, which uses the penalized or predictive quasi-likelihood (PQL) method. They found that the variance estimates produced by the HLM-5 for the level 3 parameters are substantially negatively biased. This study extends their work by using a Bayesian approach to obtain more accurate parameter estimates. Two different approaches to modeling the DIF will be presented, referred to as relative and mixture distribution, respectively. The relative approach measures the DIF of a particular item relatively to the mean overall DIF of all items in the test. The mixture distribution approach treats the DIF as independent values drawn from a mixture distribution, which is a mixture of a normal distribution and a discrete distribution concentrated at zero. A simulation study is presented to assess the adequacy of the proposed models.

A New Method for Assessing the Statistical Significance of Item-Level Differential Functioning Indices in the Differential Functioning of Items and Tests (DFIT) Framework

Nambury S. Raju, T.C. Oshima, and Alice Nanda

A new method is proposed for assessing the statistical significance of non-compensatory differential item functioning (NCDIF) indices associated with the framework of differential functioning of items and tests (DFIT). In this new method, a cutoff score for each item is determined by creating a frequency distribution of NCDIF values under no DIF condition by generating a large number of item parameters based on the item parameter estimates and their variance-covariance structures from a computer program such as the BIILOG-MG3. This cutoff for each item can be used as the basis for determining whether a given NCDIF index is significantly different from zero. This new method applies to dichotomously-scored items and has definite advantages over the current method where an NCDIF index greater than .006 with a statistically significant chi-square is deemed significant (Raju, van der Linden, & Fleer, 1995); this new method yields cutoff values that are tailored to a particular dataset and a particular item. A Monte Carlo assessment of this new method is presented and discussed.

Locating DIF-Free Items as anchors Using the Iterative Constant Anchor Item Method

Ching-Lin Shih and Wen-Chung Wang

This study proposed an iterative constant anchor item method for locating a DIF-free item as the anchor item for subsequent DIF analysis, and investigated its performance through Monte Carlo simulations. Five independent variables were manipulated: (a) item response model: the Rasch model and the 3-parameter logistic model; (b) test length: 20, 30, and 40 items; (c) percentage of DIF items in the test: 10%, 20%, 30% and 40%; (d) magnitude of DIF: 0.4 and 0.6; (e) DIF favor: one-sided (all DIF items favored the reference group), dominant (most DIF items favored the reference group) and balance (one half of DIF items favored the reference group and the other half favored the focal group). The dependent variable was the accuracy of DIF-free items being located.

The results showed that when the magnitude of DIF was 0.4, the accuracy of locating DIF-free anchor items for the Rasch model was greater than 84.9% under all conditions, and greater than 96% under most conditions, especially under the balance condition. For the 3-parameter logistic model, the accuracy was greater than 92.2%, except at the first iteration under the one-sided DIF condition. The accuracy was increased as the number of iteration was increased. The accuracy was perfect under all conditions when the magnitude of DIF was 0.6.

Poster session, 11.00h-11.40h

A Three-Latent Class Variable Model of Attention Problems

Levent Dumenci

In clinical practice and research, diagnostic status is often determined by a cutpoint imposed on either a symptom-count or a symptom sum-score distribution. Neither procedure is satisfactory because the cutpoint procedures assume that (a) there is a categorical variable that underlies the continuous distribution, (b) the threshold of this categorical variable can be determined by expert judgment, and (c) the confidence with which individuals are assigned to each diagnostic category is the same for all individuals. To overcome these problems, a three-latent class variable model was proposed to make diagnostic decisions for attention problems. The model was applied to a nationally representative sample (N = 2,619) of the U.S. population ages 6 to 18. Fit statistics indicated that the three-latent class variable model provides better fit than (a) an independence model, (b) a traditional one-latent class model (Attention Problems: disordered vs. nondisordered), and (c) a two-latent class model (Inattention and Hyperactivity-Impulsivity: disordered vs. nondisordered). In addition to the two latent classes of Inattention and Hyperactivity-Impulsivity, the three-latent class variable model included a third latent class variable of Certainty (most-likely vs. somewhat-likely) to reflect the confidence with which the diagnostic assignment is made.

A Study Of Improvement Of EQ Test Used For Employment

Ikko Kawahashi, Hideki Toyoda, and Nobutake Matsushita

We examined the robustness of EQ test against response bias in recruit situations, using paired comparison methods expressed by structure equation modeling, and then standardized the test.
A paired comparison EQ test assumes social desirability that Edward's EPPS(1959) defined to be included in the constant term of each item in the paired comparison method presented by structure equation modeling. Thus the robustness to response bias is obtained by separating constant term from individual latent factor score.
In experiment 1, paired comparison type and Likert type EQ tests were adminisitered to 140 subjects in recruit and non-recruit situations, with the purpose of evaluating the above robustness. As a result, standard deviations, effect size and convergent validity showed that the paired comparison test was more robust against response bias than the Likert type test in both situations.
In experiment 2, we added data obtained from another set of 383 subjects to the data in experiment 1 and analyzed them with the purpose of stabilizing the scoring. Then, the items in the test were divided into 10 subsets to yield parallel tests, and the correlation of scales between parallel tests were calculated. As a result, parallel tests supported from a viewpoint of concurrent validity was obtained.

A State-Trait Model for Individual Differences in Reasoning with Conditionals

Vautier Stephane, Bonnefon Jean-François, and Eid Michael

Insofar as solving Modus Tollens (MT) problems (if p then q; not-q; is p true? Correct answer: No) is considered a matter of general reasoning competence, the question arises whether the solution of such problems indicates a general and temporally stable ability. In order to answer this question a latent state-trait model for dichotomous variables was applied to 6 MT problems (representing concrete everyday problems) answered by 486 adults (mean age = 31, SD = 13) on two occasions of measurement with a time lag of 26 days. Correct answers were coded 1, incorrect 'Yes' and 'Perhaps' answers were coded 0. A model specifying 6 inter-correlated trait factors, 2 uncorrelated state factors and measurement invariance of thresholds and loadings fitted the data well, $\chi^2(30, N = 486) = 39.24$, $p = .12$, RMSEA = 0.025. Estimates suggest that there is substantial stability of the ability to solve these problems (temporally stable variance accounted for 58 to 80% of the total variance). However, the ability strongly depends on the context in which a problem is embedded and generalizes rather weakly across contexts (inter-correlations of traits ranged from .31 to .61).Consequences of these results for the measurement of reasoning abilities are discussed.

Reliability and Factor Analysis of The Beck Depression Inventory (BDI-21) on Shahid Chamran University Students

G. Rajabi

The original version was administrated the Beck Depression Inventory (BDI-21) on 392 Shahid Chamran university students. The aim of adminstration this inventory was reliability and factor analysis investigation. The results obtained of principal components analysis showed that the first factor accounted for the greatest part of the variance above inventory items. In addition, Cronbachs Alph and split-half coefficients were 0.85 and 0.82 for whole inventory, and test-retest (with 3-week interval) reliability coefficient obtained 0.53. Also, a coefficient of 0.55 was obtained between the BDI-21 and MMPI-D scale.

Item Difficulty Parameter Estimation Using Paired Comparison and the Idea of Computerized Adaptive Testing

Koken Ozaki and Hideki Toyoda

In test operations using IRT, items are included in the test before being used to rate subjects and the response data is used to estimate their item parameters. And, after equating, the items may be included in an item pool that can be used for several tests. However, this test operation method contains the probability of item content leakage. Thus, estimating item parameters while keeping the item contents secret would be useful. To address this problem, Ozaki and Toyoda (2004) developed an item difficulty parameter estimation method that does not use data from subjects but utilizes paired comparison data from the perspective of the difficulty of items by a rater familiar with the field. In this article, the idea of Computerized adaptive testing (CAT) is applied to that method. Thus, smaller number of comparison will be enough for item difficulty parameter estimation with good accuracy. The estimation accuracyof this model was confirmed in a simulation study, and the feasibility of its use in practical settings is demonstrated using actual data.

True Intraindividual Change Models: An Example Of Stress Appraisals And Coping

Vesna Buško Alija Kulenović

The study examines intraindividual variations in self-report measures of coping with low-control stress by means of the true change modeling procedures. Within the particular class of structural equation models, as put forward by Steyer et al. (1997, 2000), the true intraindividual change scores between two occasions of measurement correspond to the values of endogenous latent variables. Such a specification further enables examining correlates and predictors of true change between the measurements. Transactional stress and coping theory views the processes of coping as context-specific and varying across time and the attributes of concrete stressful situations. Yet, substantial interindividual differences in the change scores can still be found in the measures of coping with any kind of situational demands. The data on cognitive appraisals and three broad situation-specific coping measures were collected on two occasions on the sample of 421 military basic trainees. The tested structural equation models and the parameters obtained point to different patterns of relationships of latent state and latent change coping measures with particular antecedent and appraisal variables.

A Case Study In Psychometric Epidemiology - Latent Structure Analysis Of The "Scaled" 28-Item General Health Questionnaire (Goldberg And Hillier, 1979) In A Birth Cohort Sample In Mid Adult Life

Tim J. Croudace, George B. Ploubidis, Rosemary A. Abbott, Diana Kuh , Peter B. Jones, Michael E.J. Wadsworth, and Felicia A. Huppert

Most measures of mental health in current use were developed as screening tools to discern a level of morbidity near a threshold at which illnesses are diagnosable. Epidemiological studies require instruments that yield reliable scores across a wider range of health. We examined the latent structure of the 28-item, scaled version of the General Health Questionnaire (GHQ-28). We applied methods for confirmatory factor analysis of categorical data and item response theory (ULS, WLS and WLSMV estimation of the probit-probit factor model in Mplus; ML estimation of two parameter logit-probit model in MULTILOG, TWOMISS and Mplus). A single continuum of "General Health" was invoked and tested by a second-order factor model. Data were drawn from an adult midlife follow-up of a British birth cohort study. Confirmatory analysis supported the *a priori* factor structure under traditional scoring 0-0-1-1 for both men and women samples. The second-order factor was able to explain almost all of the associations among the first-order factors (all loadings > 0.78). Test characteristic curves revealed the limits of the effective measurement range for screening questionnaires for binary (0-0-1-1) and trichotomous (0-1-2-3) scoring. Implications for epidemiological research on life-course antecedents of adult mental health problems are explored.

The Development of an Item Parameter Calibration Program for IRT Modelswith the Improved Initial Estimates and Ability Distribution Estimation Method

Mitsunaga Haruhiko and Mayekawa Shin-ichi

In this article, the ways for calculating initial values of the item parameters and for estimating the latent ability distribution under the normal assumption were developed. The first method for the initial values uses the observed item correlation matrix and decomposes its logarithm into the sum of log item discrimination parameters. The second method decomposes the logit of the item-responses, which are inversely weighted by the corresponding item discrimination parameters, into the differences of the ability parameter and the item difficulty parameter. Simulation studies showed that using proposed initial value requires less iterations than using existing ones. In most of the IRT programs, the ability distributions are treated as discrete when the item parameters are estimated in the m-step while it is customary to assume those distributions are continuous when the ability distributions were estimated. Since this may make numerical calculation process instable, a new method was proposed to estimate the ability distribution under the assumption that it is multinominal whose probabilities are proportional to the normal density. Simulation studies showed that the proposed method correctly recovered the true structure of the ability distributions. A program that implements these new methods was developed and compared with the existing IRT programs.

Bartlett Correction of Likelihood Ratio Statistics in Structural Equation Modeling

Okada Kensuke, Hoshino Takahiro, and Shigemasu Kazuo

Model selection is one of the most important steps in the application of structural equation modeling (SEM). The likelihood ratio test statistic, T, is a commonly employed index in this process. Hypothesis testing can be performed on the basis that T asymptotically follows the chi square distribution. A number of fit indices have been proposed, most of which are based on the assumption that T asymptotically follows the chi square distribution. When the sample size is small, however, the distribution of T deviates considerably from the chi square distribution. This problem, especially pronounced when the number of indicators per factor is large, is serious because it violates the theoretical justification of utilizing not only T but also many other fit indices for model selection. Here we propose a Bartlett correction in a strong sense of T to improve its approximation to the chi square distribution. The efficacy of our method has been evaluated by Monte Carlo simulation. The result shows that our method is superior to the current standard.

Mean And Covariance Structures For Specific Factors With Longitudinal Data

John Tisak and Marie S. Tisak

Based on Spearman's two factor theoretical model, we highlighted the role of the specific factors and related their parameters to the important notions of stationarity or invariance, in longitudinal research designs. In particular, several methodological concerns are presented such as (a) the impact of specific factor parameters on the stationarity of the regression parameters of intercepts and slopes, (b) the procedure for the identification of the model's parameters, including estimable and non-estimable parameters, (c) a method for the estimation, testing, and fit of these longitudinal models, and (d) the influence of variable-level, whether the variables represent items or scales. Two studies were presented to exemplify these methodological issues. Specifically, a homogeneous instrument on affective commitment was used to assess this construct in the military and the second illustration was based on the Primary Mental Abilities tests measured on adolescents.

The Application of Multilevel Structural Equation Modeling to Investigate Students and Classrooms' Effects on Mathematics Achievement

Sungworn Ngudratoke and Haniza Yon

It has long been believed that educational context has hierarchical or nested structure which implies that independent assumption, which is required for almost analyses is always not satisfied. Multilevel data analysis such as hierarchical linear modeling (HLM) has been proposed to handle such drawback. However, HLM cannot be applied to nested data in which variables in the model form causal relationship. Structural equation modeling (SEM) has been proposed to solve the problem with using HLM in analyzing nested data with causal relationship.

The primary purpose of this study was to investigate the effects of students and classrooms on mathematics achievement in the United States by using multilevel SEM. In addition, this study also intended to compare the effectiveness of the results from the HLM and SEM analyses. Data from 7,624 students and 315 teachers participated in the Third International Mathematics and Science Study (TIMSS) were used in the analyses. The model composed of two levels: student and classroom levels. Multilevel SEM was applied to link the path models at the two levels. Preliminary analyses indicated that the model fit the data well. Comparisons of the effectiveness of the results between HLM and SEM showed that each program had different strengths.

Gini's Generalized Index And Perceived Scatter

Nicolas Gauvrit and Jean-Paul Delahaye

Many index of scatter have been suggested in mathematical applications. These paramaters are intended to measure the clustered or, on the contrary, scattered characteristic of a two-dimensional patterns of points.

But two different notions of scatter must be distinguished. In a classical mathematical point of view, scatter means departure from central trend - measured, for instance, by standard deviation, but also by most of the usual scatter parameters. However, in a more intuitive point of view, 'scatter' means 'regularity'. To the best of our knowledge, no usual parameter is an accurate measure of scatter-as-regularity on geometrical two-dimensional material.

We will (1) define a general class of modified Gini's index (a mean of f(d) where f is an homeomorphism from R+* to itself, and d stands for one-to-one distances between points of the pattern), (2) show that mathematical relevance implies that a sub-class should be considered, and (3) prove that psychological relevance leads to the use of a special Gini's index, which is the mean of all exp(r.ln(d)) where d stands for the one-to-one distances between points of the pattern, and r lies between 0 (excluded) and 0.3.

IRT Model Selection By Genetic Algorithm

Kojiro Shojima

The situation in which the model to be applied differs through items must not be suprising, even natural. The way of applying a model apriori to a test or questionnair should not be recommended, especially to items which propeties are not known well. At least in the phase of developing a test, applying various models to items is effective to deepen the findings about items. Especially, it is necessary to try various patterns of models for the items with categorical ordered scores, since there are some candidates for polytomous models. However, there are K^n patterns of model combinations when the numbers of items and models are n and K, respectively. It is impossible to explore all patterns in practice when n and K are large. Moreover, a lot of patterns that need not be even tested exist. Therefore, a efficient method is required to search the best or acceptable pattern. The purpose of this study is to propose a framework for searching such patterns for the test by genetic algorithm.

The Impact of Linking Methodology on MIRT Parameter Recovery in Vertical Scaling

Haniza Yon and Mark D. Reckase

The primary purpose of this study was to investigate the impact of two multidimensional linking methods on multidimensional item response theory (MIRT) parameter recovery in vertical scaling. The two linking methods were a method developed by Oshima, Davey & Lee (2000) and a new method which was based on the non-orthogonal procrustes solution. The two methods were compared in term of the degree of MIRT parameter recovery in vertical scaling across various testing conditions such as number of common items and sample sizes. A compensatory multidimensional extension of the three-parameter logistic model with three-dimensions was employed to analyze both real and simulated data . Real data were obtained from a set of science test battery given to grade 3 to grade 8 students in the United States. Two summary statistics, namely bias and root mean squared error, were used as evaluation criteria to summarize item parameter recovery. As for the recovery of the vertical scale, the results were summarized using the properties of (1) average grade-to-grade growth, which examined the differences of the means of the scale score distributions at adjacent grade levels and (2) grade-to-grade variability, which examined the differences in standard deviations between adjacent grade levels.

Optimizing Sample Size And Power In Sequential Hypothesis Testing: The CLAST Rule.

Carmen Ximénez, Javier Revuelta, Juan Botella and Manuel Suero

Experimental psychologists usually establish the sample size by using the fixed-sample rule (FSR), in which the number of observations is determined in advance. However, several studies have demonstrated that sequential stopping rules are more efficient than FSR in terms of sample size and power.
Mathematical sophistication has limited the use of sequential rules. For this reason, we propose the CLAST rule (Composite Limited Adaptive Sequential Test). CLAST is a very simple sequential rule that can be easily implemented by experimental psychologists performing statistical tests. We compare CLAST with FSR and other sequential rules by using a Monte Carlo study. The conditions of the simulation involve different variants of the CLAST rule, effect sizes and sample sizes. Two statistical tests are used: the one-tailed *t* test of mean differences with two matched samples and the chi-square independence test for twofold contingency tables.
Results show that the CLAST rule requires about 35% less observations than FSR to achieve the same power. In addition, the efficiency of CLAST increases with effect size. In conclusion, sequential rules may improve the design of psychological experiments. However, more research should be conducted since we have used a limited number of test statistics and conditions.

Behavior Genetic Analysis Of Longitudinal Data By Inteligence Test

Kentaro Nakamura, Hideki Toyoda, and Yukimasa Muraishi

A genetic ACE(Additive genetic, Common environment, non-shared Environment) model combined with a model analyzing for longitudinal data applies to twin and non-twin data about Tanaka B intelligence test(new edition) simultaneously. The sample, examined at the time of junior high school and high school, analyzed by this method, which is one of the structural equation models, includes 115 monozygotic twin pairs, 32 dizygotic twin pairs and 881 non-twins. Intelligence score(IS) and seven scores of subordinate tests are examined. Non-shared environment accounts for the variance of the each score considerably, at both of the period of junior high school and the high school.

A Statistical Test of Within-Friendship Similarity

Kentaro Hayashi, Yiyuan Xu and Jo Ann Farver

Whether reciprocated friends are similar in their individual characteristics remain an open question in the psychology literature largely due to a lack of formal statistical test procedure. The analysis of friendship data has encountered difficulties because (i) number of reciprocated friendships may vary across individuals; and (ii) friendship pairs and non-friendship pairs are usually not independent. In this paper we propose a new method to analyze such reciprocated friendship data. A test statistic was derived based on matrix algebra and matrix differential calculus. The standard error formula was examined using two re-sampling methods, bootstrap and jackknife. We also analyzed a real data set and showed that within-friendship similarity in individual characteristics such as temperamental negative affectivity can be addressed using this new statistic.

Keynote speaker, 11.40h-12.40h

Latent Variable Models for Categorical Responses: New Developments and Trends

Irini Moustaki

In this talk, we review a variety of latent variable models for categorical responses in a general framework. We distinguish between static and dynamic models, models that incorporate effects of nonlinear terms as well as covariates. Different estimation methods have been proposed in the literature such as limited information methods, full maximum likelihood methods that uses algorithms such as the E-M, Newthon-Raphson or hybrid ones as well as Bayesian estimation methods.We give emphasis on model testing and outliers searching. Finally, we discuss the issue of identifying response patterns that are over represented in a sample. For example, if all items are binary, giving 'correct' or 'positive' answers to all items can be the result of guessing or avoidance to think properly each question. We propose a methodology that uses the model to predict the proportion of extreme response patterns in the data and simultaneously estimates the latent variable model parameters using the outlier-free part of the data.

