The Third Workshop on Statistical Methods in Cognitive Assessments

2014 Program

Short courses: May 27-28

Workshop: May 29-31

Shanghai Center for Mathematical Science,

Fudan University, Shanghai, China

SHORT COURSE SCHEDULE

Tuesday, May 27	Room 2201, Guanghua East Building, Fudan University
Morning Session	
9:00 - 11:30	Jimmy de la Torre, Rutgers University Young-Sun Lee, Columbia University
Afternoon Session	
13:30 - 17:00	Jimmy de la Torre & Young-Sun Lee
Wednesday, May 28	Room 2201, Guanghua East Building, Fudan University
Afternoon Session	
13:30 - 17:00	Chun Wang, University of Minnesota Ya-Hui Su, National Chung Cheng University Ping Chen, Beijing Normal University

WORKSHOP SCHEDULE

Thursday, May 29	Room 2201, Guanghua East Building, Fudan University
Morning Session	
10:00 - 10:30	Hua-Hua Chang , University of Illinois and East China Normal University
10:35 - 11:05	Matthias von Davier, ETS
11:10 - 11:40	Jinming Zhang, University of Illinois
Afternoon Session	
14:00 - 14:30	Daniel Bolt, University of Wisconsin Madison
14:35 - 15:05	Gongjun Xu, University of Minnesota
15:10 - 15:25	Coffee Break
15:30 - 16:00	Chun Wang, University of Minnesota
16:05 - 16:35	Francis Tuerlinckx, University of Leuven
Friday, May 30	Room 2201, Guanghua East Building, Fudan University
Morning Session	
10:00 - 10:30	Jimmy de la Torre, Rutgers University
10:35 - 11:05	Ya-Hui Su, National Chung Cheng University
11:10 - 11:40	Ping Chen, Beijing Normal University
Afternoon Session	
14:00 - 14:30	Wen-Chung Wang, Hong Kong Institute of Education
14:35 - 15:05	Hong Jiao, University of Maryland
15:10 - 15:25	Coffee Break
15:30 - 16:00	Sophia Rabe-Hesketh, University of California, Berkeley
Saturday, May 31	Room 2201, Guanghua East Building, Fudan University
Morning Session	
10:00 - 10:30	Tao Xin, Beijing Normal University
10:35 - 11:05	Jian Tao, East China Normal University
11:10 - 11:40	Matthew S. Johnson, Columbia University

Short Course Abstracts

Tuesday, May 27, Pre-Workshop Short Course

[9:00 - 17:00]

Jimmy de la Torre and Young-Sun Lee

Cognitive diagnosis modeling: A general framework approach

The primary aim of cognitive diagnosis is to develop and analyze assessments that provide information with more diagnostic value compared to traditional approaches. Its main objective is to identify individual students' specific strengths and weaknesses in a particular domain. Using a general framework, this workshop aims to provide both the theoretical underpinnings and practical experience necessary for participants to use cognitive diagnosis modeling (CDM) in applied settings.

The theoretical component of the workshop will provide participants a comprehensive overview of CDM. Topics covered will include an introduction to the CDM paradigm and how it differs from traditional unidimensional frameworks, some commonly used cognitive diagnosis models and their relationships to each other, estimation of model parameters, evaluation of model-data fit, and model comparison and selection.

The practical component of the training session will provide participants a hands-on experience on the different aspects of CDM through various exercises. Participants will learn how to identify and validate attributes, run and interpret result of codes for CDM, evaluate the appropriateness of cognitive diagnosis models, and empirically validate Q-matrices.

Wednesday, May 28, Pre-Workshop Short Course

[13:30 - 17:00]

Chun Wang, Ya-Hui Su, and Ping Chen

Introduction to multidimensional item response theory

There has been a tremendous amount of progress in item response theory (IRT) in the past two decades. Multidimensional item response theory (MIRT) is a special case of IRT that assumes the item response function includes as parameters a vector of multiple person characteristics that describe the skills and knowledge a person brings to a test and a vector of item characteristics that describes the difficulty of an item and the sensitivity of an item to differences in the characteristics of the persons (Reckase, 2009).

