出國報告（出國類別：其他）

參加第七屆ITC國際學術研討會

服務機關：國家教育研究院籌備處

姓名職稱：曾建銘 助理研究員

派赴國家：香港

出國期間：2010/07/19-2010/07/21

報告日期：2010/10/17

摘要

報告者于2010/07/19-21 赴香港參加由香港中文大學主辦的第七屆ITC (International Test Commission)國際學術研討會，並發表海報論文 ”A Study on Differential Item Functioning (DIF) of The Basic Mathematical Competence Test for Junior High Schools in TAIWAN.” 。ITC-國際測驗學會是由心理組織、測驗學會及其他相關機構與出版社所組成，任務是提昇教育與心理工具測驗的效度、測量政策和適當的發展與使用評量。

ITC 國際學術研討會今年的主題為 ”Challenges and Opportunities in Testing and Assessment in a Globalized Economy.” ，另外五個次子題為：

1. Developments in psychometrics and test theory for international testing
2. Indigenous, second language, and cross national test development
3. Geotrends in testing: making use of technology advances in test administration and data management
4. Issues of policy, ethics, professionalism and training in multinational testing
5. Test security and privacy concerns when testing internationally.

邀請七位學者進行七場專題演講。

本次會議共有全世界教育測驗學者與研究生共629人參與，大會內容如下：專題演講2場次、主題演講5場次、特別主題5場次、鑽石贊助2場次(7篇)、應邀座談11場次(42篇)、座談25場次(106篇)、口頭發表20場次(109篇)、海報發表7場次(152篇)，共406篇。

目次

壹、目的……4
貳、過程……4
參、心得……16
肆、建議事項……16
附件 ……17

本文

壹、目的

一、參加第七屆ITC (International Test Commission)國際學術研討會。
二、發表海報論文 ”A Study on Differential Item Functioning (DIF) of The Basic Mathematical Competence Test for Junior High Schools in TAIWAN”。
三、聆聽28篇的論文發表(含口頭與海報)。

貳、過程

本屆(第七屆)ITC 國際學術研討會第一次在非西方國家舉行,由香港中文大學主辦。場地位於香港中文大學校內，就在港鐵大學站旁邊交通便利。

研討會共舉行三天（7/19-7/21）。於7/19 早上7:30就開始辦理報到，領取會議資料8:30即開始正式進行各項議程。7/19晚上並由大會舉辦歡迎餐會，讓與會的各國學者相見歡。

大會開幕分別由ITC主席Born博士致歡迎詞後，即由荷蘭Maastricht大學教授Roe博士進行專題演講，題目為Testing for travelers: Past and future。本屆研討會的主題為 “Challenges and Opportunities in Testing and Assessment in a Globalized Economy”，專題演講者共有七位：

1. 荷蘭Maastricht大學教授Roe博士，題目為 ”Testing for travelers: Past and future”，論述過去與未來各國文化差異，對測驗所產生的影響。
2. 紐西蘭Auckland大學教授Hattie博士，題目為 ”*Global testing, global opportunities, global challenges, and a global future for assessment*”，論述全球測驗所面臨的問題、轉機、挑戰與未來如何因應。
3. 香港中文大學教授Cheung博士，題目為 ”*From indigenous to cross-cultural personality assessment: the usefulness of the combined emic-etic approach*”，論述跨文化測驗引進後，如何調整與發展成適合於當地測驗的相關議題。
4. 荷蘭Tilburg大學教授van de Vijver博士與南非North-West大學教授Fons博士，題目為” *Recent developments in international testing*”，內容為跨文化心理測驗的翻譯、調整與文化負荷。
5. Kryterion公司Foster博士,題目為 ” Foster *International high-stakes online testing: best practices for test security and data privacy*”， 內容為國際性高風險線上測驗與個人資料的保密性的相關議題，如須考量哪些因素、如何降低風險、保全的新方法等。
6. 美國Florida大學教授Oakland博士，題目為 ”*Ethical and other professional issues: what to do when working in the absence of local standards*”， 內容為提供跨國測驗一些關於倫理、道德、政治等所須注意的規準。

7. 美國Michigan State大學教授Schmitt博士，題目為 ”*Validation support for selection procedures*”， 內容為過去一世紀美國與歐洲效標關聯效度之後設認知(meta-analyses) 研究，提出對不同國家與組織實際上有意義的構念(constructs)與限制。