THURSDAY JULY 7

Invited speakers, 9.00h-9.40h

Applications of Sampling-based Computing Methods in Analyzing some Psychometric Models.

Sik-Yum Lee

Sampling-based computing methods have been widely appreciated in analyzing complicated statistical models in various fields. In this talk, applications of these methods in solving some psychometric problems will be discussed. Basic ideas in applying the data augmentation, Expectation-Maximization (EM) algorithm, Gibbs Sampler, importance and path sampling to achieve the maximum likelihood and Bayesian solutions will be presented in the context of a framework that involves a generic model with latent variables and/or complex data structures. The methodology will be illustrated via applications to structural equation models, robust estimation and classification.

A New Approach To The Analysis Of Conditional And Average Mean Differences Between Groups

Rolf Steyer

Can there be anything new in the analysis of mean differences after almost a whole century of analysis of variance, regression models and comparison of means? The intriguing answer is “yes”. The first example in case is non- orthogonal analysis of variance. I will show that none of the programs available in the standard statistical packages such as SAS, SPSS, Systat, etc. provides a correct analysis of the main or average effect of a treatment variable even though all these programs offer three or more ways of computing the sums of squares (type I, II, III, and IV), at least if “main” or “average effect” is defined in a way compatible with the Neyman-Rubin approach to causality. A second example is analysis of covariance. I will show that Ancova does not, in general, yield correct results, if there is an interaction between the covariate and the treatment variable. Although this finding will not be a surprise to many, because, the model assumes “no interaction”, the question remains “How to analyse such a model?” A linear regression analysis with interaction terms also does not give the correct results either, if the covariate is stochastic, which is the standard situation in observational studies. This is my third example. And what about the fourth example, in which the covariate is latent and there is an interaction between the covariate and the treatment? Just take the manifest proxis? Wrong again! The new approach presented provides correct solutions for the analysis of conditional and average mean differences in all these examples. It uses standard programs for structural equation modelling that can handle nonlinear constraints.

Symposia and sessions, 9.40h-11.20h

Density Estimation by Total Variation Regularization: A Pilot Study

Ivan Mizera, and Roger Koenker

L1 type penalties have proven to be an attractive regularization device in various problems of non-parametric regression, image reconstruction, and model selection. In particular, penalties based on total variation are known to be capable of capturing sudden changes in the fitted function while still maintaining the general smoothing character. We explore the use of penalties penalizing the total variation of the estimated density, its square root, and its logarithm --- or their derivatives --- in the context of univariate and bivariate density estimation, and compare the results to some other density estimation methods --- penalized likelihood methods using penalties of L2 type, or others. Our objective is to elucidate an approach that would exhibit a unique philosophy regardless of dimension, would be capable of detecting qualitative features, like the number of modes of the fitted density, and last but not least, would possess an efficient and widely accessible implementation --- to this end we explore finite-element strategies in the framework of modern interior-point methods for convex optimization.

Regularization Methods in Multivariate Analysis

Yoshio Takane, and Heungsun Hwang

We discuss simple regularization methods in two representative multivariate data analysis techniques, generalized (multiple-set) canonical correlation analysis (GCANO) and redundancy analysis (RA). CANO and RA subsume a number of existing multivariate data analysis techniques including multiple correspondence analysis (MCA), 2-set canonical correlation analysis (CANO), discriminant analysis (DA), and so on. We incorporate a ridge type of regularization in these techniques and demonstrate the situations in which it is most effective. Regularization is deemed important as a way of supplementing insufficient data by prior knowledge, and/or of incorporating certain desirable properties in the estimates of parameters. We also discuss some mathematical properties of a matrix operator involved in the ridge type of regularization methods and their implications.

Gradient Directed Regularization for Linear Regression and Classification

Jerome H. Friedman
(with collaboration of: Bogdan Popescu)

Regularization in linear modeling is viewed as a two-stage process. First a set of candidate models is defined by a path through the space of joint parameter values, and then a point on this path is chosen to be the final model. Various pathfinding strategies for the first stage of this process are examined, based on the notion of generalized gradient descent. Several of these strategies are seen to produce paths that closely correspond to those induced by commonly used penalization methods. Others give rise to new regularization techniques that are shown to be advantageous in some situations. In all cases, the gradient descent pathfinding paradigm can be readily generalized to include the use of a wide variety of loss criteria, leading to robust methods for regression and classification, as well as to apply user defined constraints on the parameter values, all with highly efficient computational implementations.

Individual, Conditional And Average Causal Effects. An Introduction To The Symposium

Rolf Steyer

A new formulation of the Neyman-Rubin concepts of individual, conditional and average causal effects is presented which is the theoretical background to the individual contributions to this symposium. In line with Neyman, but in contrast with Rubin's original formulation the individual causal effect is defined as the difference between two true scores: the true score of an individual under treatment and its true score under control. Hence, an individual causal effect is the value of a latent variable. This does not only provide a more satisfactory nondeterministic definition of individual effects but also opens the floor to latent variable modelling and to new models that also allow estimating the variance of the individual causal effects.

Analysis Of Causal Effects With Mixed Models

Safir Yousfi

Recent work of Steyer and colleagues focused on analysing individual and average causal effect by latent variable models. It is shown that some of these analyses can be performed with mixed models that allow for the specification of random individual effects. The different approaches to causal effect analysis are compared with respect to generality and extendibility of the model, practicability, easiness of model specification, distributional assumptions. It is shown, how to specify a mixed causal model and how to relate the output of software packages to the theory of causal effects.

A Didactic Tool For Teaching The Theory Of Individual And Average Causal Effects

Benjamin Nagengast

We will present a didactic tool that illustrates the basic concepts of the theory of individual and average causal effects. The user can set up a three-group design by specifying expected outcome variables under treatment and under control for each of up to 10 units. Individual sampling and assignment probabilities can be fixed independently for every unit. The program then calculates - for the whole group and for each subgroup if a covariate is specified - the causally unbiased expected values under treatment and under control as well as individual and average causal effects. Furthermore the expected values under treatment and under control, the prima facie effects, baseline and differential effects biases are computed automatically. Additionally, several criteria for causal unbiasedness of the unconditional expected values are tested. The presentation will provide an introduction into using the tool for teaching and autodidactic learning . Core concepts of the theory of individual and average causal effects (e.g., unconfoundedness, conditional unbiasedness) will be illustrated with instructive examples.

Analyzing The Effect Of Negatively Formulated Items. A First Application Of The Individual Causal Effects Model

Steffi Pohl, and Rolf Steyer

Using the individual and average causal effects model (Steyer 2005), the answers of 503 persons at four points in time on the “Multidimensionaler Befindlichkeits-fragebogen”, a questionnaire measuring three dimensions of mood states (good-bad, awake-tired, calm/restless), were analysed using LISREL. The aim was to identify the individual effects of negative item formulation such as ‘bad’ instead of ‘good’, or unwell instead of well. To analyse these effects the participants were randomly splitted into two groups. In the control group the answers to the positively formulated items (good, well, etc.) at the four measurement times were analysed. For the analysis of the treatment group the answers of the positive formulated items were only used for the first two measures; in the last two measures the answers to the negative items were used. The individual causal effects variable was then related to traits measured by the Freiburger Personality Inventory explaining a part of its variance.

Simultaneous Procrustes Transformation of Components and Loadings Obtained From Three-Way Data

Kohei Adachi

Component analysis of an n-objects by m-variables matrix $\mathbf{X}_k$ (k = 1, …, N), which is a frontal slice of a three-way array of N sources by n by m, gives component $\mathbf{G}_k$ and loading $\mathbf{B}_k$ with $\mathbf{X}_k \cong \mathbf{G}_k\mathbf{B}_k' = \mathbf{G}_k\mathbf{S}_k\mathbf{S}_k^{-1}\mathbf{B}_k'$. Here, $\mathbf{S}_k$ is an arbitrary nonsingular matrix. In order to evaluate inter-component and inter-loading differences among sources, I consider a method in which $\mathbf{G}_k$ and $\mathbf{B}_k$ are transformed symmetrically so that $\mathbf{G}_k\mathbf{S}_k$ and $\mathbf{B}_k\mathbf{S}_k'^{-1}$ are matched to consensus component $\mathbf{F}$ and loading $\mathbf{A}$, respectively. In this method,

$$\phi(\mathbf{F}, \mathbf{A}, \mathbf{S}_k) = \frac{1}{n}\sum_{k=1}^{N}\left\|\mathbf{F}-\mathbf{G}_k\mathbf{S}_k\right\|^2 + \frac{1}{m}\sum_{k=1}^{N}\left\|\mathbf{A}-\mathbf{B}_k\mathbf{S}_k'^{-1}\right\|^2 \quad (1)$$

is minimized alternately over [$\mathbf{F}$, $\mathbf{A}$] and over each $\mathbf{S}_k$. The minimization of (1) over $\mathbf{S}_k$ seems not easy, since (1) is a function of $\mathbf{S}_k$ and its inverse. I thus reparametrize $\mathbf{S}_k$ and $\mathbf{S}_k'^{-1}$ using their SVD as $\mathbf{S}_k = \mathbf{P}_k\Lambda_k\mathbf{Q}_k'$ and $\mathbf{S}_k'^{-1} = \mathbf{P}_k\Lambda_k^{-1}\mathbf{Q}_k'$. Then, (1) is rewritten as

$$\phi(\mathbf{P}_k, \Lambda_k, \mathbf{Q}_k | \mathbf{F}, \mathbf{A}) = \frac{1}{n}\sum_{k=1}^{N}\left\|\mathbf{F}-\mathbf{G}_k\mathbf{P}_k\Lambda_k\mathbf{Q}_k'\right\|^2 + \frac{1}{m}\sum_{k=1}^{N}\left\|\mathbf{A}-\mathbf{B}_k\mathbf{P}_k\Lambda_k^{-1}\mathbf{Q}_k'\right\|^2 . \quad (2)$$

The minimization of (2) over Λ_k amounts to solving quartic equations, and minimizing over $\mathbf{P}_k$ and $\mathbf{Q}_k$ can be attained using the existing procedures. We perform a simulation study to assess the above algorithm.

Warped Image Factor Analysis

Sungjin Hong

The primary motivation of three-way factor analysis is to fit systematic data variance underlying a three-way data array (e.g., I subjects' rating on J variables at K occasions), instead of a two-way facet from it. In factor analysis of sequential data (e.g., time-series or digitized images), the measurement sequence remains "intact" and is assumed to be consistent across all measurement conditions. Otherwise, recovered sequential factors would be distorted. Shifted and Warped Factor Analyses (SFA and WFA) explicitly fit such measurement-sequence inconsistency. Warped Image Factor Analysis (WIFA) combines two ideas: (a) fitting systematic shape variation of image factors, and (b) decomposing many 2D images into a few image factors. WIFA allows image factors to change shape *independently*, unlike what is assumed in a data-level sequence adjustment: *synchronized* shape changes of image factors. The latent-level shape variation modeled in WIFA seems to make recovered factors "unique" in some two-way cases, as in SFA and WFA. The shape variation of image factors is parameterized as linear *warping* of segmented images. A quasi-ALS (alternating least squares) algorithm for WIFA will be described, which uses alternating linear regression for factor weights and nonlinear optimization for warping-size parameters. A simulated example will be presented.

A Parsimonious Model of the Interactional Space of Service Encounters: Measure Development

Marc A. Tomiuk, Blake E. Ashforth, Ursula Hess, and Heungsun Hwang

Service encounters represent interpersonal/social encounters which are also goal-oriented, economic activities. It is well established that during service encounters frontline personnel are encouraged to actively engage in *emotional labor*, a form of impression management; thereby modulating their affective displays to clients. It is argued that during service encounters, emotional labor represents the primary input to impression formation in customers and that interpersonal trait inferences will take place over the dimensions of *affiliation*, *dominance*, and *competence/mastery*. In turn, these dimensions are taken to represent the *Interactional Space of Service Encounters*. The primary purpose of this paper was to develop valid and reliable measures of each dimension. An initial pool of 150 face valid items was submitted to a series of exploratory factor analyses (ML with Varimax rotation). In the final solution, each item boasted a loading of at least .6 on its intended factor and significantly lower crossloadings. Corrected item-total correlation estimates were all greater than .6 and the Cronbach's Alpha estimate for each composite scale was greater than .9. Subsequent confirmatory factor analysis results led to the elimination of an additional number of items. Final model fit was deemed acceptable.

Leaving Flatland: Tests for Violations of Euclidean Assumptions in MDS

Jesse Spencer-Smith

Most applications of multidimensional scaling (MDS) assume that proximities are related to distances in a flat geometry (e.g. Euclidean or city block). While some variations of MDS allow scaling on sphere or in constant-curvature space, little use is made of these applications. I discuss three tests for uncovering curvature in psychological proximity data prior to scaling. In the first test, data are assumed to lie on a ratio scale, and can be applied to any such data. In the second test, data are assumed to lie on a ratio scale and it is assumed that some points are collinear with and between pairs of points. In the third test, data are assumed to lie on an ordinal scale, with additional assumptions concerning the configuration of points.

Sequencing a Computerized Adaptive Test Battery

Roberto Ricci and Wim J. van der Linden

The problem of sequencing a battery of computerized adaptive unidimensional tests is addressed. Making a test battery adaptive involves two different kinds of adaptation. Firstly, the choice of the each next test in the battery is adapted to the performances of the test taker on the earlier tests. Secondly, once a test is chosen, the items in it are selected adaptively with respect to the updates of the test takers ability estimate. This paper focuses on the first problem. We use a Bayesian procedure with an empirical prior conditional on the test taker's score on one or more background variables. In essence, the problem is an empirical Bayes problem with independent unidimensional likelihood functions of the ability parameters but a dependent joint prior distribution. Because of the prior distribution, the posterior distributions of the ability parameter do not factorize. Using extensive simulation studies the practical features of the procedure are explored.

Eligibility functions for exposure control in Computerized Adaptive Testing

Bernard P. Veldkamp

Over exposure and under exposure of items in the bank are serious problems in operational computerized adaptive testing (CAT) systems. These exposure problems might result in item compromize, or point at a waste of investments. The eligibility function method is developed to deal with both types over exposure problems. In this method, exposure control parameters based on observed exposure rates are implemented as weights for the information in the item selection procedure. The method does not needs time consuming simulation studies, and it can be implemented conditional on ability level. The method is compared with Sympson Hetter method for exposure control and with alpha-stratified testing. The results show that the method is succesfull in dealing with both kinds of exposure problems, at the costs of an increased RMSE, while the bias remains comparable for all methods.

Poor Man's CAT

Gunter Maris

In a multistage test students are administered different blocks of items depending on the responses to earlier blocks of items. It has often been observed that the method of conditional maximum likelihood is not applicable with data from a multistage test. It will be shown that, using the correct conditional likelihood, this is not true. Attention will be given to the assessment of model fit and results will be illustrated with real and simulated data.

Controlling Item Exposure and Test Overlap on the Fly in Computerized Adaptive Testing

Shu-Ying Chen

Even though traditional exposure control methods (e.g., Sympson & Hetter, 1985; Chen & Lei, in press) can control item exposure well such that items are administered with item exposure rates (and test overlap rate) less than a pre-specified value, the iterative simulations required to find stabilized item exposure parameters prior to operational CATs for these methods are very time consuming. Furthermore, the tedious iterative simulations need to be re-conducted whenever there are changes in CAT settings or the examinee population of interest. The purpose of this study is to propose a new item exposure control method, which is designed to control item exposure rates and test overlap rate simultaneously on the fly without conducting any time consuming iterative simulations prior to operational CATs.
A simulation study will be conducted to investigate the effect of the modified procedure on item exposure control and precision of trait estimation where the maximum item exposure rate will be set at 0.2 and test overlap rate at 0.15. Evaluation criteria will include root mean squared error, average bias, maximum item exposure rate in the item pool, test overlap rate, and number of items not used.

Evaluation Of Bayesian Model Selection Procedures In The Context Of Inequality Constrained Contingency Tables

Olav Laudy and Herbert Hoijtink

In recent years, much attention is devoted to inequality constrained modelling. However, model selection procedures in this area are not common. The reason is that many model selection procedures require a function of the number of parameters as penalty for model complexity, while in inequality constrained models the number of parameters is not known. We present a frequency evaluation of the available fit measures for inequality constrained contingency tables. Model selection is performed using either posterior predictive inference (L-criterion, DIC and posterior predictive checks) or prior predictive inference (posterior model probabilities). Three simulations are performed. Firstly, it is investigated how the fit measures perform in the presence of more models that partially describe the data. The second simulation investigates the effect of the sample size if there are two competing models. The third simulation investigates the effect size. The results show that the posterior model probability has the most stable performance. The DIC can perform better, but can also perform worse, specially when an unconstrained model is in the model set. The performance of the L-criterion and the posterior predictive checks are similar: occasionally, they show a good performance, but in general, they are less able to select the correct model.

A New Distribution-Free Hierarchical Bayes Approach to Logit Regression Models

Daniel M. Rice

In psychometrics, logit models such as conditional logit, multinomial logit, and Cox proportional hazards survival models are traditionally solved within a standard Maximum Likelihood framework. For these types of problems, it has been widely communicated that the standard Maximum Likelihood solutions are equivalent to a standard Maximum Entropy approach. These solutions are generally not in a form that can be related to Bayes Rule.

In this new Bayesian Entropy formulation, all response and non-response outcomes are modeled simultaneously as linear count constraints with the critical caveat that prior weights have the properties of prior probabilities. Solutions are regularized by the standard errors of the measures. Solutions are now in a form that is a direct manifestation of Bayes Rule. With small samples, solutions appear flatter and less subject to overfitting than standard Maximum Likelihood solutions.

Because of the linkage to Bayes Rule, a Hierarchical Bayes formulation based upon the Kullback-Leibler norm results. Posterior probabilities from a higher level serve as priors for a lower level. There is no assumption of a known parametric prior distribution. Rapid gradient descent computation is now tractable in SAS with the Beta version of Proc Entropy. Case studies point to advantages with small samples and large numbers of variables.

Bayesian Factor Analytic Analysis of Behavioral Genetic Data

Kazuo Shigemasu

A structural equation model of behavioral genetic data is analysed using Bayesian methodology. Multivariate data of MZ twin and DZ twin is modeled as a linear combination of additive genetic factors, a common (shared) environmental factor, and a non-shared (unique factor). Behavioral genetic model appears to be complicated but reorganization of the model in view of the hierarchical modeling makes the model building and statistical inference easier. Bayesian analysis using MCMC method provides a proper inference for the parameters involved in the hierarchical model. Besides, the distribution of the genetic coefficient which is a complex function of parameters can be obtained easily by Bayesian numerical method.

Bayesian MCMC approach for models with many random effects

Eisuke Segawa

Models that generalize multilevel, longitudinal item response, and structural equation models have been proposed (Muthen and Muthen, 2004; Skrondal and Rabe-Hesketh, 2004) and implemented in software such as Mplus 3 and GLLAMM. Numerical quadrature is used to integrate out random effects and obtain marginal maximum likelihoods in the software. When the number of random effects is large, e.g., more than 5, integration with quadrature is computationally demanding and prohibitive. Although adaptive quadrature, which is available in this software, eases the problem, the approach is computationally prohibitive for models with a very large number, e.g., more than 10, of random effects. Bayesian estimation analysis with Markov Chain Monte Carlo (MCMC) does not have such computational difficulty for high-dimensional random effect models. Its computational advantage is shown through two models of actual problems, the first is a growth model of three latent variables, with each latent variable measured by several ordinal items. The second is a 2nd order confirmatory factor model for ordinal response variables.

