Appendix
The Contribution of Lower-Level Processing to FL Reading Comprehension with Chinese EFL Learners

Presenter
Ms Feifei Han
A PhD candidate at Faculty of Education and Social Work
The University of Sydney
e-mail: feifei.han@sydney.edu.au

Aims and Keywords
This presentation reports an ongoing research project, which investigates the two competing hypotheses: whether inefficient lexical access (LA) and small working memory (WM) inhibit text comprehension in FL reading (inhibition hypothesis) or whether readers could use strategies to compensate for processing and language problems so that text comprehension are not influenced much (compensation hypothesis).

Sample
Four hundred and two Chinese university students in their second year participated in the study.

Methods
The larger project adopted a mix-method design collecting both quantitative and qualitative data, but the focus of the presentation is only on the quantitative data. The reading text were analysed using RUMM2030 for both checking both misfitting items and persons. The data were then analyzed by doing correlation analyses between LA, WM and reading comprehension in two reading conditions: untimed reading and timed reading.

Results
Firstly, there was no significant correlation between students’ LA and reading comprehension in untimed reading condition ($r = -.09, p = .07$), whereas a small and negative relationship was found between LA and reading comprehension in timed reading ($r = -.22, p < .01$). Secondly, WM showed to be correlated positively with students’ reading comprehension in both untimed ($r = .11, p < .05$) and timed reading conditions ($r = .20, p < .01$), both the values of correlation were small. The magnitude of correlation between WM and comprehension in untimed reading was smaller than that between WM and comprehension in timed reading.

Conclusions
The preliminary results from the quantitative data seem to support the compensation hypothesis that when readers are allowed sufficient time, inefficient word processing and small working memory do not inhibit text comprehension in FL reading.
Generating a Learning Scale for Evidence Based Health Care: What Rasch Measurement Says Is Evidence?

**Presenter**
Dr Ian Blackman & Ms Tracey Wachtel
School of Nursing & Midwifery
Faculty of Health Sciences
Flinders University
Adelaide, South Australia

**Background**
Australian health care practitioners continue to be challenged by the diverse health needs of consumers. In its mandate to prepare competent Registered nurses for the Australian nursing workforce, undergraduate nursing programs continue to articulate nursing skills development based on the concepts of evidence based practice, which in turn impacts on clinical practices.

**Aims**
The study seeks to identify if a scale of learning about evidence based health care practices can be generated, based conjointly on the self rated abilities of completing undergraduate nurses, to the different complexities of evidenced health care practices that are expected of contemporary nurses.

**Sample**
A group of Australian-based undergraduate nurses (n = 275) completing a 3 year undergraduate nursing program has been selected for the study.

**Methods**
A self-rated survey (using a four point Likert scale) is used by the completing undergraduate nurses to rate their self-efficacy in understanding and applying evidence based health care practices. To demonstrate this, Rasch scaling and in particular, the partial credit model has been employed. Differentiated item functioning will also be used to detect if nursing students from different subgroups score differently from each other.

**Results**
Outcomes indicate that indeed a robust scale of learning about evidence based health care practice can be established based on undergraduate nurses’ self-rated estimates and the differing levels of complexity required for Australian evidence based health care practices.

**Conclusions**
The study suggests student learning about evidence based health care practices can be robustly measured using Rasch analysis. In its bid to account for as much invariance as possible, the Partial Credit Model together with Differentiated Item Functioning can give rise to reliable self- efficacy estimates for health care practitioners.
Further Implementation of ACER ConQuest User Defined Fit Statistics

Presenter
Daniel Urbach
Research Fellow
Psychometrics and Methodology
Australian Council for Educational Research
19 Prospect Hill Road, Camberwell VIC 3124
t: +61 3 9277 5683
e-mail: urbach@acer.edu.au

August 8th (11:00–11:25)

Background
This presentation further investigates results from Adams and Wu (2004) who developed a User Defined Fit Statistic to test the fit of linear combinations of items.

Aims and Keywords
The aim of the presentation will be to showcase the advantages and uses of ACER ConQuest User defined Fit Statistics (UDFS). The advantage of using UDFS’s is the ability to identify violations of local item independence and/or violations of uni-dimensionality using a-priori knowledge of item response data. In a lot of cases, individual item based fit statistics are unable to identify such violations.