除了專題演講，大會另安排兩場座談會及各國學者、研究生的口頭發表、海報發表。大會內容如下：專題演講2場次、主題演講5場次、特別主題5場次、鑽石贊助2場次(7篇)、應邀座談11場次(42篇)、座談25場次(106篇)、口頭發表20場次(109篇)、海報發表7場次(152篇)，共406篇，手冊與會共629人。

聆聽28篇的論文發表(含口頭與海報)，茲將對個人研究與TASA、教檢較有影響之主題與摘要臚列如下：

1.*Future directions of testing in the United States*

Chair Hambleton, Ronald K. (University of Massachusetts, USA)

Symposium Abstract

The importance of test uses continues to grow. In the USA today, students from the 3rd grade to high school are administered achievement tests, for a total of about 80 million tests per year. Add several times 80 million tests to account for the diagnostic tests that are being administered to support the assessment of student progress, and it is clear that the growth of testing in the schools has been substantial. Also, the number of admissions tests and credentialing exams continues to grow. At the same time, because of the importance of these tests, and the desire to improve test score validity, technical advances in many directions are occurring. In this symposium, the presenters will focus on three important directions in the USA: First, we will consider the impact of cognitive psychology on testing. The impact has been discussed in the measurement literature for 30 years but now that impact is being seen, and substantial amounts of research are underway. Second, the impact of technology has been increasing, and now it is likely to fundamentally change our approaches to what is measured, as well as to test design and test administration. Advances in technology
will be the focus of the second presentation. Finally, nothing is more important, ultimately, than the production of valid test scores, that can be reported on meaningful score scales, and in ways that they are understood and used correctly by practitioners. Score reporting in the future will be the focus of the final presentation.

A new generation of DIF studies

Chair

Elosua, Paula (University of the Basque Country, Spain)

Hambleton, Ronald K. (University of Massachusetts, USA)

Symposium Abstract

Numerous DIF studies have been published in specialized and applied psychometric journals during the last two decades. In addition to the development of statistical procedures for detecting differential item functioning that are highly efficient in spotting problematic items, the research on DIF to date has also

focused on applications of DIF analyses in a range of testing contexts. All of this work is critical because of the extent to which DIF analyses are a fundamental part of item analysis. However, it is important to note that any analysis of differential item performance should not be narrowly focused on the detection of DIF: once DIF is detected, the task turns to understanding it, the study of effects of item type on examinee performance, or the study of the practical consequences. It is this idea of extending DIF studies with new methods and approaches that forms the basis of this proposed symposium: A new generation of DIF studies. The new perspective involves multilevel latent models, mixed models, consequences and new robust procedures for the detection of DIF. The symposium consists of four presentations given by researchers from four countries. The first study illustrates a new approach to detecting DIF based on using robust statistics ; the second one uses a simulation to evaluate the effects of factorial partial invariance on group comparisons ; the third and fourth presentations incorporate mixture models to evaluate the presence of latent classes and novel applications of multilevel IRT .

2. *"Robust anchoring and posterior anchoring as procedures for DIF and measurement equivalence"*

de Boeck, Paul A. L. (University of Amsterdam, Netherlands)*

An important issue in the process of identifying DIF and also in the process of obtaining measurement equivalence is the choice of anchor items. The basis for this choice is commonly either prior knowledge or iterative purification based on the data. Two alternatives are presented here: (1) robust anchoring, using tools from robust statistics, and (2) posterior anchoring, based on posterior DIF probabilities of the items. The robust approach can be implemented in a parametric way, for example with a robust version of the Raju distance, or in a nonparametric way, for example with marginal proportions correct. The posterior approach requires a mixture model for the items, with a DIF class and a non-DIF class of items. These two alternatives do not require prior knowledge and neither do they make use of iterative purification. They both rely on a one-step statistical procedure. Simulation studies show that their performance is excellent. Apart from their practical use in dealing with DIF and obtaining measurement equivalence, they are also novel IRT approaches in a more fundamental statistical sense.

3. *"The effect of Partial Factorial Invariance on group comparisons"*

Elosua, Paula (University of the Basque Country, Spain)*

Zumbo, Bruno D. (University of British Columbia, Canada)

Factorial invariance studies examine the equivalence among factorial structures across groups. Conclusions about partial factorial invariance mean that some of the model parameters (loadings, thresholds, error variances) are different for groups. It is difficult, however, for a researcher to quantify the effects (i.e., impact) of this lack of invariance on subsequent statistical decisions based on group mean comparisons or coefficient alpha comparisons across groups.