Performance of the Four-Parameter Extended Beta-Binomial Models In Scaling and Equating Composite of Test Scores

Miao-Hsiang Lin

Two versions of the four-parameter beta-binomial model are examined with respect to the adequacy in scaling complicated data distributions. The two versions, differing in the use of either the binomial or the compound-binomial conditional distribution, regard the upper limit of the beta as an additional parameter. The method-of-moments fitting procedure is employed to real data sets in which the average of students' ability varies with schools' levels. Initial findings clearly indicate that the fourth parameter of the proposed models gives better fits than the two- or the three-parameter beta-binomial models. The four-parameter of compound-binomial model relative to the binomial counterpart, however, does not provide considerable improvements in fit. Factors such as the sample size and the variability of item difficulty are to be further examined with respect to gains in the data-smoothing purpose. In addition, performance with respect to scaling and equating composites will be examined as well.

Equating and Scaling with Matched Samples

Anli Lin, Don Meagher, and Christina Stellato

When we want to equate two or more tests, we often use the design with common item or common person. In a common person design, we need extra effort to collect data, in which each person needs to do test twice or more with balancing order. In a common item design, sometimes there is a big difference in demographic information between two test takers (e.g., race, education, etc.). This will effect on results of equation or scaling. If we could match samples from two tests by some demographic variables, e.g., race, education, age, and sex, we will remove the influence of difference in equating or scaling from demographic variables. Also, each person does test only once.
This paper presents a research to do equating and scaling using matched samples. The purpose of this research is to verify the link constant (LC), which is used in linking paper – pencil test (PPT) and computer based test (CBT) for Miller Analogies Test (MAT). The results from unmatched and matched samples are presented. The LC from matched samples is more closed to LC from previous study with common person design.
The results from research show us that matched samples improved performance equating or scaling.

Bayesian IRT Equating With Correction For Unmotivated Respondents On The Anchor-Test

Anton A Béguin

Two forms of a test with no items in common and administered to different populations can only be equated using additional data. For example by using an anchor test. It is necessary that the behavior on the additional test is similar to the actual test performance. In practice this last condition can be violated if the actual tests are high-stakes tests while the additional tests are not. In that case some of the respondents could behave differently on the additional tests, due to a lack of motivation. This could lead to lack of fit of the model and incorrectly equated tests.
In this paper a unidimensional IRT model is used to describe the performance on the actual forms of the test. The respondents on the anchor test are divided into two groups: those who give responses in accordance with the model on the actual forms, and those who give responses that must be modeled with different IRT parameters. The split of respondents in groups and the parameters describing the results for both groups are estimated using a Markov Chain Monte Carlo estimation procedure. The identification of the model, alternative ways of modeling and the performance of the model are discussed.

Observed Score Equating Using Empirical Test Characteristic Curves

Guido del Pino, Ricardo Aravena, Rianne Janssen, and Jorge Manzi

A new method for equating different test forms under a common-item nonequivalent groups design is proposed. The method is comparable to IRT true score equating (Lord & Wingersky, 1984; Kolen & Brennan, 2004), but instead of deriving the analytical test characteristic curve (TCC) for each test form, an empirical TCC is estimated. The empirical TCC gives the mean score of the students on the test form conditional on their specific score on the common items. Smoothing of the empirical TCC is done using a piecewise linear model. Test score x on form X is equated with test score y on form Y iff x and y are the conditional means for the same specific score on the common items.
Like the equipercentile equating methods, the new equating method allows for controlling for the whole distribution of scores and not just for the mean and variance, as in the linear methods. However, the new method is not influenced by the distribution of ability of the students, as the equipercentile method is. The method is applied to a national university entrance examination on science in Chile. The results are contrasted with those obtained with IRT and with other equating approaches.

Symposia and sessions, 11.50h-13.30h

Shape Constrained Smoothing with Asymmetric Penalties

Paul Eilers

Quadratic difference penalties are powerful tools to enforce smoothness when estimating curves with B-splines. Sometimes we want more than smoothness alone: a monotone increase, or a convex shape, or even a combination of the two. This can be achieved easily with additional asymmetric difference penalties. The individual penalty terms are weighted differently, depending on the signs of (higher order) differences. This approach results in a simple and robust algorithm: iteratively re-weighted smoothing. I will present several applications. In the context of distribution modeling this approach leads to a new way to model mixtures of unimodal densities.

Estimating The Smoothing Parameters Of Penalized Likelihoods

Simon N. Wood

Many statistical models are estimated by maximization of a penalized likelihood or quasi-likelihood $l(\beta)-\frac{1}{2}\sum_i \lambda_i \beta^T S_i \beta$ where the β is a vector of unknown coefficients, S_i a matrix of known coefficients and λ_i is an unknown smoothing parameter. Given smoothing parameters, the estimation of β is usually straightforward, but computationally efficient and reliable estimation of the smoothing parameters is more problematic. I will discuss computational methods for this problem for the case in which the penalized likelihood can be maximized by an iteratively reweighted least squares algorithm, and GCV or Mallows Cp are appropriate smoothing parameter estimation criteria.
The methods are particularly useful for the estimation of Generalized Additive Models and mildly non-linear models formulated in terms of unknown smooth functions.

Using the Data to Regularize the Model

Jim Ramsay

We want to fit data **Y** by a model $\mu(\theta)$, where the model depends on model parameters θ. We have a fitting criterion $F[\mathbf{Y},\mu(\theta)]$, and we also have a penalty functional $P[\mu(\theta),\beta]$ that measures some kind of undesirable variation or "roughness" in the model, and that is defined by penalty parameters β. We estimate θ by minimizing the composite criterion $F[Y,\mu(\theta)] + \lambda P[\mu(\theta),\beta]$, where λ controls the amount of emphasis on the "roughness" in the model as defined by $P[\mu(\theta),\beta]$.
But suppose we want to estimate β? Then the roles of the two criteria are reversed, and, in effect, the fitting criterion F regularizes the model through the penalty P. The model is "data-regularized" rather than "roughness regularized".
Examples important for applications are offered that are drawn from multivariate statistics and from functional data analysis. This approach is used to estimate differential equation models from noisy observations of a dynamic process.

Applications Of Cognitive Diagnosis Modeling: From Simple To Complex Data

Jimmy de la Torre

Although unidimensional item response models have been useful in scaling and ordering students on a latent proficiency continuum, they rarely provide enough information to permit more detailed and meaningful comparisons between students. A promising alternative can be found in latent variable models for cognitive diagnosis which are developed precisely for diagnosing the presence or absence of multiple fine-grained attributes. The importance of cognitive diagnosis modeling is increasingly recognized in the field of testing and measurement showing growing awareness of the benefits of this approach. One particular cognitive diagnosis model that has been shown to provide good model-data fit despite its simple formulation and interpretation is the higher-order deterministic inputs, noisy "and" gate (HO-DINA) model. The HO-DINA model and its subsequent reformulations are attempts to address varying data complexity while preserving the intuitive psychological interpretation of the model. The formulations of the model have included a higher-order latent trait structure that simplifies the joint distribution of the attributes, and a mechanism for generating the latent responses that accounts for the possibility of multiple strategies. All these are done under the simplest sampling design involving dichotomous responses only. Several layers of complexity accrue when more intricate sampling designs and mixed test formats are involved. Such is the case with the NAEP data. Further reformulations of the HO-DINA model are proposed to analyze data consist of answers to multiple-choice and constructed response items collected from clusters of students with differential weightings.

Development and Individual Differences in Transitive Reasoning: a Fuzzy Trace Theory Approach

Samantha Bouwmeester

According to Piaget's theory an understanding of logical rules is a prerequisite for drawing transitive inferences. Information rocessing theorists, however, showed that no logical rules have to be used to infer a transitive relationship. This theory assumes that memory for the premises is the only ability involved when drawing transitive inferences. Fuzzy trace theory (Brainerd & Kingma, 1984), however, assumes that nor logical rules nor memory for the premises are needed to draw transitive inferences. According to fuzzy trace theory performance on transitive reasoning tasks can be explained by fuzzy trace ability, which is the ability to process and retrieve patternlike, reduced information on different levels of abstraction, simultaneously. In the context of transitive reasoning this fuzzy trace may be, for example, "objects get longer from right to left". Individual differences in transitive reasoning were investigated in 4 to 13 year-old children. The performance on three kinds of tasks which differed with respect to their presentation ordering and position ordering was studied in an effort to determine the use of fuzzy trace theory as a framework for explaining the development of transitive reasoning. The data were analysed using a multi-level latent class model. The results from a sample of 409 children ranging in age from 64 to 159 months showed that the two-dimensional classification of performance patterns agreed with the expected performance groups according to fuzzy trace theory. Task format had a stronger effect on performance on transitivity test-pairs than on memory test-pairs. Furthermore, the developmental effects showed more improvement in fuzzy ability than in verbatim ability.

Statistical inference for ACT-R models of cognitive task performance

Rhiannon L. Weaver

ACT-R computational models of cognition are used increasingly to inform the design of educational technology, providing interpretable models of the thought processes that generate an individual's observed response to a cognitive task. Currently, competing cognitive models are evaluated based on least squares or correlation metrics comparing low-order moments of aggregated data to moments calculated from model simulations (Schunn and Wallach, 2001). Parameters are often set {\em a priori} based on previous studies, or estimated using techniques such as iterated gradient descent based upon the evaluation metric. While this predictive approach can be useful for average population simulation, results are difficult to interpret when applying ACT-R as a direct explanation for observed behavior in sub-populations and individuals. For instance, evaluation metrics like correlation can be high even when the observed data has probability 0 under the assumed model. In this presentation I develop a Bayesian approach to parameter estimation for a constrained class of ACT-R models, using Markov Chain Monte Carlo methods to obtain posterior distributions for parameters. This likelihood-based approach allows for the use of Bayes Factors for model selection. I apply the new methodology to model selection using ACT-R models for scatterplot generation (Baker, Corbett and Koedinger, 2003) and to an estimation experiment using an ACT-R model for list memory (Anderson, Bothell, Lebiere, and Matessa, 1998).

References:

Anderson, J., Bothell, D., Lebiere, C. and Matessa, M. (1998) List memory. Chapter 7 in Anderson, J. and Lebiere, C. (1998). *The Atomic Components of Thought*. Lawrence Erlbaum Associates.

Baker, R., Corbett, A., Koedinger, K. (2003). Statistical techniques for comparing ACT-R models of cognitive performance. Proceedings of the 10th Annual ACT-R Workshop, pp. 129-134.

Schunn, C. and Wallach, D. (2001). Evaluating goodness of fit in comparison of models to data. Online manuscript available at http://www.lrdc.pitt.edu/schunn/gof/index.html

Posterior Predictive Model Checking For Cognitively-Based Psychometric Models

Roy Levy

Bayesian inference and cognitively-based statistical models are two emerging and converging lines of research in psychometrics. Model criticism remains a relatively unexplored aspect of Bayesian approaches to psychometric modeling. This presentation will discuss posterior predictive model checking, a powerful approach to model criticism, with an emphasis on cognitively-based models. Examples will be drawn from both methodological and applied work. Finally, I will address the role of posterior predictive model checking in the larger scope of the design, development, and use of assessment systems that are grounded in the substantive analysis of the domain.

New Approaches For Analyzing Multitrait-Multimethod Data

Michael Eid

This symposium presents some new and seldomly applied approaches for analyzing multitrait-mulitmethod data. Three mode models, multilevel models and structural equation models for structurally different and interchangeable models will be presented. The theoretical foundation of each model will be explained. Moreover, each model will be illustrated by the same data set making it possible to compare the different approaches not only theoretically but also with respect to the story they tell the applied researcher. This talk gives a short introduction into the analysis of MTMM data and presents the data used in all three applications. The data concerns the measurement of three traits with four items and three raters (one self-rating, two peer ratings).
After the three presentations the differences between the three approaches will be summarized.

Three-Mode Models For Multitrait-Multimethod Data

Frans J. Oort

Multitrait-multimethod data are characterised by three modes: traits, methods, and subjects. Considering subjects as random, and traits and methods as fixed, stochastic three-mode models can be used to analyse the multitrait-multimethod covariance structure. Stochastic three-mode models can be written as linear latent variable models with direct product restrictions on the parameter matrices (Oort, 1999), yielding three-mode factor models (Bentler & Lee, 1979) and composite direct product models (Browne, 1984) as special cases.
Using example data of three personality traits (measured with four items each), three methods (self-report and two peer-reports), and 482 students (Eid et al., 2001), it is shown that direct product restrictions on factor loadings and factor correlations facilitate interpretation of the results and enable easy evaluation of the validity requirements of multitrait-multimethod correlations (Campbell & Fiske, 1959). Additional direct product restrictions on other model parameters allow further distinction between trait (co)variance, method (co)variance, item specific (co)variance, and error variance.
Fitting a series of stochastic three-mode models to the present data shows trouble-free convergence of parameter estimation, attractive interpretation, fulfilment of validity requirements, substantial trait, item specific, and error variance, but almost no method variance.

Analyzing Complex Multitrait-Multimethod Data With Multilevel Models

Gerty Lensvelt-Mulders, Joop Hox, and Cora Maas

The classical multitrait-multimethod matrix can be viewed as a two-dimensional cross-classification of traits and methods. Several approaches have been proposed to analyze such data, notably structural equation models, generalizability theory, and even facet analysis. Since recent multilevel modeling software can also analyze cross-classified structures, multilevel analysis can be added to this list. If the focus is on analyzing classical MTMM data, the multilevel approach is not really attractive, because the resulting model is equivalent to a confirmatory factor model with additional restrictions imposed by the multilevel parameterization. However, it the data contain further complexities, such as additional information on the traits or persons, or if interactions are assumed, multilevel analysis of MTMM data offers a flexible analysis tool. In this paper, we will show how multilevel analysis applies to classical MTMM data, using a small data example data set. Next, we show how the multilevel model can be extended to analyze the German Big 5 MTMM data, which have a more complicated structure because additional information is available and the data were collected using triplets of raters.

Analyzing Multitrait-Multimethod Data Having A Nested Method Structure

Fridtjof W. Nussbeck, Michael Eid, and Tanja Lischetzke

The multitrait-multimethod matrix (Campbell & Fiske, 1959) is widely seen as the via regia to determine convergent and discriminant validity of psychological scales and questionnaires. The practice of analyzing MTMM data usually does not take structural differences between methods into account. Method-specific influences are considered as random fluctuations around an underlying trait. In this paper it is argued that different types of methods imply different types of method factors. Particularly, it will be shown how complex MTMM data consisting of structurally different and interchangeable raters can be analyzed. The empirical application consists of one self-report and two interchangeable peer reports which are nested within the self-report and thus structurally different. The model allows to determine consistency and method specificity coefficients. Moreover, it will be demonstrated how external variables can be will be incorporated.

Introduction to statistical models for social network analysis

Marijtje A.J. van Duijn

An overview of research questions common in social network analysis and the statistical models available to answer these questions will be given. The overview will serve as an introduction to the other presentations of the session.

Special attention is given to the formulation of statistical models as random effects (or multilevel) models, of which the social relations model, for social relations measured at interval level, is an important example. The social relations model can be extended to a longitudinal SRM for the analysis of multiple measurements of a social relation.

Accuracy and Efficiency in Estimation of the p_2 Model

Bonne J.H. Zijlstra and Marijtje A.J. van Duijn

The p_2 model is a model for the analysis of social network data. The dependent variable is a (binary) network, or directed graph, on a given set of actors, or nodes. Differences between actors in attractiveness and productivity (outgoingness) are modeled using random effects. The model also includes parameters for density and reciprocity. Explanatory variables on the levels of the actor and the ordered pair of actors can be used to further explain these parameters.
Estimation of the p_2 model is not straightforward. Quasi likelihood procedures like IGLS are fast, but produce biased estimates. MCMC algorithms are slow, but give nearly unbiased estimates. Within the MCMC methods, hybrid Metropolis-Hastings algorithms that use independent or random walk proposals are compared. Usually, independence chain algorithms are much more efficient than random walk algorithms. Because of the correlated random effects in the p_2 model, the advantage of the of the independence chain algorithm over the random walk algorithm is strongly reduced.

Random Effects Models For Digraph Panel Data

Michael Schweinberger & Tom A.B. Snijders

We consider the modeling of digraph panel data, generated by some latent Markov process evolving in continuous time. The infinitesimal generator of the Markov process is parameterized in terms of random utility modeling, meaning that the digraph changes due to nodes ("actors") changing the arcs to other nodes. The choices made by nodes are based on "regression" functions, depending on the future digraph and covariates linked by multinomial logit link functions to the probability of changing some arc.An important but not very credible homogeneity assumption underlying the model is that the parameter values are constant across nodes. The present paper proposes to substitute it by the assumption that the values are drawn from some underlying probability distribution. Maximum likelihood and Bayesian methods are considered to estimate parameters, implemented with MCMC-based data-augmentation. In view of the high computational costs induced by random effects, one-sided score tests are used to decide whether or not to include them.Two well-known and appealing features of the score test are that the (unrestricted) random effects model need not be estimated and that the mixing distribution of the random effects need not be specified. The usefulness of the methods will be demonstrated with empirical data.

Bayesian Analysis of Longitudinal Social Network Data

Johan H. Koskinen

The dynamics of social networks may be studied by investigating change in longitudinal data. Since typically, we are not in a position to observe the complete sequence of changes between observations, and it is reasonable to describe evolution as sequences of incremental changes, continuous-time Markov chains have been suggested for modelling longitudinal network data. Defining such a model by focusing on the embedded chain with holding times and transition probabilities, we are provided with a parametric framework that meaningfully describes the actual changes in the network. Another benefit is that it allows us to write down explicitly the likelihood of a complete sequence of changes. Since the complete sequence of changes between observation can be treated as a latent variable, observed data can for purposes of estimation be augmented by the latent changes and incorporated in a reversible jump MCMC sampling scheme for exploring the posterior distribution of the parameters. Apart from the benefits of a fully Bayesian analysis in terms of the precise measure of uncertainty regarding the parameters given by the posterior distribution, the proposed model formulation lends us flexibility in defining mechanisms of change as well as allowing for more general types of network data than binary.

Missing Data In Networks: Longitudinal Data Analysis With Missing Observations

Mark Huisman

Statistical modeling of longitudinal data on complete social networks isoften hampered by missing observations on ties in the network. This may bedue to attrition, when respondents drop out of the survey, compositionchange, when actors leave the network, or nonresponse, when respondents donot complete a sociometric questionnaire. This paper examines the problemscaused by missing observations on ties, and describes methods to analyzeincomplete longitudinal network data with stochastic actor-oriented modelsfor network change. Data on friendship networks among college freshmen areused to illustrate the procedures.

On Measuring Classification Uncertainty For Model-Based Clustering Procedures

José G. Dias and Jeroen K. Vermunt

The paper discusses classification uncertainty for finite mixture models. It discusses the stability of the latent class model as a classification technique. It is shown that classification based on the optimal Bayes rule is stable and difficult to improve by aggregate classifiers, except for sparse data. Additionally, we show that care has to be taken when the bootstrap method is used to answer this question; mainly, the label-switching effect has to be carefully addressed in the bootstrap approach, something that has received no attention. Simulated and empirical data are used to illustrate the approaches put forward in this paper.

The Confidence Area Of Correspondence Analysis

Akihiro Saito and Hideki Toyoda

Correspondence analysis is used especially frequently in Japan, France, and the United States. It is used by various scenes of the psychology research, for instance, to analyze subjects' attribute and his/her response to an item. Usually, the main purpose of this analysis method is to give scores(it is used as coordinates) to each category of a row and a column of a contingency table, to put them to a coordinate system all together and to analyze relationship graphically.
In a variety of multivariate analysis methods, we confirm the confidence interval of the estimate, but when we use correspondence analysis, we don't confirm the confidence interval(It can be called "the confidence area" because we use two-dimensions).
The purpose of this study is (1) to evaluate the confidence area of coordinates of correspondence analysis, it is evaluated using large-sized data of POS(Point of Sale) and the resampling method, (2) to confirm the stability of the correspondence analysis and (3) to investigate the condition of change in the confidence area.

Taxicab Correspondence Analysis

V. Choulakian

Taxicab correspondence analysis is based on the taxicab singular value decomposition of a contingency table, and it shares some similar properties with correspondence analysis. It is more robust than the ordinary correspondence analysis, and on some data sets taxicab correspondence analysis produces more interpretable results than the ordinary correspondence analysis. Examples are provided.