MIRT also has a long history, going back to the work of Darrel Bock, Paul Horst, Roderick McDonald, Bengt Muthen, Fumiko Samajima, and others starting in the 1970s. The MIRT model gains popularity in a recent decade because of the availability of high computation power and the broad applications of MIRT models in cognitive diagnosis. The goal of this short course is to draw together the different aspects of MIRT, including various MIRT models, item and person parameter estimation, and analyzing the multidimensional structure of a test. In brief, this short course will introduce users to the conceptual/theoretical knowledge of MIRT and provide valuable hands on experience with the MIRT model estimation/applications via **R**.

Workshop Abstracts

Thursday, May 29, Morning Section

[10:00 - 10:30]

Hua-Hua Chang

CD-CAT and adaptive learning

A growing body of evidence shows that Computerized Adaptive Testing (CAT) and Cognitive Diagnostic (CD) methods have enormous potential to revolutionize classroom assessment and greatly facilitate individualized learning. In a one-to-one instructional environment, the content and pace of instruction can be completely customized to best fit the observed progress of a particular student allowing the teacher to better focus on the individual's specific needs and problems. This paper will show how CD-CAT can be used to expedite such teaching on a mass scale.

[10:35 - 11:05]

Matthias von Davier

Conjunctive attributes, linear hierarchies - do they promise more than they can deliver?

Two examples are given how the current practice of deriving more complex diagnostic modeling approach may get in the way of studying and selecting models with few assumptions that fit the data.

1) There is a lot of rhetoric regarding the assumption of conjunctive skills, mainly based on the assumption that if a conjunctive model fits some data from psychological

or educational measurement, this shows that the assumption of a compensatory attribute space is futile. However, it was recently shown that the 'deterministic-input noisy-and' (DINA) is equivalent to a special case of a more general compensatory family of diagnostic models. The equivalencies solely require a linear - compensatory - general diagnostic model (GDM) without any skill interaction terms.

2) A similar point is made about hierarchical attribute spaces recently discussed in modeling diagnostic assessments. Hierarchies are assumed to exist if the application of one attribute requires the mastery of a less complex skill or attribute. Proponents of this approach posit that this will allow deriving the cognitive structure and help researchers gather more information than traditional approaches to the analysis of behavioral data. It is shown that the assumption of linear hierarchies severely limits the ability to distinguish these models from simple uni-dimensional approaches

The talk closes with a few notes and references to work that delineates how much knowledge we can realistically derive about latent variables from data that provides a finite number of binary indicators collected on a sample of test takers.

[11:10 - 11:40]

Jinming Zhang

Integrating substantive and statistical evidence in Q-matrix construction

A Q-matrix stipulates the relationship between test items and attributes (which attributes are required by each item) and is the foundation of all subsequent data analysis. Typically, the framework of a cognitive diagnostic assessment (CDA) specifies the attributes/skills the CDA aims to measure. The item writers develop test items according to the framework. Then, one or more subject experts (called raters) construct an initial Q-matrix based on item substantive evidence. Finally, psychometricians modify the Q-matrix according to statistical evidence based on the corresponding response data. The resulting data-driven Q-matrix is generally statistically solid, but may not be substantively meaningful for some items/attributes. Thus, to finalize a Q-matrix for a CDA, a compromise has to be made between statistical and substantive perspectives. There are three shortcomings in this approach to creating a Q-matrix: (1) information obtained during the process of constructing the initial Q-matrix, other than the Q-matrix itself, is generally ignored; (2) rater reliability is typically not examined; and (3) the final Q-matrix may not be optimal due to the compromise.

It is recommended to use multiple raters. After necessary training, raters should work independently to produce all substantively meaningful Q-matrices. The set of all these Q-matrices from all raters is called the Q-matrix universe. Ideally, the universe should contain all admissible (i.e., substantively meaningful) Q-matrices. At this stage, generalization theory can be applied to investigate the dependability of raters and items. Then, the raters may work together to agree upon one or several Q-matrices (called seeds), which may be useful later, especially when the number of Q-matrices is large. The idea is to use statistical evidence from the data to select an optimal Q-matrix only from the universe so that the statistically optimal one must be substantively meaningful. The criteria (e.g., model fit) for the optimization should be carefully selected. If the number of Q-matrices in the universe is limited in the sense that the calculation burden is not an issue, one may search the whole universe for the optimal Q-matrix. Otherwise, a selection algorithm (e.g., genetic or evolution algorithm) is needed to select the optimal or near-optimal Q-matrix by using the seeds as initial Q-matrices.