4. *"Latent variable mixture modeling as a method to examine sample heterogeneity, and the related problem of DIF"*

Zumbo, Bruno D. (University of British Columbia, Canada)*

Sawatzky, Richard G. (Trinity Western University, Canada)
Ratner, Pamela A. (University of British Columbia, Canada)
Kopec, Jacek A. (University of British Columbia, Canada)

We will present an overview of a program of research that applies latent variable mixture modeling (LVMM) to examine the extent to which a sample is homogeneous with respect to a specified statistical model for ordered categorical item responses. Along the way we will evaluate the implications of sample heterogeneity with respect to the latent variable scores, and identify potential sources of sample heterogeneity. As has been shown in the literature, LVMM can be used in conjunction with IRT (i.e., an IRT mixture model) to examine sample heterogeneity, and the related problem of DIF, when relevant group differences are not assumed a priori (Cohen & Bolt, 2005 ; De Ayala et al., 2002; Mislevy, Levy, Kroopnick, & Rutstein, 2008; Rost, 1990; Samuelsen, 2008; Vermunt, 2001). Our aims are: (a) to share the lessons we have learned about LVMM, its implementation and limitations, and (b) demonstrate how looking at the typically DIF situation from this vantage point allows us to investigate whether there are other variables than the usual manifest variable in DIF studies (such as gender, age, or nationality), or interactions among variables, that distinguish homogeneous groups. Our focus will be typical psychosocial measures such as emotional wellbeing and physical functioning, and the data complexities they present.

5. *"Applications of multilevel IRT models to investigate item type effects"*

Zenisky, April L. (University of Massachusetts, USA)
Elosua, Paula (University of the Basque Country, Spain)
Zumbo, Bruno D. (University of British Columbia, Canada)*

This presentation focus on a new multilevel IRT model and on its application to study item type effects which can affect the performance across groups. A multilevel IRT model developed for group-level diagnosis was applied to study data from high school end-of-course examinations. Variability in item difficulty across ethnic groups was investigated in relation to item features associated with content and cognitive process categories. Random effects were attached to each feature type at the group level, and their variability studied across groups. The estimated feature effects were shown to provide a basis for examining cross-ethnic differences for individual features as well as cross-feature differences within individual ethnic groups, as this may be useful for diagnostic purposes. The model was fitted using Markov Chain Monte Carlo procedure by R software.

Assessment models for monitoring learning

Chair

Hambleton, Ronald K. (University of Massachusetts, USA)

Symposium Abstract

The symposium will discuss on the models for monitoring teaching and learning in three countries: Denmark, Hong Kong and New Zealand. The three systems will be reviewed and discussed in terms of their influences on learning and teacher autonomy, the stakes associated with assessments, the types of assessments used, the levels of aggregation of data from these assessments, and how data are used.

6. *"National tests in Denmark – CAT as a pedagogic tool"*

Wandall, Jakob (Danish Ministry of Education, Denmark)*

Testing and test results can be used in different ways. They can be used for regulation and control, but they can also be a pedagogic tool for assessment of student proficiency in order to target teaching, improve learning and facilitate local pedagogical leadership. To serve these purposes tests have to be low stake. In Denmark, to ensure this, test results are made strictly confidential by law. The only test results that are made public are the overall national results. Because of the test design (Rasch-model), results are directly comparable, which gives an enormous potential for monitoring added value and developing new ways of using test results in a pedagogical context. The presentation gives the background and status for the development of the Danish national tests, describes what is special about these tests (IT-based, 3 tests in 1, adaptive, etc.), how the national test are carried out and what is tested. Futhermore, it is described who are allowed to know the results, what kind of response is given to the pupil, the parents, the teacher, the headmaster and the municipality and how the results can be used by the teacher and headmaster.

7. *"Alternatives to external standardized assessments: Hong Kong example"*

Hamp-Lyons, Liz (University of Hong Kong, Hong Kong SAR, China)*

In this presentation I will 1) describe the school-based assessment system that has been introduced across Hong Kong secondary education to assess the English speaking skills of all students; 2) describe how this classroom assessment data is used to report student level data for educational planning and region-wide accountability; 3) discuss how and to what extent this school-based assessment supports learning in the classroom and contributes to teacher professional development.

8. *"Assessment models for monitoring learning: New Zealand"*

Hattie, John (The University of Auckland, New Zealand)*

New Zealand has a recent history of self-managed schools with many freedoms to make decisions about teaching and assessment. There are many options for them to choose. The session outlines the options available in an on-line assessment package (asTTle) which includes Teacher customised, comprehensive, computer adaptive, interview, and attitude assessment. Feedback is immediate to teachers and students in the form of visual reports, and while they can be used for many purposes the major use is to monitor teaching and learning.