Individual Differences Additive Tree Fitting Through (Heuristic) Iterative Projection

Hans-Friedrich Koehn

A new method for the discrete nonspatial representation ofindividual differences through fitting additive tree structures tothree-way two-mode proximity data is proposed. Following therationale that different individuals base their judgments on thesame family of trees with identical topological structure,individual variation is modelled through different branch lengths.However, distinct from existing implementations, such as INDTREES, neither gradient-based optimization, nor a penalty function toaccount for violations of the four-point condition by theestimated distances is utilized for minimizing the least squaresloss function. Instead, the method presented relies on iterativeprojection onto closed convex sets defined by the constraints ofthe four-point condition underlying the additive tree model.First, a ''communal" tree is identified through application ofiterative projection as a heuristic search strategy for the beststructural representation of an aggregated proximity data.Individual tree structures, restricted to the identical topologyas the communal tree, are then constructed for the individual datamatrices through iterative projection with constraints defined bythe communal tree structure. An application to judgments ofschematic face stimuli illustrates the new method.

Nonmetric PCA of Likert Items And Causes Of The Horseshoe

Takashi Murakami

The Horseshoe is a quadratic relationship between object scores given by procedures belonging to the family of Correspondence Analysis, which are usually corresponding to the first and the second largest singular values. Although the algebraic reason why the horseshoe appears has been clarified in the literature, it is also connected with the extreme response tendency of respondents. We made a simple numerical experiment to demonstrate that there were the two types of horseshoes with different causes.
Two sets of artificial data consisting of Likert items were generated; the first one was given by uniformly discretizing multivariate normal variables loading on two common factors, and the second was generated by the same way except for using several disretizing functions that represent the individual differences of extreme response tendency. They were analyzed by Nonmetric Principal Components Analysis (PCA) with two quantifications for each variable. While we obtained a couple of horseshoes associated with interpretable components respectively from the first data set, only one horseshoe with a somewhat blurred form was given from the second one.
The analysis of real data sets and their interpretations in the context of our research will be discussed.

MeanDiff: A Program For The Uni- And Multivariate Analysis Of Unconditional, Conditional And Average Mean Differences Between Groups

Ivailo Partchev

MeanDiffs is a PC-program for the uni- and multivariate analysis of mean differences between groups on outcome (or response) variables in designs withand without covariates. It is a convenient tool for the analysis of pretest-treatment-posttest designs with two or more treatment groups, one of which can be a control group. MeanDiffs is considerably more general than comparable programs because it:* does not assume homogeneity of variances (univariate case with a single outcome variable) or covariance matrices (multivariate case with two or more outcome variables) of the outcome variables between groups;* allows analyzing mean differences between groups with respect to several manifest outcome variables or with respect to a latent outcome variable; * allows analyzing conditional and average effects with respect to several manifest covariates or with respect to a latent covariate; * estimates and tests average effects for non-orthogonal analysis of variance designs correctly, provided that the covariates are specified as qualitative indicator variables; * produces results which are easily interpretable in the analysis of conditional and of average effects (mean differences) between groups; and * estimates and tests the average causal effect, provided that the covariates fulfil certain assumptions.

Testing The Average Effect When There Is An Interaction Between A Latent Covariate And The Treatment Variable

Felix Flory, and Rolf Steyer

We outline a procedure to test the hypothesis that the average treatment effect is zero, given an interaction of the treatment with a latent covariate. The latent covariate can only be measured by (two or more) fallible tests, i.e., the test scores consist of true score and measurement error. This combines two approaches: the analysis of latent moderators in structural equation models (e.g., Klein & Moosbrugger, 2000) with the analysis of average causal effects in models with interactions between treatment and manifest moderator variables. Structural equation models with a latent moderator variable provide the estimation of the effect function, indicating the treatment effect given a value of the latent moderator. It is shown how to test the average causal effect (i.e., the main effect of the treatment) with LISREL or Mplus. It is also shown that replacing the latent variable by a manifest proxi (such as a sum score) does, in general, not yield a valid test of the average effect. Finally, the procedure is illustrated by an empirical example from evaluation research.

Understanding Measurement Bias in the Common Factor Model: The Role of Stereotype Threat

Jelte M. Wicherts

Measurement invariance (i.e., absence of measurement bias) with respect to biographical groups is important for the valid measurement of cognitive abilities and aptitudes. Various methods have been developed to detect measurement bias. However, the use of such bias detection methods is rarely accompanied by theoretical expectations regarding why and how measurement bias occurs. Understanding the sources of measurement bias can increase the chances of measurement bias being detected and can aid in more efficient test development. The aim of this talk is to relate the effects of stereotype threat on test performance to measurement invariance.

Stereotype Threat (ST) is defined as a situational pressure on a test-taker arising from stereotypes concerning the academic proficiency of his or her social group. Recent developments within experimental social psychology indicate that ST may negatively affect test performance of members of stigmatized groups. For instance, implicitly reminding female test-takers of the stereotype that women are bad at math may lower their consecutive math performance. In this talk, ST theory is related to measurement invariance as defined within the common factor model (i.e., factorial invariance). The modeling approach is described and applied to data of an ST experiment involving female math performance.

Adaptive Robustified Estimation in Gaussian MAG models

Masashi Miyamura and Yutaka Kano

Using a graphical Markov model, it is possible to describe relationship among variables precisely. Confounding bias or selection bias in estimation could, however, occur if unobserved variables would exist that relate to the variables in the model. Recently some researchers have discussed graphical models with unobserved variables and proposed maximum ancestral graphs (MAG) which consists of nodes and three types of edges, namely undirected, directed and bi-directed. MAG describes not only conditional independence but also marginal independence among observed variables, where unobserved confounding and selection variables are taken into account. Maximum likelihood (ML) estimation has been developed for the Gaussian MAG. Although the ML estimation is a standard technique in statistics, it is sensitive to outlying observations. In this talk, we propose an adaptive robustified estimation for the Gaussian MAG, where each observation is weighted with its own likelihood and robustness tuning parameters are adaptively chosen.
Simulation results and real examples are presented. It is shown that this adaptive selection of parameters works nicely to give robust estimators with smaller mean squared errors.

A Latent Class Approach To Deal With Non-Compliance In Experimental Research

Marieke Spreeuwenberg

In experimental settings where subjects are assigned to either a control group or a treatment group, participants do not always follow their treatment requirements. For example, participants do not show up in the intervention or control participants follow the treatment outside the experimental set-up. As a consequence of non-compliance estimates of the treatment effect could be biased. In most medical research “Intention to treat analysis” (ITT) and “As treated analysis” (AT) are the most frequently used statistical methods to deal with non-compliance. It seems, however that in the view of causality these methods are not always valid.
In this presentation, I will give an overview of alternatives to ITT and AT. A latent class approach for dealing with non-compliance will be proposed. Thereby it is shown that this model provides a basis for latent class models with indicator variables for compliance. Compliance indicator variables are for example measures of the pill counts or responses on a Compliance Questionnaire. Results from simulation studies will be presented and an empirical example will be discussed.

Presidential Address, 14.50-15.50

Fitting Psychometric Models using Gradient Methods and Automatic Derivatives

Robert Cudeck

A major activity of quantitative scientists is developing and applying mathematical models to interesting data. In practice this means that a lot of time is invested in writing computer programs to fit models. A number of optimization algorithms are used, including the EM algorithm (in several variants), Markov Chain Monte Carlo (with related methods), derivative-free function minimizers and searchers, and gradient methods such as steepest descent, quasi-Newton and Gauss-Newton. For quickly moving from problem to solution, one of the most efficient general classes of methods is a gradient-based approach to maximum likelihood estimation that exploits automatic differentiation. With automatic differentiation, numerical evaluations of the gradient vector and Hessian matrix are obtained easily as the log-likelihood function is computed. The procedure works by applying the chain rule for composite functions at each step of the function calculation. It can be done quite accurately . Equally appealing, it can be done mechanically. The derivatives can be used in a Newton-Raphson iteration. Approximate standard errors of parameter estimates are available from the final Hessian obtained at the solution. For many problems, this method is competitive compared to alternative estimation strategies. After surveying the basic ideas of automatic differentiation, several examples will be reviewed to illustrate the approach.

FRIDAY JULY 8

Symposia and sessions, 9.00h-10.40h

Permutation Tests in PCA for Assessing Significance of Shared Understanding

Patrick J.F. Groenen and Marco de Haas

In various areas of the social sciences it is important to assess whether or not a group of people share the same believes. A particular example concerns the believes of employees and management of a company. To see if the shared believes could be made more homogenous, De Haas and Algera (2001) studied the effect of an intervention program aimed at increasing the shared believes of the employees and management. Based on a PCA solution, they introduced a measure for quantifying the amount of homogeneity within a group. The results suggested that the homogeneity within the group had increased by the intervention.
However, as the group of people studied was rather small (n = 14), it is not clear whether the difference in homogeneity between the pre-test and post-test is significant. Therefore, the question remains whether a difference should be attributed to random fluctuations in the sampling process or whether the difference reflects a true difference. For the measure of homogeneity there is not a simple statistical test that can be applied. To solve this problem, we propose to use permutation testing. In addition, we propose two alternative measures of homogeneity all based on the PCA solution and see how well they perform in detecting a true difference.

Visualizing Dependence of Bootstrap Confidence Intervals for Methods Yielding Spatial Configurations

Henk A.L. Kiers, Patrick J.F. Groenen

Several techniques exist for summarizing data by means of a graphical configuration of points in a low-dimensional space. The most common examples are multidimensional scaling (MDS) and principal component analysis (PCA). Usually, such analyses are applied to data for a sample drawn from a population. To assess how accurate the sample based plot is as a representation for the population, confidence intervals or ellipsoids can be constructed around each plotted point (representing, say, a variable). For this purpose, it has been proposed to use a bootstrap procedure. This procedure gives a full configuration for each bootstrap sample, so we end up with a large number of configurations that jointly display variation of the configuration of the variables upon resampling. Usually, the variation is displayed by considering different variables separately. However, such a procedure ignores the dependence of variation of *different* variables across bootstrap samples. To display how the variations of different variables depend on each other, we propose to visualize bootstrap configurations in a temporally smooth way (movie). Problems with this approach, and some solutions, will be discussed.

Bootstrap Estimates Of The Parameter Uncertainty In HICLAS Analysis

Eva Ceulemans, Peter Beeusaert, and Iven Van Mechelen

The hierarchical classes (HICLAS) model, proposed by De Boeck and Rosenberg (1988), is an order-preserving Boolean decomposition model for two-way two-mode binary data. In particular, HICLAS reduces the two modes of the data to a few binary components. Furthermore, HICLAS represents the quasi-order relations that can be defined among the elements of a mode, yielding hierarchically organized classifications of the two modes. In practice, HICLAS analysis is often applied to data that are obtained from a sample from a larger population, assuming that the results for the sample may be generalized to the population from which the sample was drawn. However, HICLAS analysis does not provide any estimates of parameter uncertainty due to sampling fluctuations. In this paper, it will be discussed how such estimates can be obtained by means of bootstrap analysis. The proposed bootstrap method is innovative, because we focus on the dependency between parameters instead of considering the different parameters separately. Therefore, this method is potentially relevant for a broad class of classification methods.

Stability Of Nonlinear PCA In Comparison To Linear PCA

M. Linting, P. J. F. Groenen, and J.J. Meulman

Nonlinear principal components analysis, performed by the program CATPCA from the SPSS Categories module, is a generalization of principal components analysis (PCA). Like linear PCA, nonlinear PCA reduces the variables in a data set to a (small) number of principal components that represent the information in the variables as closely as possible. However, in contrast to linear PCA, CATPCA can handle variables of different analysis levels (nominal, ordinal, and numerical) simultaneously, and can deal with nonlinear relationships between variables.
Although (CAT)PCA can be seen as an exploratory technique, it need not be deprived of confirmatory diagnostics, like confidence regions. In this presentation, we show how the bootstrap procedure can be applied to assess stability of the parameters of nonlinear as well as linear PCA. In former research, we showed that the CATPCA solution is quite stable, if there are no categories with relatively small marginal frequencies. We will start this presentation by briefly recapturing these results. Subsequently, we will focus on the comparison of the stability of CATPCA on the merged NICHD data to the stability of PCA. We show confidence ellipses for the eigenvalues and component loadings for both solutions.

A Comparison Of Strategies For Estimating Bootstrap Confidence Intervals In PCA

Marieke E. Timmerman, Henk A.L. Kiers, and Age K. Smilde

The bootstrap methodology can be used to estimate confidence intervals (CI's) for the estimated parameters in a Principal Component Analysis (PCA). First, an overview of possible strategies for bootstrapping in the simple case of PCA of scores obtained from independent observations is offered. Attention is paid to the resampling scheme, methods for estimating the CI's, and possible non-uniqueness of the estimated parameters. The quality of the estimated confidence intervals using various bootstrap versions is examined in a comparative simulation study. The bootstrap version using non-parametric resampling and the bias-corrected and accelerated method appears most recommendable in practice. Second, approaches towards bootstrapping in the more complicated case of functional PCA, where the successive observations are dependent, are discussed.

Aspects Of Multilevel Modelling For Item Response Models

Murray Aitkin

This talk will touch on a number of current issues in the MLM formulation of IRT models. These include:
* Allowing for clustered survey designs by additional levels of the MLM
* Efficent algorithms for many levels
* Generalizing the distributional model for the latent ability

Multilevel Modeling Of Complex Survey Data

Sophia Rabe-Hesketh, and Anders Skrondal

In multistage surveys clusters or primary sampling units are sampled at stage 1, then subclusters at stage 2, etc. until elementary units are sampled at the final stage. Multilevel models take into account the clustered nature of such data by including random effects representing unobserved heterogeneity between the clusters sampled at each stage. However, two other common features of survey designs, stratification and unequal probabilities of selection, are often ignored in multilevel modeling. This talk describes pseudo maximum likelihood estimation where inverse (selection) probability weights are used to take into account unequal probabilities of selection at each stage. Unfortunately, this approach can yield biased parameter estimates if there are weights at the lower levels of the model. Several methods for scaling the weights to reduce the bias will be discussed and simulations presented to evaluate their performance. The sandwich estimator is used to obtain standard errors appropriate for complex survey data with stratification, multistage sampling and unequal selection probabilities.

Multilevel Analysis -- Beyond The Software

N. T. Longford

The talk presents multilevel analysis as a general method for dealing with hierarchically structured populations and multiple contexts. A perspective is discussed in which a distinction is drawn between fixed and random effects. However, a contradiction in the generally accepted wisdom of analysing with the 'correct' model is identified, and shown to be frequently breached in some applications of multilevel analysis. Some common problems with using the standard output of the multilevel analysis will be diagnosed; they are related to dealing with uncertainty in a collection of estimates and non-informative selection of the target of estimation. A gold-standard solution is proposed -- formulating inferential goals prior to data inspection. Although it is difficult to adhere to consistently, departures from it should be considered as a source of bias in estimation of both the target and the sampling variation of its estimator.

Exogeneity Of Regressors In Multilevel Models

Peter Ebbes, Ulf Bockenholt, and Michel Wedel

This presentation provides an overview of currently available methods that can be used to test for different types of dependencies between random effects and regressors in a multilevel context. We present results from Monte Carlo studies designed to investigate the performance of these methods. In addition, we discuss estimation methods that can be used when some but not all of the random effects and regressor independence assumptions are violated. Because current methods are limited in various ways, we will also present a list of open problems and suggest solutions for some of them. As we will show, the issue of regressor random-effects independence has received some attention in the econometrics literature, but this important work has had little impact on current research practices in the social and behavioral sciences.

Lord's Paradox Revisited: Implications For Panel Data Analysis Using Hierarchical Models.

Jeroen K. Vermunt

At a seminar a colleague from the sociology department presented his research on the effect of life course events (divorce) on psychological wellbeing (depression). Data were collected by a two-wave panel study, and the event of interest (divorce) occurred between the first and second panel wave. This yields the well-known non-equivalent control group design. Although such a design makes it possible to infer causal effects, it is crucial that one corrects in the right way for selection bias caused by the fact that the composition of the treatment group (divorced) may differ from the one of the control group (married).
In the divorce-depression study, the researchers tried to take into account selection bias by using depression at the first occasion as a control variable in the model in which depression at occasion 2 was regressed on divorce. Such an approach, which is referred to as a conditional or an ANCOVA analysis, is fully wrong in this case, while a simple regression analysis of change scores would have been a much better choice.
Although a lot has been written about Lord's paradox – change score regression and ANCOVA yield different results – I did not find a good (fully convincing) explanation in the literature for why the change score is wrong is the current application. This can, however, easily be demonstrated by specify explicitly the assumed selection mechanism.
What is the link to hierarchical modelling? Well, longitudinal data analysis using multilevel techniques can be seen as a generalization of the change-score method that 1] makes it straightforward to deal with more than 2 time points, 2] can be used to detect differential treatment effects, and 3] can be used with categorical and limited dependent variables. Moreover, by incorporating the multilevel approach in a SEM-like framework, one could specify the selection mechanism, which makes it possible to test the basic assumptions underlying the change score and the ANCOVA methods.

Efficient Designs For Paired Comparisons

Heinz Holling

In psychological research linear as well as logistic models for paired comparisons are often used to measure preferences. In the recent past, efficient designs for these models have been extensively studied. A great amount of theoretical as well as empirical evidence is provided which shows that efficient design strategies outperform established ones. In this talk current results and some new findings will be presented. Special attention is paid to paired comparisons based upon partial profiles in which alternatives are described by subsets of attributes.

The Design Of Conjoint Choice Experiments

Roselinde Kessels, Martina Vandebroek, and Peter Goos

Conjoint choice experiments are frequently used for learning about the preferences of consumers in marketing, patients in health economics, or commuters in transportation studies. In psychology, this type of experiment can be used, for example, to evaluate respondents' attitudes towards the severity of crimes. Currently, there is still a lot of research on the analysis of conjoint choice experiments as well as on the design of this type of study. The optimal design of conjoint choice experiments is not only a hard combinatorial optimization problem, but its solution also depends on the values of the unknown parameters in the statistical models utilized to analyze the experimental results. The purpose of the talk is to present a Bayesian approach to designing conjoint choice experiments to cope with the dependence of the optimal experimental design on the unknown model parameters. Several criteria for evaluating the conjoint choice designs will be compared. A set of practical recommendations will conclude the presentation.

Optimal Test Design In Polytomous IRT

Valeria Lima Passos & Martijn P.F. Berger

With the objective of improving estimation of IRT model parameters, two cases of design optimisation in IRT can be distinguished: *optimal calibration design* and *optimal test design*. The former focuses on the estimation of item parameters whereas the latter targets the latent trait. While statistical techniques of optimal experimental design are commonplace for item calibration, they are comparatively less used for handling issues of optimal test assembly. This presentation will show that underlying methods of *optimal test design* can be productively applied as item selection procedures to assemble a test with multi-categorical items. The polytomous IRT models at stake are the Nominal Response Model (NRM) and the Graded Response Model (GRM). Different item selection criteria (ISCs) based both on local and global information measures are evaluated in terms of their selection strategies and performances. The former is associated with the parametric characteristics of selected items, the latter is assessed via efficiency, accuracy and stability of estimation. Furthermore, both test premises, group-based and adaptive test settings, are taken into account. Finally, based on the obtained findings, general guidelines are presented that can help to pave the way towards an efficient, objective-driven test construction.