Sample
Two simulated data sets both containing 1,000 cases as well as 10 and 20 items are investigated as well as two real data sets. The first comes from the Australian Graduate Medical School Admissions Test (GAMSAT) and the second, comes from a survey of Student Attitudes for a Victorian Learning Difficulties Project.

Methods
The simulated data was generated with model violations (namely local dependence and multi-dimensionality). Two real data sets were also used, one with item bundles (which was tested for item dependence violations), and one with known sub scales (which was tested for dimensionality violations). For each data set, individual item fit statistics were also compared to the UDFS’s. This was done in ACER ConQuest.

Results
The simulations showed the UDFS’s ability to identify violations of item response model assumptions, when local item dependence and multi-dimensionality were introduced in the generated simulation data. Two real datasets were also analysed, and such violations were again identified.
Conclusions
When a priori knowledge of item combinations of instruments are known, which is often the case in commercial assessments which contain item bundles or sub-scales, UDFS’s may be extremely useful in identifying violations of local item independence and uni-dimensionality in practice.
Participation in international assessment programs that are designed to monitor student progress has increased in recent years. Such assessment programs provide important information on the achievement of students in a specific country relative to students in other countries, and identify factors that have contributed to this achievement. Since curriculum and language of instruction varies across different participating countries, it is often essential for the original tests to be adjusted for curriculum differences and be translated into different languages. However, it is widely recognised that curriculum differences and differences between source and target languages may have a significant impact on the equivalence of test items. As a result, determining the cross-lingual and cross-country equivalence of the international test items is fundamental. This study is designed to investigate the impact of language and curriculum Differential Item Functioning (DIF) on the construction of an International Science Scale (ISS). A total of 12,762 students (Grade 4–Grade 10) took a Science Test in 2011. Subjects were classified into three groups: (1) English native speakers who took the English version, (2) Chinese native speakers who took the Chinese version and (3) Chinese native speakers who took the English Version. The Rasch Simple logistic model was applied to analyse the data using RUMM 2030 software. RUMM 2030 uses the analysis of variance of the residuals to identify those items with language DIF. Once DIF was detected, the ‘top-down purification’ approach (Tennant and Pallant 2007) was adopted to deal with items displaying DIF. The results of this study show that DIF due to curriculum and language can change the relative position of examinees on the latent variable. This study has important implications for international assessment programs, particularly when data obtained from different countries are to be pooled to construct a common scale.
Extended Mantel-Haenszel Procedure in DIF Detection

Presenter August 8th (9:50–10:15)
Dr. Xiaoxun Sun
Psychometrics and Methodology Research Program
Australian Council for Educational Research
19 Prospect Hill Road, Camberwell, VIC, 3124
t: +03 9277 5419
e-mail: xiaoxun.sun@acer.edu.au

Background
Many methods have been developed for studying Differential Item Functioning (DIF). An approach developed by Holland and Thayer (1988) adapted the Mantel-Haenszel (MH) statistic (Mantel and Haenszel 1959). Xiaoxun Sun has implemented MH procedure and its extension in an experimental version of ACER ConQuest which can handle DIF analysis for multiple focal groups and polytomous items.

Aims
This presentation compares the extended MH procedure in DIF detection to the standardized item difficulty difference method based on the Rasch estimates for both dichotomous and polytomous items. The use of MH method in DIF analysis with multiple focal groups will be illustrated.

Keywords DIF • MH statistics • ACER ConQuest

Sample
Grades 3–10 data from International School Assessment (ISA) are used in the study. The assessment data were from 25,000 students.

Methods
The DIF analysis results based on the standardised difference method has been compared with results using extended MH procedure. In using the MH procedure, stratification based on ability is studied in terms of appropriate number of strata. ACER ConQuest has been used for the analyses.

Results
The Rasch DIF method has a higher DIF detection rate than the MH procedure. The appropriate number of strata used in MH method is found to depend on the test length and ability distribution.