9. *"Comparing and contrasting models for monitoring learning in three countries"*

Ercikan, Kadriye (University of British Columbia, Canada)*

This presentation will review, compare and discuss models for monitoring teaching and learning in three countries that will be presented in the first part of the symposium: Denmark, Hong Kong and New Zealand. The three systems will be reviewed and discussed in terms of their influences on learning and teacher autonomy, the stakes associated with assessments, the types of assessments used, the levels of aggregation of data from these assessments, and how data are used.

Advanced issues in computerized adaptive testing and computerized classification testing

Chair

Wang, Wen Chung (Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong SAR, China)

Symposium Abstract

Computerized adaptive testing (CAT) has been widely implemented in recent years, mainly because of the significant progress in computer technology and item response theory. A major advantage of CAT is that it yields person estimates more efficiently than paper-and-pencil tests. In some cases where a classification of test-takers into a few categories is sufficient, CAT can be adapted to fulfill this goal. The adapted procedure is called computerized classification testing (CCT). Nowadays, many conventional paper-and-pencil tests or non-adaptive tests have been gradually replaced by CAT or CCT. Computerized adaptive testing (CAT) has been widely implemented in recent years, mainly because of the significant progress in computer technology and item response theory. A major advantage of CAT is that it yields person estimates more efficiently than paper-and-pencil tests. In some cases where a classification of testtakers into a few categories is sufficient, CAT can be adapted to fulfill this goal. The adapted procedure is called computerized classification testing (CCT). Nowadays, many conventional paper-and-pencil tests or non-adaptive tests have been gradually replaced by CAT or CCT. The symposium contains four papers which address some advanced issues in CAT and CTT. In the first paper, CCT was implemented under the generalized graded unfolding model. In the second paper, CCT was implemented under the higher-order item response model. In the third paper, CAT was implemented under the two-parameter testlet model with ability-based guessing. In the fourth paper, a new method was developed to increase efficiency in the expected a posteriori ability estimation under multidimensional item response models. All the four papers adopted simulations to evaluate the performances of the new algorithms under a variety of conditions.

10 *"Implementation of computerized classification testing under the generalized graded unfolding model"*

Liu, Chen-Wei (Department of Psychology, National Chung Cheng University, Taiwan)
Wang, Wen-Chung (Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong SAR, China)*

The generalized graded unfolding model (GGUM) has been recently developed to describe item responses to Likert items (agree-disagree) in attitude measurement. In this study, we developed two item selection methods in computerized classification testing under the GGUM, the current-estimate/ability-confidence-interval method and the cut-score/sequential-probability-ratio-test method and evaluated their accuracy and efficiency in classification through simulations. The results indicated that both methods were very accurate and efficient. The more point each item had, the fewer the classification categories were, the more accurate and efficient the classification would be. However, the latter method may yield a very low accuracy in dichotomous items with a short maximum test length. Thus, if it is to be used to classify examines with dichotomous items, the maximum text length should be increased.

11 *"Computerized classification testing under the higher-order IRT model"*

Lee, Kung-Hsien (Department of Psychology, National Chung Cheng University, Taiwan)*
Wang, Wen-Chung (Assessment Research Centre,The Hong Kong Institute of Education, Hong Kong SAR, China)

Many CCT algorithms have been developed under unidimensional item response theory (IRT) models. In practice, a test battery may contain multiple tests, and a test may contain multiple subtests. For example, an English test usually includes subtests of listening, speaking, reading, and writing, and each subtest consists of several items. Besides, it is theoretically justifiable that there is an upper-layer "overall" English proficiency that governs the four "domain" abilities of listening, speaking, reading, and writing. Under such a case, the higher-order IRT model is needed, because it accommodates such a hierarchical structure in latent traits. The major advantage of the higher-order IRT model is that both the overall and domain abilities can be estimated simultaneously, which is not applicable for standard unidimensional or multidimensional IRT models. Until now, CCT has not yet been developed under the higher-order IRT model. In this study, we developed such algorithms and evaluated their performances under a variety of situations through simulations. The results showed that the developed CCT algorithms were efficient and accurate in the categorization of test-takers by taking into account both the overall and domain abilities, especially when the loadings of the overall ability on the domain abilities were high.