On The Design Of Cluster Randomized Trials For The Comparison Of Group Versus Individual Therapies

Mirjam Moerbeek, and Weng Kee Wong

Many interventions studies in the behavioral sciences use cluster randomization so that complete clusters of subjects are randomized to treatment conditions. In trials where the experimental condition is a group therapy, the data in the experimental condition have an additional level of nesting, because the experimental is delivered in focus groups of subjects. This may induce the response to vary across focus groups within clusters, resulting in an additional intra-class correlation coefficient at the focus group level. Although much attention has been paid to the design of cluster randomized trials, knowledge on the design of trials where the number of levels of nesting varies across the treatment conditions is underdeveloped. The aim of this presentation is to study the optimal allocation of subjects to treatment conditions for studies with binary responses. The focus group size is fixed, and the costs vary across treatment conditions. The optimal design is shown to depend on the values of the response rates in both conditions, and the intra-class correlation coefficients at the focus group and cluster level, and it is therefore a locally optimal design. Attention will be paid to the robustness against incorrect prior estimates of these model parameters, and to the construction of robust optimal designs through the maximin principle. The optimal design will furthermore be compared to the design with equal sample sizes in both treatment groups.

Relative Efficiency Of Unequal Versus Equal Cluster Sizes In Multicenter Intervention Studies

Gerard J. Van Breukelen, Larissa Kotova, Math J.J.M. Candel, & Martijn P.F. Berger

Experimental evaluations of public health programs and medical treatments often involve a nested (multilevel, multi-center) design. Examples are prevention of depression among pupils of several schools, prevention of burnout among employees of companies, and clinical trials with patients from many general practices. Equal cluster sizes; that is, sampling the same number of persons within each cluster (school, company, general practice), is optimal for estimating and testing a treatment effect, at least assuming homogeneity of variance and of sampling cost across clusters. Unfortunately, equal cluster sizes are seldom feasible.

This paper addresses the relative efficiency (RE) of unequal versus equal cluster sizes for treatment effect evaluation, assuming a quantitative outcome and a 50:50 randomized allocation of clusters or of persons within clusters. It is shown how the RE depends on the intra-class correlation and the amount of cluster size variation, attaining a minimum of 0.90 or 0.80 at worst for realistic distributions of cluster sizes. The results confirm and generalize those by Kerry and Bland (2001) for English and Welsh general practices. Also, they suggest a simple correction for optimal designs as derived by Moerbeek, Van Breukelen and Berger (2000) assuming equal cluster sizes. Generalization to other than 50:50 randomization and to dependence of outcome variance or sampling cost on the treatment condition is straightforward.

Bayesian Parametric Modelling With A Nonparametric Prior

Stephen Walker

The idea behind the talk is to think about Bayesian inference from a point of view which suggests that all priors should be considered on spaces of distribution, or density, functions. Reasons for this will be explained. From this perspective the standard approaches to parametric model choice and comparison become problematic, even incoherent. However, from a nonparametric point of view, or when the prior is selected so that no comparison or checking of the support of the prior will be undertaken, a natural approach using decision theory can be used to undertake parametric procedures. That is, even though the prior may be large, it is possible to present Bayesian inference from a reduced model.

Bayesian Non-Parametric Model Selection and Model Testing

George Karabatsos

This presentation examines a general Bayesian nonparametric approach to model selection and model testing, useful for evaluating the predictive-utility of models that are either probabilistic (Bayesian or classical-frequentist), or even deterministic. In this approach, conditional on an observed set of data, the "best" model is identified as the one that predicts a sampling distribution (or density) that is nearest in Kullback-Liebler distance to the posterior estimate of the true sampling distribution (density), and the decision is made to reject a model when it has a distance that is "significantly" large (significance is determined from a straightforward calibration of the Kullback-Liebler distance). The posterior estimate of the true sampling distribution is based on a Bayesian non-parametric prior that gives positive support to the entire space of possible sampling distributions (densities). This presentation discusses the theoretical and practical advantages of Bayesian nonparametric approach over all other types of model selection procedures, and over any model testing procedure that depends on interpreting a p-value. The approach is also illustrated on real data sets, for the comparison and testing of (1) models of item response theory, (2) models under stochastic order constraints, and (3) cognitive models.

Hierarchical Bayesian Models for Ordinal Data from Factorial Designs

Jay Myung

For data from a factorially designed experiment, one can often make a priori order predictions on the cell means (e.g., theoretical predictions or previous findings). However, the analysis of variance, the most widely used statistical procedure for analyzing such data, does not allow one to incorporate order restrictions into the procedure. The likelihood ratio test framework (Barlow, Bartholomew, Bremner & Brunk, 1972) has been developed for order-restricted inference in frequentist settings. Obtaining the required sampling distribution of the likelihood ratio statistic under arbitrary ordinal constraints on the parameters, however, has been a major obstacle to applications of this framework in practice. In the present study we introduce a hierarchical Bayesian approach for analyzing ordinal data from factorial designs. The basic idea of this approach is to employ a prior distribution representing orderings on the parameters and then to estimate the posterior distribution by Markov chain Monte Carlo methods. Specifically, we adopt the Bayesian p-value as the criterion to assess model fit and we use the Deviance Information Criterion as a model selection criterion. We illustrate the approach for data from the two-way layout under various order constraints on the parameters.

Order-Restricted Inference and Bayesian Model Selection illustrated via Tversky's study of Intransitive Preference.

Geoffrey Iverson

In the behavioral sciences the vast majority of experimental data are analyzed by the methods of analysis of variance, as matter of routine. Such analyses often suffer from one or other of a number of shortcomings. ANOVA models typically do not capture the expectations of the experimenter in any but a vague sense; the reporting of complicated interaction effects is accompanied by qualifications requiring many words, a sure sign that there is a mismatch between what is expected in the data and the outcome of an analysis; furthermore, the ability to detect genuine experimental effects is compromised (i.e. statistical power is often quite low). On the other hand it is common for experimental hypotheses to be couched in terms of order restrictions, and in recognition of this a sophisticated statistical methodology has been developed to address problems of order-restricted inference. Whereas that methodology should be a commonplace, in practice it is rarely used. Rather than speculate on the reasons for this curious neglect, we illustrate the power of Bayesian methods to select between a pair of order-restricted models that, years ago, were analyzed incorrectly in Tversky's classic study of intransitive choice behavior.

Order-Constrained Inference Using Reparameterized Models

William H. Batchelder

Suppose a parametric model has S parameters, $\theta = (\theta_1, \ldots, \theta_S)$, each free to vary in some positive interval. In general, for order-constrained inference, one might have a partial order P of parametric constraints, where (θ_s, θ_t) in P represents the constraint $\theta_s < \theta_t$. Also there could be numerical constraints such as $\theta_s < x$, where $x > 0$. We explore an approach to order constrained inference based on reparameterizing a model to reflect the order constraints. For example, any single constraint, (θ_s, θ_t), can be represented by introducing a new parameter τ in (0,1), and replacing θ_s by $\tau \cdot \theta_t$ in the model. The new model has S unconstrained parameters, but it is equivalent to the old model with S parameters subject to the parametric constraint. Inference done with the reparameterized model will inform inference for the original, order-constrained model. Several ways to reparameterize a model to reflect special sets of order constraints are presented. Among other things, we examine the family of binary multinomial processing tree models (see Batchelder & Riefer, 1999, *Psychonomic Bulletin & Review*, and Hu & Batchelder, 1994, *Psychometrika*) and show that it is closed under the class of reparameterizations we develop.

A Reconciliation Of Pearson And Yule Through Generalized Correlation, And Some New Tests Of Independence

Wicher P. Bergsma

A generalized correlation coefficient was introduced by Jupp and Mardia in 1980, but has not been studied further outside some specialized applications. We show that the sample version of this coefficient yields both the chi-square and Hotelling's T-square test. That is, the two tests are the same, but applied to different sample spaces. It follows that seemingly unrelated "Yulean" categorical and "Pearsonian" continuous data methods are, in fact, closely related. Recognition of this fact opens up possibilities for more flexible hypothesis tests than are currently available. We shall show that going beyond the usual interval, ordinal and nominal measurement levels it is possible, through generalized correlation, to define flexible "tailor made" tests, which are, in a well-defined sense, intermediate between the chi-square test and ordinary t-test. In an example we demonstrate improved power for detecting differences in political preference rankings for men and women.

Some Statistical Artefacts In Heredity Coefficients.

H. Kelderman and P.C.M. Molenaar

Heredity coefficients may be invalid if the model, e.g. the standard Martin and Eaves model, under which they are computed, is incorrect. In this paper we look at some assumptions of the usual models and the degree to which their violation influences heredity coefficients. of their violation; the influence of randomness of factor loadings, Heteroscedasticity, and the modelling of test scores rather than item scores. For example, for the Martin and Eaves model, there are pervasive substantive arguments in favour of individual differences in factor loadings. It is shown that, under liberal assumptions, data from a heterogeneous 1-factor model yield the same population covariance structure as the corresponding standard 1-factor model. Furthermore, it is shown that the sample covariance matrices and the standard goodness of fit statistics are asymptotically same. A small simulation study yields similar results for realistic sample sizes. So under standard normality assumptions, one does not discover the lack of fit of the model that is the result of misspecification of the probability model, here the randomness of the factor loadings. The effects on the heredity coefficients are substantial. Thus choosing and applying the wrong model may have adverse consequences in practice, even if it provides a more parsimonious description of the data.

Nonparametric Adaptive Estimation of the Threshold

Amrit Pratap and Yaser S. Abu-Mostafa

An important concept in Psychophysics is that of threshold, the boundary between the detectable and the undetectable. We consider the problem of estimating the threshold of a Psychometric Function in a completely nonparametric setting. The Psychometric function is known to be monotonic and if the sample size is small, the naive estimate often violates the monotonicity constraint and better performance is achieved by enforcing the monotonicity constraint on the estimation problem. Better performance is also obtained by sampling sequentially and using the current sample to decide the location of the next example. In this paper, we give the maximum likelihood nonparametric estimator of the threshold and present an adaptive algorithm to estimate it by using small number of examples.

On the Detection of Influential Subsets of Categories in Goodness of Fit Testing

David Magis

To compare a vector of discrete counts from a *k*-category population to a theoretical (known) one, one often performs a statistical test of goodness of fit. It consists in computing Pearson's X^2 statistic, with an asymptotic chi-squared null distribution. However, if the hypothesis of adequacy is rejected, it is not easy to determine which subset of categories leads to that statistical conclusion.
The aim of this paper is to present a method for detecting such influential subsets of categories. We start by building a set of nested intervals for category counts, such that any sample belonging to this "region" corresponds to a (sufficiently) small value of X^2. A fast and simple detection routine is derived. These intervals can also be helpful in other contexts, which are briefly described. Some selected examples illustrate our purpose.

A Model for Gradual Process Change

Paul De Boeck, James Wollack, Allan Cohen

An IRT model will be presented with two processes, so that a second process gradually takes over from the first, while both the *change point* and the *change rate* are considered random parameters in order to model individual differences in both respects. Various applications can be thought of, such as a gradual shift to guessing due to speededness, and a gradual learning process starting from a guessing state. The two random effects play between the two processes. Within one of the two processes a Rasch model is assumed to play, so that an additional random effect is included, for the *proficiency* of the individual. The other process is a guessing process with no individual differences. The model is formulated by Wollack & Cohen (2004) as a model for the simulation of speeded test data. It can be estimated with SAS NLMIXED and WINBUGS. We will report on a small simulation study and a speededness application to real data. We will also discuss a possible extension with a different LLTM for both processes, so that the model can be used also for within-person change between two processes with individual differences (as in a strategy shift).

Reference:

Wollack, J.A., Cohen, A.S. (2004). A model for simulating speeded test data. Paper presented at the AREA meeting, San Diego.

On The Theoretical Relationships Between PLE And JMLE In The Rasch Model

Ernesto San Martín and Paul De Boeck

Joint maximum likelihood estimates (JMLE) have become less used in IRT models, for the Rasch model were extensively used in the early IRT packages, but their popularity has decreased in favor of the MMLE and the CMLE, and they are not widely available. Both Mellenbergh and Vijn (1981) and Verhelst and Molenaar (1988) relate the JMLE to a two way logit model, with items and person raw scores as factors, which we call a pseudo-logit (PL) model. In the first paper the actual fitting is done through the associated log-linear model, while in the second a weighted least squares solution is implemented. In this paper we review the underlying theory and discuss various alternatives for finding the JMLE directly, using a standard statistical package, rather than specialized IRT software. Furthermore, both an exact formula and a useful upper bound on the standard errors of the JMLE are given in terms of the those of the PL-model, with the consequent gain in computational efficiency. Finally, it is shown how can the Fisher Information Matrix of the JMLE be recovered from the PLE. To this end, a special class of partitioned matrices is introduced and their properties explored.

Copula IRT Models To Deal With Local Item Dependency

Johan Braeken and Francis Tuerlinckx

The Rasch model, and item response theory (IRT) models in general, are not robust to violations of local independence (i.e., local item dependencies (LID) can affect the estimation and reliability of the different model parameters by for instance decreasing the standard errors of the persons' proficiency parameters). However, there exist several modelling approaches (e.g., usage of sum scores, conditioning on other responses, additional random effects) that try to incorporate LID, but they have some drawbacks (e.g., information loss, non-reproducibility of marginal probabilities, interpretation problems or difficult parameter constraints). We will introduce a new class of IRT models that extend the Rasch model with copula functions to deal with LID problems. These models belong to the larger class of marginal models in which the marginals and the association structure are modelled separately (in our case by means of the Rasch model and copula functions, respectively). These copulas can offer a solution for some of the flaws haunting the existing models.

Modeling Partial Knowledge In Multiple-Choice Tests Using Elimination Scoring

Rianne Janssen, Eddie Ip, Paul De Boeck and Michel Meulders

The elimination scoring technique of multiple-choice questions originally proposed by Coombs, Milholland, and Womer (1956) requires respondents to eliminate all the response alternatives they think are incorrect. In the present paper, each response alternative of a multiple-choice question is scored dichotomously as to whether the subject's response (eliminate or not) is correct. Following Huynh (1994, 1996), it is shown that if the binary items that correspond to the response alternatives are summed for each multiple-choice question, then the sum of these items follows the partial credit model (PCM; Masters, 1982), if the binary items have equal discrimination, but independent of possible interactions among the binary items of the same multiple-choice question. However, following Meulders, Ip, and De Boeck (2004), a more general IRT model can be derived by considering the response patterns on the binary items of each multiple-choice question as a partially ordered set (poset). The general IRT model encompasses the PCM, the ordered partition model (Wilson, 1992; Wilson & Adams, 1993), as well as componential poset models, where the probability of particular score pattern is modeled as a linear function of the item parameters of each binary item. The approach is illustrated with a test on synonyms.

A Latent Class Extension Of The Dimension/Category Framework

István Hidegkuti, Paul De Boeck

In psychology it is rather common to use categories, for example psychiatric categories. An important question is whether what is considered to be a category is really categorical and not just dimensional. A new approach to investigate this issue is the Dimension/Category (Dimcat) framework that makes use of item response theory. In its original form Dimcat is used for dichotomous indicators and manifest categories. However, in a number of cases the indicators are polytomous and the class membership of the respondents is unknown.

We will present an extended version of the Dimcat framework: extended to latent classes, and graded responses. The basis for this is a discrete mixture model where each mixture component is a graded response model. The decision of dimension- or category-likeness is based on the fit statistics of restricted versions of this model. The extended dimcat framework will be presented through a simulation study and a real data application. A comparison will also be made with the Taxometric approach that is the most popular method thus far for the differentiation of categorical and dimensional structures.

Poster session, 10.40h-11.20h

Using Latent Curve Modeling For Analysis Getting Not From Longitudinal Research.

Olga Mitina

Latent Curve Modeling (LCM) is included into an arsenal of methods of SEM and in the certain degree integrates ideas of regression, factor analysis and ANOVA. LCM is applied to the analysis repeating measurements, most usual for the data received from longitudinal researches. The purpose is revealing latent variables determining character of changes of investigated characteristics in due course.

But LCM can be use not only for longitude researches but in cases when we want to study how one variable depends on any independent ordinal or interval variable if we can repeat measures of the dependent variable sequentially (in order) changing values of this independent one. Measures should be done for each subject at each value of the independent variable.

In our paper we will present two examples of using LCM for non-longitude research.

In the first one we studied implicit stereotypes in Russia concerning how Internet access (independent variable) determines Internet activities and how gender and Internet self-assessment determine these stereotypes.

The second (nontraditional) use of LCM allowed us to reveal differences in implicit stereotypes about drugs among Russians. Using this method of data analysis allowed to determine implicit interval scale to measure attitudes to this or that drug.

Influence Of Sample Size, Estimation Method And Model Specification On The Recovery Of Weak Factors In Confirmatory Factor Analysis: A Monte Carlo Study.

Carmen Ximénez

This poster summarizes the results of a large Monte Carlo simulation study assessing the separated and combined effects of estimation method (ML and ULS), sample size (100, 300 and 500), number of factors (1, 2 and 3), correlation between factors (0 or 0.50), model specification (correct vs. incorrect) and loading size in the weak factor (0.25, 0.35 and 0.50) upon the recovery of weak factors in confirmatory factor analysis. Results show that the ULS estimation method produced more convergent and proper solutions. As expected, the recovery of the weak factor improves for correctly specified models and gets worse especially for models where misspecification implies omitting one factor. However, if factors are correlated the weak factor is satisfactorily recovered both in correctly and incorrectly specified models. Also, the recovery of the weak factor improves as the number of factors decreases and the loading size in the weak factor increases. The recovery is especially poor for misspecified models by omitting one factor and for models with loading sizes of 0.25 in the weak factor.

Response Bias due to Mode of Data Collection: Validation and Application of a Structural Equation Model of Response Styles

Bert Weijters, Maggie Geuens and Niels Schillewaert

We investigate bias of survey results due to response style, i.e. a respondents tendency to (dis)agree with items irrespective of the content of the specific item, e.g. yeah-saying or acquiescence bias. We compare and validate two operationalizations of response styles (Greenleaf 1992; Baumgartner and Steenkamp 2001) to assess the bias they cause in different modes of data collection. While both typologies originally used single indicators for each response style, we compute multiple indicators in order to validate the measures. We assess discriminant validity and internal consistency. We provide a method to specify a structural equation model operationalizing these response style measures as latent constructs with multiple indicators. The model allows us to compare the level of bias among four comparable samples. These samples filled out the same questionnaire via different modes of data collection: a mail survey, a survey by telephone, a web survey among a panel of respondents and a web survey with banner recruitment. We find that the latter sample shows the severest bias due to response styles, especially yeah-saying. The telephone survey shows the highest intra-subject standard deviation (or response range).

Correspondence Analysis Of Square Contingency Tables

Anoek Maan and Mark de Rooij

Correspondence analysis of square contingency tables is in general problematic due to the large values on the diagonal of such a table. The diagonal takes the solution to be a symmetric one, whether the data are symmetric or not. If it would be possible to minimize the influence of the diagonal, the asymmetry in the data can be really appreciated. We propose a new algorithm, based on the ideas of Greenacre, where the influence of the diagonal cells is minimized. The new method is compared to two existing correspondence analysis procedures: The first was proposed by Van der Heijden where CA is applied to the residuals from the quasi-independence model; The second was proposed by Greenacre who proposed to analyse an extended matrix using correspondence analysis.

Estimating the π* Goodness Of Fit Index For The Finite Mixture Nominal Response Model.

Javier Revuelta

Testing the fit of finite mixture models is a difficult task since asymptotic results on the distribution of likelihood ratio statistics do not hold. This poster describes the application of the π* goodness of fit model (Rudas, Clogg & Lindsay) to finite mixture nominal response models. The π* statistics assumes that the population is composed of two subpopulations, those that follow a parametric model and a residual group outside the model; π* is defined as the population proportion in the residual group. In this proposal, the population is divided in two or more groups or classes. Several groups follow a nominal response model, and there is also a residual group. We present maximum likelihood algorithms for estimating item parameters for each group, the probabilities of the groups and π*. The paper includes also a simulation study on goodness of recovery for the two-class two-parameter logistic model. Finally, an example is presented with real data from a multiple choice test.