Thursday, May 29, Afternoon Section

[14:00 - 14:30]

Daniel Bolt

Item complexity and Samejima's logistic positive exponent (LPE) model

Item complexity is an item feature that is increasingly considered in the design of test items and the construction of tests. Less frequently considered is the implications complexity may have for the psychometric modeling of test items, especially in cases where the test is largely unidimensional. In this talk we consider practical issues in the application of Samejima's logistic positive exponent (LPE) model as a basis for accounting for item complexity, as well as some consequences of ignoring item complexity when in fact it is a feature that varies across items.

[14:35 - 15:05]

Gongjun Xu

Statistical analysis of Q-matrix based diagnostic classification models

Diagnostic classification models have recently gained prominence in educational assessment, psychiatric evaluation, and many other disciplines. Central to the model

specification is the so-called Q-matrix that provides a qualitative specification of the item-attribute relationship. In this talk, we develop theories on the identifiability for the Q-matrix under the DINA and the DINO models. We further propose an estimation procedure for the Q-matrix through the regularized maximum likelihood. The applicability of this procedure is not limited to the DINA or the DINO model and it can be applied to essentially all Q-matrix based diagnostic classification models.

[15:30 - 16:00]

Chun Wang

Multidimensional analysis of student change/growth using item response theory

In educational testing, reporting overall ability for accountability purposes along with finer-grained domain scores for diagnostic purposes has become standard. Two item response theory (IRT) models that have been proposed and recently studied extensively for such purposes are the compensatory multidimensional IRT model (MIRT) and the higher-order IRT (HO-IRT) model. While previous research has shown, through simulation studies, that both models can provide reliable overall and domain ability estimates (Yao, 2010), this study analytically shows the common and differential performance of these two models. Furthermore, we consider a novel situation when data are collected from two time points. For this objective, we first extend both the MIRT and HO-IRT models to their longitudinal versions with correlated latent traits, and show when and how the reliability of overall and domain change scores are improved by using the proposed longitudinal models. Theoretical results are derived and a simulation study is conducted to support our findings—in most cases, using longitudinal models by pooling responses from two time points together, both the overall and domain abilities and their change scores can be estimated with higher precision. Finally, multivariate hypothesis testing methods are proposed to statistically check if individual change/growth is significant.

[16:05 - 16:35]

Francis Tuerlinckx

Diffusion-based item response modeling

Psychometrics makes use of elegant and generally applicable measurement models. Cognitive psychology develops mathematical process models for explaining specific

cognitive phenomena. In an ideal world, measurement of cognitive abilities should be based on insights into these cognitive abilities from cognitive psychology. Consequently, measurement models should incorporate the principles of the mathematical cognitive process models. There are many reasons why this is not (yet) the case. In this talk, I will illustrate how mathematical models from cognitive psychology can be used in psychometrics. Specifically, I will make the connection between diffusion (or related) models of decision making and item response models. This is joint work with Dylan Molenaar and Han van der Maas (University of Amsterdam).

Friday, May 30, Morning Section

[10:00 - 10:30]

Jimmy de la Torre

New item selection methods for cognitive diagnosis computerized adaptive testing

This paper introduces two new item selection methods, the modified-posterior-weighted Kullback-Leibler (MPWKL) index and the generalized deterministic inputs, noisy "and" gate (G-DINA) model discrimination index (GDI), that can be used in cognitive diagnosis computerized adaptive testing. The efficiency of the new methods is compared with the posterior-weighted KL (PWKL) item selection index using a simulation study in the context of the G-DINA model. The impact of item quality, generating models, and test termination rules on attribute classification accuracy are also investigated. The results of the study show that the MPWKL and GDI perform very similarly, and have higher correct attribute classification rates or shorter average test lengths compared to the PWKL. In addition, the GDI has the shortest implementation time among the three indices. The proportion of item usage with respect to the required attributes across the different conditions is also tracked and discussed.