12 *"Computerized adaptive testing under the two-parameter testlet model with ability-based guessing"*

Huang, Sheng-Yun (Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong SAR, China)*
Wang, Wen-Chung (Assessment Research Centre, The Hong Kong Institute of Education, Hong Kong SAR, China)

Testlet response models were developed to fit item responses to testlet based items. As the guessing in multiple-choice often involves ability, a new IRT model with ability-based guessing was proposed. To analyze multiple-choice items with testlet design, in this study we incorporated the modeling of ability-based guessing into the two-parameters testlet response model, implemented CAT algorithms, and compared the performances of three item exposure control methods through a series of simulation. In the first simulation study, three independent variables were manipulated: (a) testlet effect (small and large), (b) . size (small to large proportion of ability on guessing), and (c) testlet length. In the second simulation study, we compared the performances of three item selection procedures. The results indicated that the CAT algorithms and the three item exposure control methods for the new IRT model were successfully developed and implemented. The smaller the testlet effect and the longer the testlets were, the smaller the root mean square error would be; the Sympson and Hetter online method and the Sympson and Hetter online with progression method could maintain a well-controlled item exposure rate as their pre-specified rate without substantial loss in measurement accuracy. Although the progression method could maintain control of item exposure rate, it had a higher bank usage and higher measurement accuracy.

13 *"Improving the expected a posteriori (EAP) ability estimation in multidimensional computerized*

adaptive testing"

Chen, Po-Hsi (Department of Educational Psychology and Counseling, National Taiwan Normal University, Taiwan)*

In multidimensional computerized adaptive testing (MCAT), the computer time for traditional expected a posteriori (EAP) ability estimation methods increases exponentially as the number of dimensions increases linearly. For example, in four dimensional MCAT, it took about 15 seconds to yield ability estimates on the four dimensions and to select the next item using traditional EAP estimation method when there were 30 quadrature points in each dimension (Chen, 2006). The computer time should be largely reduced to make MCAT feasible. In this study, I proposed a new EAP estimation method and evaluated its performance through simulations. Six sets of simulated data, constructed by three different numbers of dimensions and two different types of correlations between dimensions, were used to compare the performances of the new and traditional EAP estimation methods. The dependent variables were conditional bias and root mean square of error (RMSE) after administering 5 and 10 items in each dimension, and the mean computer time for ability estimation and item selection. The results indicated that the new method needed much less computer time than the standard method, especially when there were as many as 6 dimensions. In addition, the new and standard methods yielded a similar degree of conditional bias and RMSE. In conclusion, the new method can improve the proficiency of the EAP ability estimation to some degree, and can be easily implemented in MCAT.

14*"Considerations in developing vertical scales using IRT: Link methods, link items and sample size issues"*

Mok, Magdalena Mo Ching (Assessment Research Centre, & Psychological Studies Department, HKIEd, Hong Kong SAR, China)*

Yan, Zi (The Hong Kong Institute of Education, Hong Kong SAR, China)

Lau, Doris Ching Heung (Assessment Research Centre, HKIE, Hong Kong SAR, China)

The purpose of this paper is to contribute to recommendations for the development of vertical scales using IRT methods. The recommendations will be based on three sources, namely, test development literature, simulated data, and empirical data. Considerations in vertical scale development include in this study include method (concurrent, stepwisechain) of linking, properties of link items (goodness of fit, adherence to curriculum), number and proportion of link items (15%, 30%), and model of analysis (Rasch, 2-parameter model). A number of data sets will be simulated to test effect on scale properties of different methods in the construction of vertical scale. Merits of these methods are compared by comparing the match between parameters of the estimated and the original (generated) samples. The real data set comprise a sample of 5,755 primary students between primary 2 and primary 6 from 24 schools, and 3,621 secondary students between secondary 1 and secondary 3 from 11 schools in Hong Kong. The mathematics competencies of participants were assessed using booklets with linked items between adjacent year levels. Different methods of constructing the vertical were compared in terms of estimated population mean (grade-to-grade growth), estimated population standard deviation (grade-to-grade variability), separation of grade distributions by effect size, and differential item

functioning. The paper will combine findings from the literature review, analysis on simulated and real data to derive a set of recommendations regarding the development of vertical scales for charting growth across year levels.