Conclusions
The MH results showed a more conservative DIF detection rates compared with the standardized difficulty difference method except for very easy or very difficult items. The MH method can be used to detect non-uniform DIF. The extended MH procedure can better handle multiple focal groups and polytomous items.
The Attitudes of School Leaders to the Relationship Between School Registration and School Improvement

In 2004, the Government of Western Australia introduced a mandatory inspection-type registration process for all Non-Government Schools. Part of the aim of this registration process was to help schools improve twelve educational and administrative aspects. These were: (1) School Governance, (2) School Financial Viability, (3) Enrolments & Attendance, (4) Number of Students, (5) Instructional Time, (6) School Staff, (7) School Infrastructure, (8) School Curriculum, (9) Student Learning Outcomes, (10) Care for Students, (11) Disputes and Complaints, (12) Legal Compliance. A questionnaire based on these twelve aspects was designed with five items per aspect (60 items total), conceptually ordered from easy to hard, and given to 110 administrators. It was completed by 65 administrators for a usable response rate of 59%. The data were analysed with the Rasch model computer program RUMM2030 which accommodated the small numbers by estimating parameters even when some response cell frequencies are zero or low. It does this by re-parameterising the thresholds into principal components (not the factor analysis kind), but components that make up the structure of the threshold parameters where there are data. The frequencies are not used directly, but rather functions of the frequencies are used as the sufficient statistics for these parameters and the thresholds are recovered from these. A unidimensional, linear scale, School Administrators’ Beliefs That Actual School Improvements Were Due to Formal School Registration, was created with 48 items. The Person Separation Index of 0.86 was highly satisfactory. The item-trait interaction was 83.76, df = 96 with p = 0.81 supporting the creation of a unidimensional scale. The results showed that there was a group of items that administrators said were relatively easy to say that actual school improvements were due to formal registration and another group that administrators said were very hard to say that actual school improvements were due to formal registration. This study produced a new Rasch measurement for a key aspect of school improvement. It provides new insight into the policy and practice of school registration.
The Impact of Unobserved Extreme Categories on Item and Person Estimates: A Simulation Study

Presenter
Dr. Edward Feng Li
Educational Assessment Australia
UNSW Global
e-mail: edwardfli@hotmail.com.

Background
For any polytomous items, it sometimes occurs that an extreme category, which is logically possible, is not observed in a particular sample. For example, in education when the performance tasks by the students from different year levels are judged by the same set of criteria, it is likely that none of the lower year level students would achieve the highest marks in some criteria. In health, it may happen when a group of generally healthy participants are measured by an instrument designed to detect some particular symptoms.

Aims and Keywords
This paper uses a simulation study to investigate the impact of unobserved extreme categories on item and person estimates.

Keywords Polytomous items • Unobserved extreme categories • Item estimates • Person estimates

Sample
The sample size for this simulation study is 1,000, with a mean of 0 and a standard deviation of 1, N(0,1).

Methods
Based on the polytomous Rasch model, the Partial Credit Model (Master 1982), data were simulated for 1,000 persons and 10 polytomous items with five categories under two scenarios, one with unobserved extreme high categories and the other with unobserved extreme low categories. The generated data sets were analysed with the RUMM2030 software (Andrich, Lyne, Sheridan and Luo 2009).

Results
The results show that unobserved extreme high categories in the data tend to lead to underestimated person measures and unobserved extreme low categories in the data tend to lead to overestimated person measures. The results suggest that collapsing unobserved extreme high categories, even when no disordered thresholds are present, improves person and item estimates, compared with leaving them
unchanged. However, collapsing unobserved extreme low categories only improves item and person measures when a large proportion of items have unobserved extreme low categories.

**Implication**
These results have implications for designing and measuring performance tasks that need to be carried out across a wide spectrum of ability groups.
A Comparison of Ratings Between American and Australian Nurses Using Differential Item Functioning (DIF)

Presenter
Dr Patricia Nicholson
Melbourne School of Health Sciences, Faculty of Medicine
Dentistry & Health Sciences
The University of Melbourne
e-mail: pfnich@unimelb.edu.au

Background
Following validation of a Performance Based Scoring Rubric using Item Response Modeling the analysis showed that the Analytical Observation Form produced acceptable reliability estimates (0.94) and the fit statistics for most of the items were acceptable. Using a video-clip as a fixed prompt, the factors likely to influence the accuracy of the performance ratings of nurse educators were explored using the validated instrument.

Aims and Keywords
The aim of the study was to determine to what extent the ratings of nurse educators were influenced when using an Analytical and Holistic rubrics to judge the performance of an instrument nurse observed in a video-clip.