Modelling The Relationship Between Self-Concept And Mathematics Performance In EU Countries

Maria Eugénia Ferrão, Maria de Fátima Simões, and Helena Carriço

The self-concept is a construct widely studied in the educational context. Its interest is due to the influence that it has on a lot of behaviour variables, which motivate people to action, especially in those actions prone to successful performance demonstration. There is a general consensus amongst researchers about the idea that when students have an adjusted selfconcept, their social and personal characteristics show a better adaptation independently of the real performance level or individual capacity. In fact, that construct interferes on motivation making individuals to prosecute aims and objectives. A multilevel model is applied on European data of PISA collected in 2003. Controlling by gender, schooling trajectory, motivation for studying maths, socio-economic level, estimates show a pattern of association between personal self-concept and performance in Mathematics. A logistic model let us identifying risk groups that inadequately evaluate their personal competence. Authors found similar empirical evidences for Brazilian educational context.

Validating of Measuring Attitude Toward Smoking scale among Shahid Chamran Students of University Ahvaz

G. Rajabi

The aim of this research was validating of Measuring Attitude Toward Smoking Scale among Shahid Chamran's Students of University Ahvaz. In this research 250 Shahid Chamran,s students that were randomly selected using a multi-phase method have completed Measuring Attitude of Toward Smoking Scale. The results showed that: (1) Cronbach's and Split-half reliability coefficients (internal cinsistency) were satisfied for the while scale and in extracted forth factors. And (2) factor analysis this scale using the Principale Components Analysis (Varimax rotation) four factors were extracted, these are: relationship with smokers, smoking restrictions and laws, health and non-smoke settings, and non-name. In finally, researchers and students of university can be used of Measuring Attitude of Toward Smoking Scale as reliable and valid instrument in educational and research settings, particular of in investigation of attitude smoker and non-smoker peoples toward smoking.

Treatment Of Missing Data In Nonlinear PCA: Standard Options And A New Proposal

Pier Alda Ferrari and Paola Annoni

The presence of missing data in real datasets is often a limiting feature when interpreting results by MVA techniques, especially if missing data show specific patterns.
The goal of the paper is to study and compare various options for the treating of missing data when dealing with ordinal data to be analyzed by Nonlinear PCA. Firstly, available standard options are analyzed with the aim of detecting limiting aspects and underlying hypotheses. Secondly, a different proposal is presented which integrates the full capability of Nonlinear PCA technique and an imputation method for missing observations based on ‘city block’ distance.
Results both from simple fictional case and a real case are discussed and compared. The fictional case has the advantage that, extreme situations can be defined in order to highlight every possible interpretations. On the other end, analysis of real data is helpful in showing interesting implications of various treatments of missing data.
The study represents a suggestion for those who have to apply Nonlinear PCA to ordinal variables in the presence of specific missing data patterns.

Incomplete Data In Consumer Research

Maaike J. Velderman, Frank M.T.A. Busing, and Mark de Rooij

In consumer research subjects are often confronted with many objects which must be scaled or ordered according to preference. For example subjects have to judge several mobile phones or different steam irons. Problems arise when subjects are confronted with digestable objects. Imagine to taste a large number of ice-creams or a lot of glasses of whisky. It would be more reasonable if we could present less objects to the persons.
Instead of asking a subject to judge all the objects one could present a subset to this person. In our research we investigated several data deletion patterns to create these subsets; moreover, within each of the patterns the percentage of deletion was also varied.
The analyses to be shown are based on empirical data obtained from a company in the Netherlands. The data are perfectly suitable for an unfolding analysis. Optimal combinations of patterns and deletion percentages will be discussed. Perhaps more important, fallible designs will be addressed.

Dealing With Nonnormality: A SEM Approach To An Organizational Study

Maria de Fátima Salgueiro, and Sílvia Silva

In recent years statisticians and SEM methodologists have studied the implications of non-normal data in terms of biased estimates, standard errors and incorrect test statistics. Different types of solutions have been proposed, including normalizing data transformations and the use of robust or asymptotically distribution free estimation methods. Applied researchers, however, often ignore the characteristics of the real data they intend to model, namely nonnormality, thus compromising results and conclusions.
Occupational safety is an important issue in applied organizational psychology. A theoretical framework was built and SEM was used as the statistical tool for testing the proposed model. Data were collected in several big and medium sized Portuguese companies. Latent constructs of interest include organizational safety climate, team safety climate, risk perception and workers' safety behaviors. Workers had to fill in a self-report questionnaire, in which most of the questions were measured on a seven-point Likert type scale. Due to the nature of several questions, data collected turned out to be highly skewed. Kurtosis is often also a problem.
Considering our organizational safety study, this talk illustrates how different methodological options influence results and conclusions. Particular attention is devoted to measurement scales (continuous *versus* ordinal), data nonnormality and sample size.

Predicting Changes In Depression With Interacting Person Characteristics: A Comparison Of MARS, CART, And RTA

Paul van Schuppen, Elise Dusseldorp, and Cornelis M. van Putten

During multiple regression analysis, when interacting variables are allowed, it is sometimes difficult to decide which variables actually should be used. When the ratio of variables/cases is high, not all variables should be allowed to interact, because a larger variables/cases ratio will cause loss off power. In order to find out which variables to use, several modeling techniques have been created. Among these are: Multiple Adaptive Regression Splines (MARS), Classification and Regression Trees (CART) and the Regression Trunk Approach (RTA). In order to see how these models perform with real data, data from an intervention study for elderly with a depression have been used.
In this regression problem with many possible interaction variables these models where compared on their cross-validated prediction error. We also looked at the interpretation possibilities of the models, and how well they will perform when outliers are present.

Intraindividual Variability In Self-Esteem Measures: An Application Of The Latent Change Variable Models

Vesna Buško and Orjana Marušić Štimac

The developmental changes in general self-esteem were analyzed on a sample of 308 seventh and eight grade primary school participants of both sexes. The data on the general self-esteem were collected in the preadolescent sample on four occasions of measurement within the two year period. Following the assumptions of the latent state-trait theory, the measurement models were defined to represent the degree to which variations in self-esteem measures were due to individual dispositions and/or to occasion-related influences. Further, the models of LST theory were specified to include the latent variables corresponding to the state variability and the trait change in self-esteem scores. By use of the true change modeling procedures the identified latent change variables were found to be related to the measures of students' academic achievement as well as their engagements in sports.

Clustering Nominal Data with Equivalent Categories: a Simulation Study Comparing Restricted GROUPALS and Restricted Latent Class Analysis

Marian Hickendorff, Cornelis M. van Putten, Willem J. Heiser, Norman D. Verhelst

This study concerns methods for clustering data of nominal measurement level, where the categories of the variables are equivalent: the variables are parallel indicators. Two techniques are compared on their cluster recovery in the analysis of simulated data sets with known cluster structure, by means of the adjusted Rand index.

The first technique is GROUPALS, an algorithm for the simultaneous scaling (by homogeneity analysis) and clustering of categorical variables. To account for equivalent categories, equality constraints of the category quantifications for the same categories of the different variables are incorporated in the GROUPALS-algorithm. The second technique is Latent Class Analysis, with the extra restriction to account for parallel indicators that the conditional probabilities are equal over the variables.

The number of clusters, the relative class size and the number of categories per variable are also varied in the simulated data sets, to study their effects on the cluster recovery performance of both techniques.

On Mixing Attitude Item Relational Structure Analysis

Hsiang-Chuan Liu, Maw-Fa Chein, Jing-Ming Ju, and Tung-Sheng Liu

There are generally two kinds of scores for non-categorical attitude items: graded rating scores and symmetrical rating scores. Makoto Takeya (1987) proposed two formulas of item relational structure analysis for these two kinds of scores respectively. These formulas can only be used for items of equal choices and are not interchangeable. The purpose of this study was to provide a new series of generalized indexes of ordering coefficients. They can be used for items with mixing rating scores, which include graded rating scores and symmetrical rating scores with equal or unequal number of choices. These coefficients include the first order generalized index of ordering coefficient, the modified and more sensitive first order generalized index of ordering coefficient, the second order generalized index of ordering coefficient, the modified sufficient and consistent second order generalized index of ordering coefficient, and the higher order generalized index of ordering coefficient. In addition to mathematical derivation and analysis, examples of attitude items related to instruction were presented to introduce the methods of item relational structure analysis.

The application of correspondence analysis to simulated unfolding data

David P. Verschoor, Mark de Rooij, Marike Polak

In this study we investigate the use of Correspondence Analysis when applied to single peaked data simulated by IRT-like models. To simulate the data two kinds of item response unfolding models were used, the hyperbolic cosine model by Luo (HCM) and the generalized graded unfolding model by Roberts (GGUM). Three different types of correspondence analysis were studied:1) The standard correspondence analysis, of which Heiser showed that it performed good when applied to single peaked data; 2) Correspondence analysis using conjoint coding which provides extra information concerning the discriminability of the items; 3) Correspondence analysis using successive categories (Nishisato) that provides thresholds between the response options.
There are two main objectives of this study. First, we want to know how well the three types of CA's describe the item response data. Secondly, we want to look at the differences between the three types of CA: do the two extended correspondence analyses approaches provide useful information?

Symposia and sessions, 11.20-13.00

Interval Estimation for Proficiency in Multiple-Format

Chiou-Yueh Shyu

Like single-format testing, the primary purpose of multiple-format tests is to make inferences about the proficiency of examinees. In applied settings, proficiency is measured imperfectly and always contains some degree of random error. Confidence intervals (CIs) for the proficiency integrate the information of sampling error into the estimation procedure.
The research on procedures for constructing CIs in mixed-item-format testing is limited, especially in the application of item response theory (IRT). Via a simulation study, the purpose of this study is to construct various CIs for proficiency under various polytomous IRT models. CIs constructed based on pivotal inference , fiducial inference, and bootstrap inference will be discussed and compared.
Computer simulation studies based on a generalized partial credit model and a generalized graded response model are used to evaluate the types of confidence intervals. Estimates of item parameters for a large-scale standardized English test containing 75 MC items and a polytomous item are assumed to be known and were used as a basis for the simulation. Under each model, one thousand simulation trials were run for each θ point from –3 to 3, with an increment of 0.2.

Estimation Of Try-Out Item Parameters In Computerized Tests With Both Try-Out Items And Pretested Items.

Marie Wiberg

Computerized achievements tests that consist of both pretested items with known item parameters and try-out items with unknown item parameters are considered. The aim is to show how the item parameters for try-out items can be estimated with help of the pretested items in a computerized test. The three-parameter logistic item response theory model is used, but a problem is that the unknown abilities might influence the estimations. This is of special importance in estimating the variances of the item parameter estimators. A few examples will illustrate this problem.

Limited vs Full Information Methods for IRT estimation: a MonteCarlo study

Carlos García-Forero, and Alberto Maydeu-Olivares

Within an IRT framework ordinal response models, such as the Graded Response Model (GRM; Samejima, 1969) are generally estimated using Maximum Likelihood (ML) via the EM algorithm. Within a SEM framework this model can be estimated using a Limited Information (LI) multi-step procedure (e.g., Muthén, 1984). Since ML estimation is asymptotically optimal, in large samples parameter estimates should be more accurate, and standard errors smaller, than for LI methods.
We perform an exhaustive simulation study using a common parameterization of the model to compare the performance of both methods in finite samples. Relative bias of the parameter estimates, standard errors, as well as coverage of true parameters, are assessed. Varying conditions of sample size, data skewness, dimensionality, number of categories, and parameter values are investigated.

Numerical Integration In IRT Models

Jorge Gonzalez, Francis Tuerlinckx, Paul De Boeck

When using the Marginal Maximum Likelihood (MML) approach to estimate IRT models, the integral appearing in the marginal likelihood is analytically intractable, so that numerical methods such as Gauss-Hermite quadrature (GH) are needed. When more than one random effect is considered in the model, the dimensionality can be a problem because the number of quadrature points is too high. A possible solution can be found among the Quasi-Monte Carlo (Q-MC) techniques, because they yield quite good approximations for high dimensional integrals with a much lower number of points, chosen for their optimal location. In this paper a comparison will be made between three techniques: GH, Q-MC, and full Monte-Carlo integration (MC).

The EM Algorithm As Relative Entropy Minimization

Alexander Weissman

The expectation-maximization (EM) algorithm has found wide applications in maximum likelihood problems when direct maximization of a log-likelihood function is difficult or intractable. Alternating between an expectation (E) step and a maximization (M) step, the EM algorithm assures that: (1) the log-likelihood function will not decrease after each EM cycle; and (2) the log-likelihood function will converge to a maximum. In practice, convergence of the algorithm is typically determined by monitoring the (marginal) log-likelihood function. Interestingly, the negative of this log-likelihood function is equal to the relative entropy, or Kullback-Leibler distance, between the joint distribution of observable and hidden (latent) variables, and the conditional distribution of the hidden variables given the observables. Thus, maximizing the log-likelihood function can be interpreted as minimizing the entropic distance between two distributions. This interpretation allows the log-likelihood function to be partitioned into two components, with each component providing additional information regarding the progress of the algorithm. Examples analyzing these two components are provided for IRT item parameter estimation problems.

Predicting Group Variables From Variables Measured At The Individual Level: A Latent Variable Multilevel Approach

Marcel Croon

Classical multilevel model pertain to what Snijders and Bosker(1999) have called macro-micro situations in which an outcome variable measured at the individual level is predicted by explanatory variables which are either measured at the same individual or the higher group level. In many applications in the social and behavioural research, (especially in the area of organizational studies and personnel psychology), one is confronted with a micro-macro situation in which outcome variables measured at the group level are predicted or explained by means of explanatory variables measured at the individual level.
Traditional approaches to analyse data in micro-macro situations involve either disaggregating or aggregating the data. Both approaches have clear disadvantages. In this presentation a latent variable multilevel model will be introduced that formalizes the process of aggregating the data. It will be shown that a regression analysis on aggregated data yields inconsistent estimates of the model parameters. Consistent estimates of these parameter can be obtained by a standard regression analysis on adjusted group means in which the group means for the aggregated variables are replaced by their BLUPs: Best Linear Unbiased Predictors.
This approach will be illustrated by analysis on simulated and real data.

Omitted Variables in Multilevel Models

Jee-Seon Kim and Edward W. Frees

This talk presents statistical methodology for handling omitted variables in a multilevel modeling framework. In many non-experimental studies, the analyst may not have access to all requisite variables, and this inevitable omission may lead to biased estimates of model parameters. By exploiting the hierarchical nature of multilevel data, a battery of statistical tools can be developed to test various forms of model misspecification as well as to obtain robust estimators in the presence of omitted variables. The methodology allows for tests of omitted effects at multiple levels jointly and separately. A simulation study suggests that omitted effects at lower levels may cause more severe bias than at higher levels. Dissimilar variance component sizes across levels also appear to intensify omitted variable bias. A real data study illustrates how the estimates and associated tests can be used for model building and respecification. It is argued that multilevel data typically contain rich information to deal with omitted variables, offering yet another appealing reason for use of multilevel models in the social sciences.

Profile Analysis by the Cross-Classification Multilevel Model

Rüdiger Mutz

In applied psychology profiles for the graphical illustration of empirical results are very common and important. A profile as the scores of an invidual or group over a series of scales can be characterized in terms of elevation (overall mean), scatter (the standard deviation of the profile) and shape (the pattern of „ups and down“) (Cronbach & Gleser, 1953). However, statistical methods for seperating the effects of elevation, scatter and shape in group profile analysis are rather cumbersome. Stemmler (1988) for instance offers a complicated multistage discriminant analysis to seperate these effects.
In this paper the problem of discrimination of group profiles will be reformulated and simplified under the heading of multilevel models (Raudenbush & Bryk, 2002). This reformulation makes it possible not only to estimate elevation, scatter and shape as parameters of a cross-classification multilevel model using standard statistical software, but also to test which of the component makes up the group profile differences. For implementation a SAS-program will be shown. A survey of n=1490 psychology students in Germany, signed up 2000, serve as data for illustrating this approach. Finally, additional applications of this approach will be mentioned for example the comparison of repeated measured profiles.

Confidence Intervals For The Amount Of Heterogeneity In Meta-Analytic Random- And Mixed-Effects Models

Wolfgang Viechtbauer

Random- and mixed-effects models are often used in meta-analysis when the effect size estimates contain additional variability (heterogeneity) that cannot be accounted for based on sampling error and the fixed effects (moderators) alone. While methods for estimating and testing the significance of the variance components (i.e., the amount of heterogeneity) in meta-analytic random- and mixed-effects models have been discussed at some length in the literature, little work has been conducted so far examining methods for constructing confidence intervals thereof. Not surprisingly, Wald-type confidence intervals typically yield inadequate coverage probabilities, especially when the amount of heterogeneity is close to zero, where the asymptotic approximation to the distribution of maximum likelihood and restricted maximum likelihood estimators tends to be poor. A variety of alternative methods for obtaining confidence intervals will be examined, including methods based on inverting the regular and restricted likelihood ratio and the commonly used Q test.

A Comparison of Parametric and Nonparametric Approaches for Two-level Random-Intercept Models

Olga Lukociene

The two basic assumptions underlying standard linear random-intercept models - normal errors and normal random intercepts - may be unrealistic in social science research. Outcome variables of interest are very often categorical variables, which makes it necessary to use non-linear mixed models. Also the distributional assumptions about random intercepts are not realistic in most applications. Latent class regression analysis provides an alternative nonparametric approach that relaxes this assumption and that makes it straightforward to deal with categorical outcome variables. No systematic comparisons of the two random-intercepts approaches have been performed so far. The nonparametric approach has several practical advantages. But that is of course not enough to prefer that particular method. The objective of my research is to provide a systematic comparison between parametric and nonparametric random-intercepts models.

Proficiency Structure Using K-Means Clustering

Seongah Im, James E. Corter, and Kikumi Tatsuoka

A partial ordering relation has been used to construct a network of knowledge which is represented graphically on dichotomously responded items. However, not all performance in education and psychology are always recorded as dichotomous responses and there exist cognitive attributes genuinely entailed in items. Thus, it would be interesting to discover a structure of cognitive attributes recorded on a continuum between all and none.

A purpose of this research is to propose an approach to constructing a structure of knowledge based on cognitive attributes. In order to do this, two steps are needed. One step is to cluster subjects into non-overlapped latent groups that reflect underlying patterns. The number of groups is determined by examining dissimilarities of groups and by using a proportion of variance components of the groups and cognitive attributes.

The other step is to set a cutoff score that indicates the level of knowledge or skill that is considered sufficient in a domain given. Setting a cutoff score depends on judgment that can lead to errors. Accordingly, the cut score should be decided optimally where the false positive- and false negative-error rates are minimized.

By doing the two steps, we can construct a structure composed of proficiency pattern vectors which satisfy a partially ordered relation. The structure can be called *'Proficiency Structure'*.

Hierarchical Classes Modeling Of Real-Valued Data

Jan Schepers and Iven Van Mechelen

Hierarchical classes (HICLAS) models constitute a distinct family of classification models for N-way N-mode data that imply a simultaneous clustering of each of the modes in the data. Up to now the HICLAS family was limited to binary and rating data. In this talk we present a novel expansion of the family to deal with real-valued N-way N-mode data. An estimation method for the novel models is explained and simulation results on its performance are presented. Finally, an application to an empirical data set will be discussed.

Model Based Clustering of Large Data Sets Using Bayesian Computational Statistics

Herbert Hoijtink

This paper will present an exploratory model based clustering approach for the analysis of large data sets. Basically a model based cluster analysis (another name that is often used is latent class analysis) searches for homogeneous groups of persons that is groups of persons that give similar responses to a set of items.
The approach proposed addresses a number of practical problems that often arise in exploratory model based cluster analysis of large data sets:
1) The application at hand concerns a large data set (4000 persons and 250 items). Obtaining clusters for large data sets is not a well developed area. Based on two conjectures a MCMC clustering algorithm that renders non-degenerate clusters will be proposed.
2) For large data sets the usefulness of the likelihood ratio goodness of fit test is questionable because the number of possible response vectors by far out-weights the number of observed response vectors. An alternative that does not suffer from this limitation is proposed.
3) In the application at hand data are missing by design: each pupil receives a specific subset of the 245 items in the test. It will be shown how to deal with the missing data both for obtaining clusters and goodness of fit evaluation.