15 *Testing the invariance of latent traits in multiple group analysis*

Zhang, Zhiyong (Department of Psychology, University of Notre Dame, USA)*

Abstract

Evaluating measurement invariance (or differential item functioning, DIF) is critical for international testing. Traditional methods require the same set of test items to be used in all populations. After obtaining invariance, the latent traits can then be compared. This study proposes to test the invariance at the latent trait level so that the test items can be tailored according to each population. For example, authoritarian behavior might be the target latent trait in a study, but in one population authoritarian behavior might be indexed by measures of physical coercion while in another population it might be indexed only via verbal behavior measures. Thus, the latent traits are the same although the test items are different. We propose to use marker variables and the likelihood ratio statistic to test the invariance of latent traits across different populations. Through well-designed simulation experiments, we demonstrate the feasibility of testing latent trait invariance when tailored items are used. The study has important implications in international testing. For example, the design of studies can be innovative in building measurement batteries that are better matched to one's experimental aims. For example, in studies involving subgroup comparisons based on age, ethnicity, gender, etc., test batteries can be tailored to the subgroup while still measuring the samelatent traits.

16 *Structural modeling of TIMSS 2007 mathematics test data*

Choi, Youn-jeng (University of Georgia, USA)*

Cohen, Allan S. (University of Georgia, USA)

Bandalos, Deborah L. (University of Georgia, USA)

Abstract

The TIMSS (Trends in International Mathematics and Science Study) testing program is designed to provide comparative information about student achievement in mathematics and science among participating countries. In this study, we focus on a structural analysis of Grade 4 mathematics with an eye to determining what student and country characteristics might be related to differences in test performance. This grade is important as it is typically when most countries begins formal education of rational numbers, particularly proportional reasoning. We focus on several sets of characteristics, including student, teacher, educational policy, and educational curriculum characteristics to try to determine which sets appear to have the most impact on student test performance. TIMSS data from the 39 countries participating in TIMSS 2007 Grade 4 mathematics will be used in this study. TIMSS 2007 used 14 test booklets administered randomly to students so we will use one test booklet among the 14 test booklets. Preliminary exploratory factor analyses from one randomly selected booklet (N = 13,446) indicates a single factor. Linear regression indicated several factors related to total math score over the 26 items in the booklet: public expenditure on education, pupil-teacher ratio, emphasis on a national

curriculum, and provision of remedial instruction. Multilevel structural equation modeling will be used to further analyze these data for the conference paper.

17 *The impacts of participant personalities on the judgments in modified angoff standard settings*

Zhang, Sheng (Faculty of Education, Beijing Normal University, China)*
Zhang, Danhui (Faculty of Education,Beijing Normal University, China)
Tang, Keran (National Center of Educational Assessment, China)
Chi, Zhaoyan (National Center of Educational Assessment, China)

Abstract

The procedure of standard setting is to collect the judgments of qualified participants with various experience and expertise about the level of knowledge and skills required for a person to be considered as above the cut-score. Many studies have investigated objectivity of judgment and the extent to which one's final judgment is influenced by the background of participants. This study investigated the impact of participant personalities on judgments. A modified Angoff method with four iterative group discussions was employed in setting performance standards for the Chinese National Assessment of Education Quality in 2008. Influential data were also provided during the group discussion to modify participants' judgment. The personalities of each participant were also measured with the Big Five Personality Scale for the purpose of measuring participants' characteristics on agreeableness, conscientiousness, openness, neuroticism, and extraversion. It was found that people with different personalities showed distinct behavioral tendencies in this setting. Those with neuroticism tended to adjust their judgments based on the provided information, but they were not easily influenced by the other participants; those higher on agreeableness were more likely to agree with others, as well as reach consensus; those higher on openness were generally active in the group discussion. Despite the discrepancies of person behaviors during the procedure, the findings suggested that the participant personality was not a significant factor in influencing the overall final standard settingjudgment.

18*CTT, IRT, and G-DINA analysis of TIMSS 2007*

Park, Yoon Soo (Teachers College, Columbia University, USA)*
Lee, Young-Sun (Teachers College, Columbia University, USA)