[10:35 - 11:05]

Ya-Hui Su

The development of the priority index in computerized adaptive testing

Computerized adaptive testing (CAT) not only enables efficient and precise ability estimation, but also increases the security of testing materials since examinees are given different sets of items from a large item bank. The construction of assessments usually

involves fulfilling a large number of non-statistical constraints, such as item exposure control and content balancing. To improve measurement precision, test security, and test validity, the priority index (PI; Cheng & Chang, 2009; Cheng, Chang, Douglas, & Guo, 2009) and multidimensional priority index (MPI; Yao, 2011, 2012, & 2013) were proposed to manage various constraints simultaneously for both unidimensional and multidimensional CATs. In practice, many educational and psychological tests are constructed under a multidimensional framework. Some of the items (multidimensional items) in a test are intended to assess multiple latent traits. For instance, an arithmetic item can be used to assess both symbolic representation and calculation. Most current constraint control methods, such as Yao's MPI method, were developed under a situation where the multidimensionality is between items. Hence, it is important to propose a modified MPI method for the item selection when a within-item multidimensional test is assembled. This talk will first review the development of the priority index from unidimensional to multidimensional, and then the modified method will be introduced. The results from a set of simulation studies showed that the modified method outperformed the existing method in multidimensional CATs.

[11:10 - 11:40]

Ping Chen

Online calibration in cognitive diagnostic computerized adaptive testing

More recently, there is great interest in how diagnostic testing can be used to inform instructional decisions and improve student learning. Thus, diagnostic tests should provide examinees' latent cognitive profiles on a given set of attributes pertinent to learning and not simply a summative score. By merging the advantages of cognitive diagnosis and computerized adaptive testing (CAT), cognitive diagnostic CAT (CD-CAT) provides each examinee with helpful diagnostic feedback, as well as improves the accuracy and efficiency of cognitive diagnostic assessment, which has received increasing attention in educational measurement.

Like all unidimensional CAT (UCAT) applications, item replenishment is an essential part in a CD-CAT for its item bank maintenance and management, which governs retiring obsolete or overexposed operational items as time goes on and replacing them by new ones (Wainer & Mislevy, 1990). Just like UCAT item replenishing, new items need to be calibrated on the same scale as the operational items, and the calibration

precision of new items will directly affect the estimation accuracy of examinees' attribute mastery patterns. Moreover, online calibration technique is commonly used to calibrate new items in UCAT; and actually in CD-CAT, online calibration also has several compelling advantages over the traditional calibration approach with common-item design. As a result, it is very natural to use online calibration to calibrate new items in CD-CAT.

In the presentation, we start with an introduction to the issues of item replenishment in CD-CAT. It is noteworthy that, in addition to requiring developing new items and calibrating new items as in UCAT item replenishing, CD-CAT item replenishing has a unique step, i.e., asking experts to identify the Q-matrix corresponding to the new items (Q_{new_item}) or estimating it based on statistical methods. Then we briefly introduce the existing online calibration designs (including random and adaptive designs) in CD-CAT. In addition, we elaborate on some online calibration methods (including CD-Method A, CD-OEM and CD-MEM) specially proposed for CD-CAT under DINA model, assuming Q_{new_item} is known and correct. Finally, given the Q_{new_item} is unknown or unidentified, we describe a data-driven joint estimation algorithm (JEA) which depends solely on examinees' responses on the operational and new items to jointly estimate the Q_{new_item} and the item parameters of the new items.

Friday, May 30, Afternoon Section

[14:00 - 14:30]

Wen-Chung Wang

A new class of item response theory models for ipsative tests

There are two major kinds of tests: normative and ipsative. Most tests are normative in that their scores can be compared between persons (e.g., who is more proficient in math, and who has a higher motivation). In contrast, scores of ipsative tests can be compared only within persons (e.g., I prefer artistic activities to enterprising activities). Typical examples of ipsative tests are pairwise-comparison tests and ranking tests, in which persons are requested to choose one statement from a paired of statements or to rank a set of given statements. In the talk, I will introduce a new class of item response theory (IRT) models that we recently developed for ipsative tests and describe relative issues, such as linking design, differential statement functioning, and computerized adaptive testing under the new class of models.

[14:35 - 15:05]

Hong Jiao

A multicomponent testlet model

Testlets, context-based or situation based items are widely used in large-scale assessments. The next generation of assessments calls for innovative items which are most often scenario or situation based items or testlet based items. Innovative items often assess higher order thinking skills which may simultaneously require multiple latent traits in solving problems in real life situations. For such complex assessments, noncompensatory multiple latent traits and the testlet structure need to be considered concurrently. This study proposes a multicomponent testlet model which is a non-compensatory multidimensional testlet model for the next generation assessments where multiple noncompensatory latent traits are required to answer items embedded in testlets. Model parameter estimation is explored using Markov Chain Monte Carlo method in OpenBUGS.