Market Segmentation Using Bayesian Model Based Clustering

Pascal van Hattum

(Marketing) Companies often have data matrices with more than 2500 records and with more than 100 items. In such datasets, records often represent persons and items are often answers to questions obtained via market research.
One multivariate analysis which can be performed on such a dataset is a cluster analysis. The main goal of cluster analysis from a market segmentation perspective is to find groups (=clusters) of persons who give (more or less) the same answers to the items.
A simple cluster model assuming within cluster independence of the item responses is often not sufficient for such marketing datasets. Some of the items have strong within cluster dependencies.
This paper will present a model based clustering approach for the analysis of large data sets, in particular from marketing companies.
The approach proposed is a latent mixture of log-linear models that contain main-effects and specific sets of two-way interactions (to account for within cluster dependencies) and deals with data that are missing by design.
In the application a MCMC algorithm is used to group persons according to their item responses. The algorithm is based on two conjectures with respect to the geometry of model based cluster models.

A General Method for Constructing Exact Confidence Intervals for True Scores

Won-Chan Lee

In classical test theory, an examinee's true score, τ, is viewed as an expected value of "hypothetical" observed scores, X, arising from repetition of the measurement procedure. Under the assumptions of randomly-parallel forms (Lord, 1955, 1957) and dichotomously-scored items, there exist several confidence-interval approaches for estimating τ based on the binomial distribution (e.g., Clopper & Pearson, 1934; Agresti & Coull, 1998). In this paper, a general method is presented for constructing "exact" (i.e., without normal approximation) confidence intervals that can be used for tests with dichotomous items, polytomous items, and mixtures of both types. Roughly speaking, this method obtains the two endpoints of a (100γ)% confidence interval by (a) finding the cumulative distribution function (cdf) given τ, and (b) determining the $(1\pm\gamma)/2$th quantiles from the cdf using a linear interpolation. Three examples are provided: (1) Items are scored dichotomously, and the total test scores are assumed to follow a binomial distribution; (2) items are scored polytomously, and the total test scores are assumed to follow a multinomial distribution; and (3) mixtures of both polytomous and dichotomous items, and the total test scores are assumed to follow a compound multinomial distribution.

Asymptotically Distribution-Free Interval Estimation Of Coefficient Alpha

Donna L. Coffman and Albert Maydeu-Olivares

Van Zyl, Neudecker and Nel (2000) have recently derived the asymptotic distribution of the sample coefficient alpha assuming that the observed data is multivariate normal.

Yuan and Bentler (2002) showed that this normal theory (NT) based standard error for the sample coefficient alpha is asymptotically robust to violations of MVN assumptions under some conditions that unfortunately cannot be verified in applications. Also, Yuan, Guarnaccia and Hayslip (2003) provided the ADF (asymptotically distribution free) asymptotic distribution of coefficient alpha. However, computing the ADF standard error of the sample coefficient alpha can be rather computer intensive when the number of variables is large.

Here we provide a computational formula for the ADF standard error of sample coefficient alpha that can be computed effortlessly for any number of variables. Also, we perform a simulation study to compare the confidence intervals for sample coefficient alpha under NT and ADF assumptions. Our simulation study shows that applied to categorical data the ADF confidence intervals are generally more accurate than the NT confidence intervals, which are too narrow.

The Validity Of Operational Scores After Rater Disagreement Resolution: A Monte Carlo Simulation Of Performance Assessment

James A. Penny, and Robert L. Johnson

When multiple raters score a writing sample, on occasion they will award discrepant scores. In order to report a single score to the examinee, some method of resolving those differences must be applied to the ratings before an operational score can be reported. Several forms of resolving score discrepancies have been described in the literature. Initial studies of the various methods (e.g., rater mean, expert, tertium quid, parity, discussion), however, have demonstrated that decisions about student performance can differ depending on the resolution method applied, suggesting that the choice of rater disagreement resolution model can influence the subsequent validity of the operational scores. Thus, studies are needed to investigate the validity of the scores associated with each resolution model. In this research, we conducted a Monte Carlo study (500 replications of 50 papers) to vary the elements (e.g., rater agreement with validation scores, expert agreement with validation scores, rater drift) associated with scoring and disagreement resolution to determine the conditions under which a particular resolution method might be superior (i.e., have higher validity) to the others.

Estimating Average and Expected Conditional SEM at Each Raw Score on Polytomously-Scored Items Using Multinomial Error Model

Rashid Saif Al-Mahrazi

Unlike dichotomously scored items, all item score patterns for the same total raw score on polytomously-scored items do not necessarily have the same estimate of conditional standard error of measurement (SEM). Practical constraints including fairness and scaling often require that all examinees with the same total raw score receive the same value for conditional SEM (Brennan & Lee, 1999; Lee, 2001). Brennan and Lee (1999) and Lee (2001) have suggested three procedures: two are based on sampled data and the third is based on all possible item score patterns. However, all three existing procedures require a large amount of computations of the conditional SEM for a number of item score patterns at each total raw score. The third procedure requires listing all possible item score patterns for each total raw score, e.g., with tests of polytomous items scored from 0 to 5, there are 780 item score patterns for a score of 12 on a 5–item test, and 4,395,456 item score patterns for a total raw score of 25 on a 10-item test. Therefore, a new simpler procedure is desirable for estimating a single conditional SEM at each total raw score. The purposes of this paper are to present a mathematical demonstration of a new procedure that employs a multinomial error model to compute a single estimate of the conditional SEM at each total raw score; and to illustrate the application of this procedure to real data.

Item Screening

Dougal Hutchison

Item screening is an important element of a test developer's task. Some basic principles are relatively uncontroversial. Items need to measure what they are supposed to measure, to be clear and unambiguous, not to discriminate unfairly against or be offensive groups of those tested, to distinguish well between those who know a concept and those who do not, to be appropriate in terms of difficulty for those tested, to represent a good use of the testee's time, and so on.
Classical Test Theory has been around for a fairly long time, and has developed a relatively standard set of criteria for assessing whether items are satisfactory. The situation for IRT is less well-established. The obvious but flawed answer is to use statistical significance, available on all standard IRT/Rasch packages. The difficulty is that statistical significance is as much a test of how large your sample is whether you have actually found a serious effect, and there is a substantial range of sample sizes used. Some other approach is needed
The proposed paper would discuss how this could be achieved in IRT settings, and describe the performance of a statistic which would assess the contribution of a proposed new item.

A Neglected Method For Variable Selections

John Sachs and Shing-on Leung

Principal component can reduce dimensions of multivariate data but components consist of linear combinations of all variables. Early in 1981, Bhargava and Ishizuka's (1981) had proposed a simple, but neglected method, the BI-method, to select subsets of variables instead of using all. It uses the trace information as a criterion, which is the same as ordinary regression analysis, and uses sweep operator which can be easily programmed. The BI-method is however stepwise in nature, and this paper suggests that there is an all-possible subsets alternative, the all-pos method, which can also be easily programmed. Both methods are applied to some real data sets. It is found that the all-pos method can give maximal results in terms of trace information. However, since computing time grows exponentially with number of variables, the all-pos method is prohibited when number of items is very large, say 50. In such situations, the BI-method can provide a fast but still good alternative. Predictive validities and reliabilities of shorten scales are discussed.

On Full And Limited Information Statistics For Goodness-Of-Fit Of Sparse 2^p Contingency Tables

Shing-on Leung

In the analysis of 2^p contingency tables, Bartholomew & Leung (2002) have proposed a Y-statistics based on two-way margins to judge the overall goodness-of-fit. Maydeu-Olivares and Harry (2003) has however pointed out that this is a limited information statistics as it only utilized the second order margin. Leung (2005) has suggested a full information statistics based on the one-to-one mapping between cells and margins suggested by Maydeu-Olivares and Harry (2003). This paper applies the full and limited information statistics to the analysis of real data, the sexual attitude data previously analyzed by many literatures. Contributions of the new statistics are discussed.

A Functional Data Analysis of Event Data and Their Intensities

Carlotta Ching, Ting Fok and James O. Ramsay

A functional data analysis model is developed for longitudinal and functional data from events that occur at random times. The model examines the event intensity, or the instantaneous rate of occurrence, and estimates the variable values of the data.

Event data are data that occur at random times, collected usually either using hand-held computer or manually by recording the event occurrence. For these events, each observation is the time t_i of an event, and the state x_i of a variable at the time of occurrence. If only event times are recorded, the data are called a *point process*. If both event times and variable values are recorded, the data are a *marked point process*. The variable values are known as the *mark*. The model, derived by the sum of the two log-likelihood functions, estimates the intensity function and the smoothed function for the mark variable simultaneously.

An example would be the application of the model to a set of lupus data that involve the medical histories of 63 lupus sufferers over several years to examine the flare intensity and severity of lupus symptoms for each patient.

Generalization of Canonical Correlation to N-way Arrays

Richard A. Harshman

Canonical correlation assesses the strength and nature of linear relationship(s) between two matrices that share a common set of rows (e.g., ${}^{1}\mathbf{X}_{IJ}$ and ${}^{2}\mathbf{X}_{IJ'},\ J \neq J'$). We consider generalizations that assess (multi)linear relationships between two or more arrays that share one or more subscripts in common (e.g., ${}^{1}\mathbf{X}_{IJ}$, ${}^{2}\mathbf{X}_{IJK}$, ${}^{3}\mathbf{X}_{IJK'}$, ${}^{4}\mathbf{X}_{IJ'K'L}$). There are a number of possible approaches, three of which will be discussed in this talk. (1) The conceptually simplest is to generalize the *method* of canonical analysis. For example, if one starts with the standard method based on $svd({}^{1}\mathbf{X}'\,{}^{2}\mathbf{X})$ one can replace the inner product $\mathbf{C}_{JJ'} = {}^{1}\mathbf{X}'_{IJ}\ {}^{2}\mathbf{X}_{IJ'}$ with a contracted tensor product $\mathbf{C}_{JKJ'K'\ldots} = {}^{1}\underline{\mathbf{X}}_{IJK} \otimes {}^{2}\underline{\mathbf{X}}_{IJ'K'} \otimes \cdots$, and then replace the use of singular value decomposition with a multilinear array decomposition such as N-way Tucker or Parafac. The canonical weight vectors that were obtained from the singular vectors of the SVD can be obtained in the generalized method from the columns of the “factor loading” or weight matrices of the array decomposition. (2) One can generalize the statement of the canonical relation to incorporate canonical-weight vectors applied to several modes/subscripts of each array, optionally in combination with a small N-way “core” array (similar to that in Tucker models) that integrates the composites. This type of canonical relation can be written compactly in a “tensor”-type notation such as the Index Formalism described in J. Chemometrics, 2001. For example, a simple canonical relation between three-way arrays ${}^{1}\underline{\mathbf{X}}_{IJK}$ and ${}^{2}\underline{\mathbf{X}}_{IJ'K'}$ could be ${}^{1}\underline{\mathbf{X}}_{IJK}\,{}^{1}\mathbf{W}_{JS}\,{}^{1}\mathbf{W}_{KT}\,{}^{1}\underline{\mathbf{G}}_{STR} = {}^{1}\mathbf{Y}_{IR} \approx$ $\approx {}^{2}\mathbf{Y}_{IR} = {}^{2}\underline{\mathbf{X}}_{IJ'K'}\,{}^{2}\mathbf{W}_{J'S}\,{}^{2}\mathbf{W}_{K'T}\,{}^{2}\underline{\mathbf{G}}_{STR}$ Here, the columns of the **Y** s are the canonical variates, the columns of the **W**s hold the canonical weights and the two **G** arrays are the “cores”. (3) The most conceptually satisfying approach is to generalize the definition of the canonical relationship itself. An appropriate generalization should be equivalent to canonical correlation in the two way case but when applied to arrays be sensitive to (and describe) the greater richness of internal and external structural relations possible because of their multilinearity. One possible generalized relation is canonical *cross-fitting*, in which optimal composites from one data array form (parts of) a best-fitting model of a canonical object (a multiply-subscripted canonical variate) abstracted from the other array(s). Alternating Least Squares can estimate weights for such cross-fitting. Work with these tensor-like objects and relations is facilitated by Bader and Kolda’s Tensor Toolbox for MATLAB, recently released by Sandia National Laboratories.

To Compare the Special and Regular Education Teachers’ Attitudes toward Children with Chronic Illness in 1995 and 2002 in Taiwan, R.O.C.

Chiu-Hsia Huang

The purpose of this study was to compare the special education and general education teachers’ levels of acceptance toward of chronically ill children (CWCI) in 1995 and 2002 in Taiwan, R.O.C. Approximately, in 1995 and 2002, one thousand one hundred and ten, residing in seven different educational settings, served as participants by randomly sampling distribution completed in both two studies. The means and standard deviations of the subject responses to each of the eight conditions (asthma, quadriplegia, diabetes, leukemia, big seizure disorder, hepatitis B, technology assisted child, and AIDS/HIV), and total by type of teacher preparation, level of teaching experience, and personal contact experience or without personal contact experience with CWCI were explored across eight chronic illness conditions using a five point Likert type rating scale. The alpha level was 0.05 for rejection of the null hypothesis in Huang’s both studies. The results indicated that the most acceptable CWCI by all teachers was children with leukemia; the less acceptable CWCI by all teachers was AIDS/HIV (Huang 1995 & 2002, and Sullivan’s).

On the use of Person-Fit Methods in Personality Assessment

Wilco H.M. Emons

Person-fit analysis is a psychometric approach for the detection and diagnosis of item-response vectors that are unlikely under the IRT model that describes the data. In this presentation, I will consider the use of person-fit statistics in the analysis of personality data. Personality assessment is different from cognitive assessment in that it involves measurement of typical performance on several broad personality domains, typically using multiple sets of Likert scales. Cognitive testing, on the other hand, is concerned with measurement of maximum performance, typically by means of a large set of dichotomously scored items. The person-fit approaches to be discussed are based on nonparametric and parametric item-response theory models. In this presentation, I will not only focus on the use of person-fit methods to detect aberrant item-score vectors in personality data, but also on how person-fit methods can be used to obtain additional information on the psychometric properties of a test. Examples include model-fit assessment and evaluation of the comparability of measurements across groups (e.g., clinical population versus the general population). Results from simulation studies, as well as real data examples from health psychology will be presented.

Bayesian Non-Parametrics For Latent Variable Models

Matthew S. Johnson

Many models developed to study multivariate data posit a low-dimensional latent variable and a mixing distribution to explain the dependence structure between the observable or manifest variables. Most implementations of latent variable models assume that some function of the mean vector is a linear function of the latent variable and that conditional on knowledge of the latent variable, the dimensions of the manifest variable are independent. However, nonparametric versions simply assume that the relationship is monotone. Monotone latent variable models imply a number of properties on the conditional distributions of each subset of manifest variables given the remaining manifest variables (Holland and Rosenbaum, 1986). This paper approximates these conditional distributions by utilizing a Markov chain Monte Carlo algorithm that augments the observed data with data from the conditional posterior predictive distribution to simplify the algorithm. The paper goes on to demonstrate the efficacy of the approach utilizing data from a large-scale assessment and discusses how posterior predictive checks can be utilized to examine the validity of the latent variable model assumptions non-parametrically.

Nonparametric Methods For Detecting Heterogeneity In Logistic Models

Katalin Balázs, Paul De Boeck

In most cases the binary person-by-item data in the context of Item Response Theory are correlated, or in other words stated, extra binomial variation occurs. Heterogeneity of the parameters can explain this variation. There is a need for proper statistical methods to indicate possible extra heterogeneity and its location, since investigating all different combinations of random parameters is not very practical or sometimes even unfeasible. The random person effects are the focus of this study.
We will study different sources and degrees of heterogeneity in order to investigate and to compare nonparametric exploratory methods for detecting heterogeneity. At present, DIMTEST, a method for testing the unidimensionality of a test, and DETECT, a method to reveal the dimensionality structure of the data, are prominent methods for revealing the latent structure of data. We will report on a simulation study with data generated with several different underlying structures, differing as to the kind and degree of heterogeneity. The diagnostic performance of DIMTEST and DETECT for detecting heterogeneity under different data structures will be compared.

Using the BTL model to validate a product acquisition pattern detected by mokken scaling analyses

Maarten Terpstra

The dynamics of consumer financial product portfolios are an important research topic in the financial industry. One line of research is aimed at the detection of prototypical product acquisition patterns by means of, for example, mokken scaling analyses on data of product possession. Ultimately these analyses yield an ordering of financial products, which are interpreted as the prototypical product acquisition pattern of consumers.
This paper deals with the validation of a particular mokken scale by means of the application of the BTL-model on data of product acquisition. First, based on data of the timing of product acquisition, a so called stimulus comparison dataset has been constructed. Second, by means of logit modelling, the BTL-model has been applied. The analyses yielded a ratio scale which fits the data well, which implies that the original dataset can be reproduced on basis of the scaling values. Most important, the analyses yield evidence that the original mokken scale represents the prototypical acquisition pattern indeed.

The Basic Local Independence Model in Knowledge Space Theory and Axioms / Properties in Nonparametric Item Response Theory

Ali Ünlü

The basic local independence model (BLIM) is considered in knowledge space theory introduced by Doignon and Falmagne as an explanation for binary response data. The BLIM is a restricted latent class model, a probabilistic version of the deterministic knowledge structure model (and Guttman model), which incorporates (errors of measurement) careless error and lucky guess probabilities as explicit model parameters. In this work, we investigate under which (ordinal) restrictions on the error rates of the BLIM, certain axioms and measurement properties in nonparametric item response theory – monotonicity, nonintersection (invariant item ordering), monotone likelihood ratio, stochastic ordering of the latent trait, and stochastic ordering of a manifest variable – are fulfilled for the BLIM.

Symposia and sessions, 14.20h-16.00h

Normal Ogive Model as a Working Hypothesis

Fumiko Samejima

Samejima (2000) has pointed out that any mathematical model with point-symmetric item characteristic functions for dichotomous responses has an intrinsic contradiction in the principle in ordering the maximum likelihood estimates of the latent trait θ , based on response patterns or sequences of binary item scores. As a solution of this problem, models that have point-asymmetric item characteristic functions were considered, and the logistic positive exponent family of models, in which the third parameter, called acceleration parameter, selects one of the two opposing principles, was proposed. The normal ogive model is a point-symmetric model that was proposed on the early stage of latent trait models, and, together with its approximation, logistic model, it has been used widely in psychological measurement. The rationale behind the normal ogive model is introduced in Load & Novick's "statistical theories of mental test scores," (1968), with two assumptions that
1) the regression of item response tendency on θ is linear, and
2) the conditional distribution of the response tendency, given θ , is normal, with the equal standard deviations for all fixed values of θ . The present paper illustrates the fact that, in spite of the deficiency of the normal ogive model that provides point-symmetric item characteristic functions, the model provides good working hypotheses in various situations. For example, a suggestion of using the normal ogive model in place of the three-parameter logistic model, that has been most widely used in computerized adaptive testing , is included with suggestions of some modifications in the selection of a new item from the item pool.

Multidimensional Extensions of the Nominal Response Model for Distractor Analysis

Daniel Bolt, and Timothy Johnson

Multidimensional extensions of Bock's nominal response model are used to investigate individual differences in distractor selection in a multiple-choice test. Analyses of a 23-item sentence correction test identify several dimensions associated with the selection of a "No correction needed" distractor across items. Logit and probit-based implementations of the multidimensional model are compared, and issues in the estimation of the models using simulation-based Bayesian and maximum likelihood methods are considered. Results from the analyses provide several considerations for the design of future versions of the test.

A Two-Level Model for Speed and Accuracy on Test Items.

Wim J. van der Linden

The current literature on test theory shows two fundamentally different approaches to modeling response times on test items. The first approach is to adopt time parameters in a regular IRT model. The reason for this approach is an assumed interaction between the parameters that govern the person's responses and response times, for example, an interaction constrained by the well-known speed-accuracy trade-off. In the second approach, such interactions are ignored and the response times are modeled assuming a stochastic process (e.g., Possoin process) operating independently of the responses.
This paper is based on a third approach, with two different levels of modeling each consisting of two different models. At the level of a fixed person, the distributions of the responses and response-times are modeled separately as independent variables ("local independence"). However, for a population of persons, ability and speed are expected to be statistically dependent, and we introduce a second-level model to represent the dependency (population model). Likewise, for the domain of items from which the test is constructed, item difficulty and time intensity are expected to be dependent. This type of dependency is represented in a separate second-level model (item-domain model).
In the presentation, we will first show how to estimate the parameters and validate the model, and then discuss the application of the model to test design, item selection in adaptive testing, item calibration, as well as the experimental study response times in psychology.