Abstract

Mathematics educators and psychometricians often wonder how best to estimate the mastery of grade-appropriate curricular skills and are also interested in examining regional differences in attribute mastery within the same country. In the 2007 administration of the 8th grade mathematics Trend in International Mathematics and Science Study (TIMSS), two U.S. states—Massachusetts and Minnesota—and three Canadian provinces— Québec, Ontario, and British Columbia—were included as benchmarking participants, which are regional entities that follow the same assessment procedures as the countries. They ranked 6, 7, 8, 9, and 13, respectively, among 56 total participants, which were all above the U.S. national sample that ranked 14th. This framework and design of the TIMSS assessment provides an ideal structure to conduct an empirical analysis of attribute mastery within the U.S. and Canada. Although various attempts have been made to make inferences on attribute mastery

based on information from student performance based on broad content domains, traditional methods of analyzing large-scale assessments such as classical test theory, item response theory, or generalizability theory have shown to be ineffective to provide attribute-level information. In contrast, employing a cognitive diagnostic modeling approach, attribute mastery as well as cognitive diagnostic information can be attained. This study examines fine-grained attribute mastery using the deterministic, inputs, "and" gate (DINA; Junker & Sijtsma, 2001) model for each regional entity that can be used as diagnostic information for educational instructors to improve student performance and learning in mathematics education.

19. *Applying mixture IRT models to TIMSS data*

Alexeev, Natalia (University of Georgia, USA)*
Cohen, Allan (University of Georgia, USA)
Templin, Jonathan (University of Georgia, USA)

Abstract

The Trends in International Mathematics and Science Study (TIMSS) is an international assessment to measure trends in mathematics and science learning. The aim of TIMSS is to improve the teaching and learning of mathematics and science by providing data about students' achievement relative to different curricula and instructional practices. A recent development in TIMSS research is application of mixture IRT models, particularly mixture Rasch models (MRM), to investigate qualitative and quantitative differences among latent groups in the examinee population. The use of the MRM is prevalent in such studies in part due to its simplicity as well as to the availability of software packages for estimating model parameters. Some previous research suggests, however, that using reduced IRT models, such as the Rasch model, may lead to over-extraction of latent classes and, therefore, to misinterpretation of results. The purpose of this study is to investigate how mis-specified mixture IRT models can affect extraction of latent classes and thus interpretation of the data. The problem and solutions will be illustrated using TIMSS 2007 Mathematics
Test for Grade 8.

20. *Combining Bayesian networks with two-tier items to modeling students' learning bugs and sub-skills*

Shih, Shu-Chuan (Department of Mathematics Education, National Taichung University, Taiwan)*
Kuo, Bor-Chen (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)
Yang, Chih-Wei (Graduate Institute of Educational Measurement and Statistics, National Taichung University, Taiwan)

Abstract

The main purposes of this study are to develop the two-tier mathematics diagnostic tests based on Bayesian networks and explore the efficiency of combining Bayesian networks with two-tier items for modeling students' learning bugs and sub-skills in time calculation after students have learned the related contents. Six steps are involved in this study: developing the student model based on Bayesian

networks that can describe the relations between bugs and sub-skills; constructing two-tier items that students can be provided an opportunity to reveal their bugs and sub-skills in time calculation contents; using two-tier items for evidence model creation and completing the cognitive diagnostic model that combining Bayesian networks with two-tier items; administering test items for sixth graders in elementary school; estimating the network parameters using the training sample and applying the generated networks to bugs and sub-skills diagnosis using the testing sample; and assessing the effectiveness of the combined models work in predicting the existence of bugs and sub-skills. The tests are administered to 300 sixth grade students. The responses of 240 samples are used as a training data set for building Bayesian networks and others are treated as a testing data set for evaluating the holdout classification accuracies of the combined model. The results show that using combined model to diagnose the existence of bugs and sub-skills in individual students can get good performance.

21. *Value madded models and accountability*

Swaminathan, Hariharan (University of Connecticut, USA)*
Rogers, H. Jane (University of Connecticut, USA)

Abstract

With the advent of the No Child Left Behind (NCLB) legislation, the issue of accountability has come to occupy center stage in public education. The policy governing Adequate Yearly Progress (AYP), designed to assess the effectiveness public education, requires the comparison of students' performance at a single point in time to pre-set standards. Recent research has documented the problems associated with and the inappropriateness of using the measure of academic growth as dictated by NCLB. On the other hand, the value added assessment procedures based on academic growth of students have been heralded as the tool for educational accountability and have been implemented in several states, notably Tennessee for documenting teacher effectiveness. While considerable research has been conducted on the usefulness of the AYP measure for accountability, the psychometric and statistical issues surrounding Value Added Models have been less well understood. The purpose of this paper is to examine the psychometric and statistical issues surrounding Value Added Assessment and Value Added Models. In particular, the growth model and projection model procedures developed by the authors for conducting value added assessment in a US state will be described and compared. In addition, the use of such models for identifying and providing resources for children and schools/districts at risk of failing AYP is described. Data from the state assessment will be used to illustrate the methods described in the paper.