[15:30 - 16:00]

Sophia Rabe-Hesketh

Using Warm's weighted likelihood estimates as response variable: An alternative to plausible values

Datasets from large-scale assessments, such as PISA, NAEP, and TIMMS, contain multiple plausible values of the latent variables that can be analyzed using Rubin's rules for multiple imputation. A drawback of this approach is that the model used to generate the plausible values must include a regression (or conditioning model) for the latent variable that is more general than the model fit by the secondary data analyst. For example, fitting a multilevel model using plausible values generated from a single-level conditioning model can lead to bias in the estimated variance components. We investigate use of Warm's weighted likelihood estimates of the latent variable as response variable instead of plausible values. Uncertainty in the Warm's estimates is taken into account by using extensions of meta-regression. Our method works well if measurement is sufficiently precise and may therefore have promise as large-scale assessments become increasingly adaptive.

Saturday, May 31, Morning Section

[10:00 - 10:30]

Tao Xin

An application of M statistics to evaluate the cognitive diagnostic modeling

The appropriateness of the cognitive diagnostic model (CDM) to the response data needs to be evaluated in support of interpreting respondents' performance and diagnostic test development. However, the most commonly used full information goodness-of-fit statistics χ^2 and G^2 in CDMs become invalid in the case of a large number of items or few respondents (sparse contingency table). Like the alternative methods in the context of item response theory, a counterpart version of limited information goodness-of-fit statistics M (Maydeu-Olivares & Joe, 2005) was used to evaluate the CDMs' model fit when contingency table is sparse. For simplicity, DINA model is taken as an example in this study. Simulations were first conducted to show that M could provide proper empirical type I error rates and good power in a variety of designed conditions. Here the empirical type I error rates and power were used to evaluate the performance of M statistics when DINA model is fit or not, respectively. There are four attributes were examined by 12 items. The associated Q matrix and DINA item parameters for the test were fixed across all simulated sample sizes $N = 250$, $N = 1000$, $N = 2000$ and $N = 4000$. Here the four sample sizes indicate different levels of sparseness evaluated in this study. There 1000 replications were conducted under each sample size to compute empirical type I error rates and power, which were compared at five significance levels: 0.01, 0.05, 0.100, 0.20, and 0.25 among M_2, χ^2 and G^2, respectively. The result showed that type I error rates of χ^2 and G^2 are too liberal or conservative when the sever sparseness occur ($N = 250$). As a matter of fact, M_2 exhibits a reasonable type I error rates as compared to the corresponding nominal significance levels. The statistics becomes an alternative method when χ^2 and G^2 are statistically infeasible. As the sample size is increasing, M_2 performed better.

[10:35 - 11:05]

Jian Tao

A generalized measurement index of the differences in response probabilities with its application in personality measurements

This study defines a generalized measurement index for the differences in response probabilities to an item. Based on the difference index, a probability-difficulty hypothesis is proposed. A general framework for modeling responses and response times (RTs) on Likert-type personality items is presented, in which the sub-model describing the item responses can be a graded IRT model, and the sub-model describing RTs is developed based on the probability-difficulty hypothesis. Meanwhile, this framework is exemplified by employing the generalized partial credit model (GPCM) for responses and a lognormal model for RTs. Furthermore, Bayesian methods for estimating model parameters and for assessing the model-data fit are described. A simulation study shows that the new approach improves the accuracy of estimating the individual trait levels with the ancillary information contained in RTs. Finally, the applicability of our approach is illustrated by an empirical example in personality measurements.

[11:10 - 11:40]

Matthew S. Johnson

A hypothesis testing procedure for Q-matrix entries

de la Torre (2008) introduced a method, which he called the EM based δ-method, for empirically evaluating the Q-matrix within the DINA framework. The method compares the pseudo-empirical proportions of correct responses between groups of examinees that either have or do not have the (proposed) required skills for an item. In this talk I present the asymptotic distribution of these pseudo-empirical differences in the proportions correct, denoted by $\delta_j(q)$. I then investigate, through simulation, the power and Type I error rates for hypothesis testing procedures for the correct specification of the q-vector for a given item, i.e., H_0: $q_j = q_{j0}$, against a simple alternative hypothesis, H_1: $q_j = q_{j1}$, and for testing the correct specification of single entry in the Q-matrix, e.g., H_0: $q_{jk} = 0$. Finally I demonstrate the method by applying it to Tatsuoka's fraction subtraction data.