A Comparison Of Traditional And An IRT-Based Scoring Rule For Time-Limit (Intelligence) Tests

Margo G.H. Jansen, H. Kuyper and M. P.C. van der Werf

Many tests that are intended to be power tests are actually more or less speeded because a considerable number of people are unable to attempt every question. The particular strategy selected for scoring missing item responses may have considerable impact on the interpretation of the results both on individual and on group-level. In this paper we compare the traditional scoring rule of counting the number of correct with alternative scoring rules based on an IRT model. For this purpose we used two subtests of the PSB-battery, the PSB-3 and the PSB-8. It was found that the correlations between number correct on the one hand and the IRT based accuracy estimates on PSB-3 and PSB-8 on the other were quite high (.85 and .91). The correlation between the accuracy estimates on PSB-3 and PSB-8 was only slightly higher than the correlation between the number correct scores. Both were relatively low.
The effect on the predictive validity of differences in scoring procedures was studied by calculating correlations with scores on a language, an arithmetic and an information test as external criteria. Rather surprisingly, with regard to the predictive validity the number correct scoring rule was superior for the PSB3, whereas the accuracy scoring rule was somewhat better for the PSB8.

Scale Development for Maximal Power with Congeneric Components

Tenko Raykov

The relationship between statistical power and measurement error for congeneric components is discussed in a class of covariance structure models. This relationship is used to develop a strategy for development (construction/revision) of multiple-component measuring instruments associated with highest power when testing particular hypotheses about latent relationships (cf. Raykov & Hancock, 2005). The strategy is also shown to lead to an instrument with maximal validity (cf. Penev & Raykov, 2005). Testing minimal effect hypotheses, as an alternative to 'null hypothesis significance testing', is outlined based on these relationships within the considered model class.

References:

Raykov, T., Hancock, G. R. (2005). Examining change in maximal reliability for multiple-component measuring instruments. *British Journal of Mathematical and Statistical Psychology* (in press).

Penev, S., Raykov, T. (2005). On the relationship between maximal reliability and maximal validity for linear composites. *Multivariate Behavioral Research* (in press).

A preview of new capabilities of the CALIS procedure for structural equation modeling

Yiu-Fai Yung

The CALIS procedure in the SAS/STAT software package is a statistical analysis tool for structural equation modeling, including error-in-variable models, path analysis and factor analysis. In this talk, we will preview some new CALIScapabilities such as a variety of modeling languages, detailed effect analysis, ad-hoc correlation model analysis, simultaneous parametric tests, standardized resultsand etc. While attempts have not been made to incoporate most recent statisticaltechniques, proposals of such techniques by audience are encouraged.

Model Specification Search with Genetic Algorithm for Practical use

Hiroto Murohashi and Hideki Toyoda

Many techniques for model specification search in structural equation modeling (SEM) have been proposed, but there is no single decisive method yet. Recently, there's been an increase of research about applying metaheuristics for model fitting task. Metaheuristics is a generic term for the optimization algorithms which are based on local search algorithms but improved to be more accurate. In preceding studies, metaheuristics showed nice result in simulation studies. But there's no actual study which put metahuristic search to practical model specification seacrch in SEM. In this study, we focused attention on the situation of real data analysis, and implemented model searching algorithm based on a genetic algorithm (GA). Through simulation study and example analysis using real data, the effectiveness and applicability of proposed method were illustrated.

Measuring Unbiased Concentration Performance Due To Negative Affect By Use Of The Frankfurt Adaptive Concentration Test (FACT)

Tim Lossnitzer and Andreas Frey

Previous research pointed to the fact that problematic bias in measures of concentration performance due to participants´ activation (Frey & Moosbrugger, 2004) and due to noise (Lossnitzer & Frey, 2004) can be controlled when concentration performance is measured with the adaptive versions of the Frankfurt Adaptive Concentration Test (FACT, Moosbrugger & Heyden, 1994). To examine if bias in measures of concentration performance due to negative affect can be controlled in the same way, 103 participants were given either an adaptive or a non-adaptive version of the FACT and negative affect was measured via self-ratings. A multi-sample LISREL model with unconstrained correlations between the latent variables ‘concentration’ and ‘negative affect’ for the adaptive and the non-adaptive sample fits the data (χ^2=6.86, df=7, p=.443). This model is compared to a model with the correlation between the two latent variables constrained to be equal for both samples. The result of the model comparison (χ^2=6.49, df=1, p=.011) suggests the relationship between the performance in the concentration test and negative affect depends on the test version used. Thus, bias in measures of concentration performance due to negative affect can be controlled when the adaptive version of the FACT is used.

Summation and Matrix Operations Package for Mathematica

Shin-ichi Mayekawa

Psychometricians or statisticians often use matrix and summation expressions when they develop their models. However, existing programs for symbolic mathematics have difficulties handling matrix and summation operations. In this paper, a Mathematica package which can do the following was developed:

1) Fully Symbolic Matrix Operations
    1. Simplification of Matrix Expression
    2. Simplification of Partitioned Matrix Expression
    3. Conversion of Matrix Expressions to Summation Expressions
    4. Derivative of a Scalar Function of Matrices w.r.t. a Matrix

2) Fully Symbolic Summation Operation
    1. Simplification of Summation Expressions
    2. Conversion of Summation Expressions to Matrix Expressions
    3. Derivative of a Summation Expression w.r.t. a Subscripted Variable

The package can be used for both practical and educational purposes.

Symbolic Computation in Generalized Structured Component Analysis

Yoshio Takane, Heungsun Hwang

Generalized structured component analysis (GSCA) has been proposed for path analysis with latent variables defined as exact linear combinations of observed variables. In contrast with the partial least squares (PLS) approach, GSCA systematically minimizes a single global loss function by an alternating least squares algorithm. Most often, simple analytic forms of updating equations, which significantly cut down the computation time, can be derived through symbolic manipulations of the loss function. However, updating equations typically depend on particular path models being fitted. To automate this process, we may use SUMOPack, Mathematica based symbolic matrix manipulation tool developed by Shin-ichi Mayekawa (presentation at this meeting). In this paper, we demonstrate possible uses of this package to develop an efficient computer program for GSCA. Examples are given to illustrate the method.

Generalized Constrained Redundancy Analysis

Yoshio Takane and Sunho Jung

We propose a method of generalized constrained redundancy analysis (GCRA) that incorporates external information in redundancy analysis (RA). In this method, both the criterion variables (Y) and the predictor variables (X) in RA are first decomposed into several components according to the external information, and RA is applied to the decomposed components. By combining the terms in the decomposition of Y and those in the decomposition of X, a variety of existing and new breeds of RA are realized including a variety of partial (non-partial, partial, semi-partial and bi-partial) and constrained (unconstrained, semi-constrained and bi-constrained) RA. Examples are given to illustrate the method.

The Combined Use of Generalized Structured Component Analysis and Fuzzy Clustering for Capturing Group-Level Respondent Heterogeneity

Heungsun Hwang, Wayne S. DeSarbo, and Yoshio Takane

Hwang and Takane (2004) proposed generalized structured component analysis (GSCA) as a comprehensive alternative method to partial least squares. The parameters of GSCA are currently estimated by pooling data across respondents under the implicit assumption that they all come from a single, homogenous group. However, it is often more realistic to assume that respondents come from heterogeneous groups, so that they are different with respect to their attitudes or preferences. In this paper, GSCA is combined with fuzzy clustering in a unified framework so as to take into account such group-level respondent heterogeneity. An empirical application is presented to illustrate the usefulness of the proposed method.

Nonlinear Principal Components Analysis: Splines in Action.

Nuno Lavado and Teresa Calapez

The title of this talk, besides telling something about its purpose, pretends to be a tribute to Ramsay's and Winsberg's previous work on I-splines.
We begin by making a review of the philosophy behind CATPCA's algorithm. We recall that the use of spline transformations within the loss function allows a global vision of the nonlinear varieties of Principal Components Analysis (PCA) accomplished by CATPCA. To stress the potential of splines, we briefly discuss our own version of the so called and well known "Cylinder Problem". This is an example in which known nonlinear functions are simulated. It is shown that splines are a privileged way to make relationships between variables linear.
We conclude by introducing some open questions on the CATPCA's philosophy: why is the coding matrix still an indicator one? Why not a matrix based directly on other type of fuzzy coding? What's the importance of something that we would like to call "piecewise loadings"?

A Neural Network Model of Metaphor Understanding based on Data Obtained from Statistical Language Analysis

Asuka Terai and Masanori Nakagawa

The purpose of the current study is to construct a neural network model that implements the human-like process of metaphor understanding, based on data obtained from statistical language analysis. This research deals with expression "A like B" as metaphor (simile). In the psychological field, it is explained that metaphor understanding is dependent on a knowledge structure, which could be measured Semantic Differential Method (SD method). However, it is practically impossible to collect enough data to construct a model that covers various metaphors sufficiently by this method, because participants are not able to respond to numerous indices within a limited experiment time. In this research, at first, the knowledge structure is represented using the probabilistic relationship between concepts and their attributes, which is computed from a statistical analysis of language data. Secondly, a computational model of the metaphor understanding process is constructed. Finally, a psychological experiment is conducted to examine the psychological validity of the probabilistic relationship and predictions from the model. The results of the psychological experiment support the probabilistic relationship and the model.

The Continuing Misuse Of The Significance Test

Rink Hoekstra, Sue Finch, Henk Kiers and Addie Johnson

The significance test is without a doubt the most frequently used statistical technique in the behavioral sciences. Nevertheless, this intensively trained technique is often misinterpreted. In 1999, the APA Task Force on Statistical Inference gave some suggestions on how to use (or avoid the use of) significance testing, and emphasized the importance of reporting confidence intervals and effect sizes. We examined whether these recommendations have led to changes in the use of significance testing and confidence intervals in published articles. We checked the occurrence of misinterpretations in 153 articles from *Psychonomic Bulletin & Review*. One of the most striking findings was that many articles, submitted before as well as after APA Task Force report, mistakenly interpreted the absence of a significant effect as proof for the presence of a null-effect. Furthermore, we found that confidence intervals are still not frequently used. These and other findings suggest that the Task Force of Statistical Inference did not effect revolutionary changes in the way researchers publish their research findings.

Concepts, Constructs and the Worlds of Possibility

Keith A. Markus

A theoretical variable such as integrity, conscientiousness, or academic honesty may correspond to either a concept or a construct but the standard idiom does not distinguish the two. One can describe the difference between concepts and constructs in terms of set theory. Constructs extend over actual cases whereas concepts extend of both actual and possible cases. As such, theoretical claims made about, say, integrity as a concept differ from claims about integrity as a construct. The restriction of constructs to a specified population plays a central role in test validation and psychometric analyses aimed at distinguishing constructs from one another. The extension of concepts over possible populations plays a central role in the adoption of non-actual possibilities as goals in making efforts toward systemic change and also in the linking of constructs across populations. The failure of the standard idiom, which conflates constructs with concepts, to provide a vocabulary that captures both population dependent and population independent aspects of variables recommends the modification of that idiom to distinguish concepts from constructs. This distinction suggests various changes in practice such as including the intended population in the names of constructs but not concepts.

Symposia and sessions, 16.20h-18.00h

Testing A Multidimensional Item Response Model For Continuous Data

Oksana B. Korobko and Cees A.W. Glas

Most item response (IRT) models are stochastic models for the responses of persons to items where the effects of the persons and the items on the responses are modeled by separate sets of parameters. Usually, IRT models pertain to discrete responses. In some testing situations, however, the responses are continuous, or the number of response categories of an item is so large that it is convenient to treat the responses as continuous. Mellenbergh (1994) proposed an IRT model to deal with these responses. In the present paper, the model is generalized to a model with multidimensional ability parameters and a maximum marginal likelihood estimation method is presented. Further, test statistics are presented to test the model against differential item functioning, violation of local independence and violation of normality assumptions. A simulation study is presented to assess the power of the tests.
Further, a real data example pertaining to equating of examination packages is presented to illustrate the use of the model in practice.

Reference:

Mellenberg, G.J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research 29,* 223-236 1994.

Linking Continuous and Discrete Ability Models Using the Saltus Assessment Model

Tzur M. Karelitz, Mark R. Wilson, and Karen L. Draney

Assessment models can describe examinee abilities in terms of the degree of proficiency they posses (quantity) or the nature of that proficiency (quality). Therefore, models that use continuous abilities to describe examinees knowledge and skills may benefit from dividing the continuous scale into a theory-driven set of ordered categories. By doing so, users can make qualitative (rather than quantitative) interpretations of the nature of differences in proficiencies between examinees at one location of the continuous scale versus another location. In the current work, the Saltus model is used as a link between continuous ability models and discrete ability models, to assess the diagnostic information gained by substituting part of the continuous ability by one or more discrete skills.

Application Of Different Explanatory Item Response Models For Model Based Proficiency Scaling

Johannes Hartig and Andreas Frey

In educational research, the anchoring of proficiency scales is often achieved by referring to IRT based item parameters. In most cases, thresholds that discriminate between certain levels of competence are identified post hoc after the item difficulties have been estimated empirically. If, however, a priori assumptions about task characteristics determining the item difficulties exist, they can be explicitly incorporated into the measurement model. One model suitable for this type of explanatory measurement is the Linear Logistic Test Model (LLTM). Another explanatory model is a LLTM including random item effects (LLTM+e) (Janssen, Schepers & Peres, 2004). The task parameters estimated within these models can both be used to define thresholds between levels of competency on the proficiency scale. The LLTM and the LLTM+e were applied to empirical data from a test for English as a foreign language (Nold & Rossa, 2005) used in a nationwide German study. A priori defined task characteristics were used as predictors for item difficulty in a LLTM and a LLTM+e. Additionally, item difficulties were estimated in an unrestricted Rasch model. The results are compared in terms of model fit, their implications for scale anchoring, and differences in the resulting item and person parameters.

On the Factor Structure of Standardized Educational Achievement Tests

Tim Gaffney

This research analyzed the factor structure at both the item and parcel level of California's norm and criterion referenced standardized educational achievement tests (SEAT) used in that state's high-steaks educational accountability assessments. It will be shown through full information factor analysis (e.g., NOHARM and Testfact) that at the item level, SEATs are highly unidimensional (i.e., they appear to tap a general test-taking ability) even when items representing broad content areas such as English, science, mathematics, and history are analyzed simultaneously as a single measure. These item- level factors also account for a relatively small proportion (1/4 to 1/3) of the variance. It will also be shown that, when items representing each content domain are combined into parcels, a richer bi-factor structure emerges that accounts for a larger portion (2/3) of the variance, in which the general and combined specific factors roughly split the common factor variance. The meaning of this lack of correspondence between the item and parcel level structures will be discussed (e.g., well articulated structure at the parcel level does not necessarily imply correspondingly clear structure at the item level), as well as the diagnostic and remedial implications of these tests' factor structure for educators and test forms developers.

Comparability of Construct across Item Formats in the Third International Mathematics and Science Study: The Case of Singaporean Students' Performance

Kim H. Koh, and Hwei Ming Wong

In many of the large-scale international assessments (e.g., TIMSS, PISA), both multiple-choice and constructed-response item formats are used to assess student achievement. In terms of measuring learning outcomes, it is widely accepted that multiple-choice items are limited to measuring factual knowledge and simple recall skills. On the contrary, constructed-response items are more effective for assessing deep understanding of content knowledge and higher-order thinking skills. However, some researchers have shown that multiple-choice and constructed-response items measured the same basic trait or proficiency. Based on the recent released TIMSS 2003 reports, Singaporean students were among the top performers in both mathematics and science at the 4th and 8th grade levels. But yet little is known about the effects of item formats on the Singaporean students' performance. Are the multiple-choice and constructed-response items tapping the same cognitive domains? This study will report the results of the construct comparability of the multiple-choice and constructed-response items in the TIMSS 2003 mathematics and science tests, Singaporean Grade 8 population. The utility of the Multi-Group Confirmatory Factor Analytic tests of measurement invariance/equivalence and Item Response Theory methods in examining the comparability of the factor structures and item parameter invariance across item formats will also be discussed.

Model to Extract Original Sources from the Variant Documents

Maki Miyake, Hiroyuki Akama, Nobuyasu Makoshi, and Masanori Nakagawa

In this paper, we propose a quantitative model to extract original sources from the variant documents. The model is evaluated by simulating a process of intertextuality between the supposed original texts and their derived and modified versions. The first step is to create some artificial texts, which we assume to be dummy "original texts"(in abbreviation, Os), and some "variant documents"(Vs) by randomly distributing all the words contained in each of the Os. Secondly we try to see whether our proposed model is able to extract the traces of the Os in the form of factors, by applying factor analysis to the lexical frequency data gathered from all the subsets of the joint Vs. Based on the result of the simulation, we apply this model to the synoptic problem, which is about the genealogical interdependence between the Synoptic Gospels as one of the controversial subjects in the New Testament (NT) studies. As the result of application of the model, we come to be led to a new possibility for the different original sources from the existing assumption of the dominant sources in the NT studies.

Re-examination of a Social Development Model Constructed via Principal Component Analysis

Masayuki Kuwata, Asuka Terai, Seiji Ohsawa, and Masanori Nakagawa

Latest social indicators (child-bearing, education, health, etc.) compiled by the United Nations was used to confirm our previous model for the social development. Principal component analysis analyzes 12 variables of 124 countries and derives two components which reflect the social development and the natural increase as our previous research revealed. Plotted countries by the scores of these two principal components on a two dimensional plane draws a curve that the previous also extracted. However, current analysis shows differences. First, the drawn curve is thicker than the previous one. Second, the eigenvalue's proportion of the first principal component is lower than the previous. Further, the difference of the eigenvalue between the second principal component and that of the third is close. These differences would imply the third dimension's emergence. Thus, we distilled the third principal component and divided the countries into two parts one of which shows relative higher and the other does relative lower scores of this component. The higher score countries' group draws the outer line of the original curve and the lower score ones does the inner line. Our goal is to elucidate the attribution of the third principal component and construct a new social development simulation model.

The Application of Item Response Theory to the Development of a Gender-Balanced Short Version CPAI-2 Depression Scale

Shu Fai Cheung, Wai Chan and Fanny M. Cheung

The item response theory (IRT) was applied to develop a short version of the 20-item Depression scale of the Cross-Cultural (Chinese) Personality Assessment Inventory-2 (CPAI-2). Based on the IRT analysis and the differential item functioning (DIF) results in a normative sample of 1,911 Chinese participants from China, 15 items were identified to function similarly for both males and females. The item parameters of these 15 items were then examined, and five items that were most informative for persons with high depression scores were selected. Sensitivity and specificity analysis of the preliminary short versions were conducted on the normative sample and another sample of over 300 Chinese major depression patients. The performance of the short version was compared to the full gender-balanced version. The potential advantages of using the IRT to develop customized assessment form were discussed.

Psychometric Evaluation of the CSUQ Using Data from an Internet Touch Point Experiment

B.L. Olson , J.A. Safar , C. Turner, and M. Zavod

The submitted paper reflects part of an ongoing research effort to develop a set of metrics to describe the "customer experience" associated with insurance and financial services. This study examined the psychometric properties of the Computer System Usability Questionnaire (CSUQ) in its application to assessing user satisfaction of interactive, consumer-based websites. The primary research question was the degree to which the CSUQ, used for research in an area different from that for which it was developed, would exhibit similar factor structure, reliability, validity and sensitivity as found in other contexts. A two-condition, between-groups design was implemented using four hundred web panel participants. Participants completed an online auto rate quote using one of two selected web sites and subsequently rated their experiences using the CSUQ. The findings from this study provide evidence of further generalizabiltiy and usability of the CSUQ instrument.

Authors Index