22. *Investigating item drift in TIMSS via Cognitive Diagnostic Modeling*

Park, Yoon Soo (Teachers College, Columbia University, USA)*
Lee, Young-Sun (Teachers College, Columbia University, USA)

Abstract

Longitudinal studies of trended international mathematics achievement have focused on the overall performance of students using trended assessments such as the Trends in Mathematics and Science

Study (TIMSS) and the OECD Program for International Student Assessment (PISA). They have employed methods such as Classical Test Theory (CTT) and Item Response Theory (IRT) to rank individuals within a latent ability continuum. Although inferences generated from these approaches have provided insights into the relative standing of students and their mathematics ability in comparison to other countries, they have yet to examine how specific attribute mastery change over time—whether unique skills required to solve a mathematics problem have grown or remained constant with time. This view is different from examining student performance in broad domains such as algebra and geometry, because investigating fine-grained attributes form the basis to students' understanding of the material as well as providing direct information to educational researchers and instructors on areas that students need improvement. Cognitive Diagnostic Models (CDM) were developed for this specific purpose—to examine whether a specific cohort has mastered skills that are required to correctly answer a problem. Using three waves of the TIMSS—1999, 2003, and 2007—this study examined attribute mastery from a CDM framework while implementing longitudinal analysis methods. This study shows that by examining item drift—whether CDM parameter estimates change over time—we can signal instructors and mathematics researchers on content areas that have deviated and detect trends in attribute mastery that cannot be estimated using traditional methods.

參、心得

1.研討會內容很多研究與想法值得學習，ITC已舉辦七屆，此次是第一次在非西方國家舉辦，研究主題多樣化，尤其很多關於測驗的新興議題與未來發展，如認知診斷、DIF新的研究趨勢，都在此次研討會會中呈現，對於個人研究方向或實務上(如TASA與教檢測驗技術的改善)助益良多。

2.此次有紐西蘭、丹麥、加拿大與香港所發表的監控學習評估模式，對於教育部、地方教育局、學校、教師、學生與家長提供有效的界面與資料庫管理，這些國家投入大量資源建立監控學習的評量模式(如紐西蘭 http://e-asttle.tki.org.nz/) 頗值得我國教育當局或各縣市學習。

3.規模盛大，但發表安排不夠周延，造成想聽的場次相衝，因篇數很多又只有三天，造成同一時段有6個session同時發表，無法兼顧而有遺珠之憾。

4.香港大學國際化值得學習，與會的所有文章包括口頭與海報皆以英文發表，雖部份局限了中國人發表的空間，但相對的也增加了國際學者的參與。

5.便利的港鐵與捷運站周遭的住家、商業規劃及香港文化水準(服務、禮貌、守法)造就了繁華、擁擠而不亂的東方明珠。香港城市的整體規畫與建設亦值臺灣當局得學習。

肆、建議事項

1.鼓勵投稿與經費補助，讓本處學術與研究成果得以在國際發聲，並獲取新知。

2.加強英語能力，將研究成果展現。

3.有機會能增取主辦類似國際研討會，以打開臺灣及本處之知名度。

附件

A STUDY ON DIFFERENTIAL ITEM FUNCTIONING (DIF) OF THE BASIC MATHEMATICAL COMPETENCE TEST FOR JUNIOR HIGH SCHOOLS IN TAIWAN

By

Chien-Ming Cheng

National Academic for Educational Research Preparatory Office

cheng@naer.edu.tw

ABSTRACT

This study investigates the relationship between a gender's group membership and performance on test items using four differential item functioning procedures - Area Measure, Likelihood ratio test, Mantel-Haenszel, and SIBTEST methods.

The basic competence test for junior high schools is a very important breakthrough in education in Taiwan because it adopts the item response theory (IRT). The DIF topic in IRT is important because of concern that the basic competence test for junior high schools be fair and impartial for every student. In the study the presence of DIF for gender groups is investigated for this new system of testing. The results of this study are the identification of items that show evidence of DIF and that are judged to be due to bias, a determination of which methods are the most accurate for detecting DIF and an investigation of the possible reasons for causes of DIF and bias.

Both real and simulation data are analyzed to compare the four detecting DIF methods. From the results, synthesis and discussion of effect size, frequency, consistency, and Type I error rate, of the four methods, SIBTEST was deemed the most appropriate to detect DIF items for the basic mathematical competence test for junior high schools in Taiwan.

Keywords: differential item functioning, Area Measure, Likelihood ratio test, Mantel-Haenszel, and SIBTEST