Keywords Differential item functioning • Rasch Model • Competency-based performance

Sample
The sample included nurse educators who were involved in providing clinical support to nurses working in the operating theatre (Australia = 186; America = 127).

Methods
A group of nurse educators were required to observe the performance of an instrument nurse captured in a video-clip using the rubric which was developed using the Australian College of Operating Room Nurse and Association of peri-Operative Registered Nurses unit of competency and standards of practice. An independent-samples t-test was conducted to compare the mean scores of American and Australian nurse educators with a significant difference (p < 0.001)in the mean scores found (26.46 and 22.74 respectively). In order to explore whether errors or bias were impacting on the quality of the data, the t-statistic of the paired item difficulty estimates (i.e. American and Australian item difficulty estimates) for each item on the Analytical Observation Form was explored using ConQuest (Wu et al. 2007). A direct comparison was then made to determine the differences in each subgroup.
**Results**
Although there were eight items with $t$-statistics smaller than $-2$ and larger than 2, only two items exhibited significant differences with $t$-values of $-3.96$ and $-7.238$ respectively.

**Conclusion**
The study had several limitations which included the use of a video-recording of practice. Although DIF was detected between the Australian and American nurses, further research is required to confirm the findings from this study and allow for the generalization of the findings in clinical practice. The use of international competency standards also have wide reaching implications for further research in examining nursing practice difference in various countries.
A Research on the Effectiveness of DynEd Computer-Assisted English Language Learning: Taking Ningbo Polytechnic as an Example

Presenter August 8th (11:25–11:50)
Jingru Huang; Baixiang Wu
Ningbo Polytechnic
Ningbo, Zhejiang, China

The Integration of Multimedia-aided English language teaching and learning is one of the trends of abroad and domestic research, and DynEd CALL (computer-assisted English Language learning) within the framework of integration is a new challenge. This paper explores the DynEd CALL model based on public English teaching reform of Ningbo Polytechnic that has been done for 3 years, which aims to find the effective model of language learning and teaching. The purpose of this study is to investigate the effectiveness of DynEd CALL model. The data were collected through DynEd Records Manager, on over 2000 non-English majors from 7 institutes of Ningbo Polytechnic, implemented for 1 year. The data includes each student’s studying time, days studied, two placement tests, two speaking tests (both were done at the beginning of learning DynEd and at the ending of learning DynEd), study score, mastery tests score of each unit and detailed studying information about the courses of New Dynamic English and First English, such as functional buttons usage and speech recognition, etc. The data also contains a questionnaire to each student followed by 1 year’s DynEd study, 30 questions using Likert 5-point scale. Pre-and post-comparative analyses of two placements tests and speaking tests show that the differences are nearly 0.5 and 0.4 respectively, which means the award-winning, multimedia content keeps students on task and engaged. And the questionnaire indicates students have adopted the concept that language is a skill, not knowledge. DynEd CALL can significantly promote students autonomy and students’ listening and speaking skills, but deficiency on reading and writing skills, which are the perspectives we should enhance in the following teaching.

Keywords DynEd computer-assisted • Language learning • Effectiveness • Data base
Validation of a Large-Scale Reading-to-Write Task: Evidence from Multi-faceted Rasch Model Analysis

Presenter
Zhang Xinling, PhD
School of Foreign Languages
Shanghai University
Shanghai, China

The new decade witnesses the advent of integrative writing tasks in large-scale tests home and abroad, with accompanying validity evidence (Asencion 2004; Messer 1997; Watanabe 2001; Xu Hao and Gao Caifeng 2007; Zhang Xinling and Zeng Yongqiang 2009). However, research adopting Rasch model to detect test score variances are fairly inadequate. The present study employed MFRM to collect validity evidence for a large-scale reading-to-write test task. Analyses of 190 subjects’ writings with FACETS and the formula of \( \log\left(\frac{P_{njmk}}{P_{njmi}}(k - 1)\right) = B_n - C_j - E_m - D_i - F_{ik} \) yielded the following findings: (1) The task can distinguish among candidates of different reading-to-write abilities, and score variance is largely attributed to the construct; (2) The task is difficult for the targeted candidates; (3) bias analysis indicated that differentiating rating behavior of individual raters necessitates rating rubrics improvement and further rater training.

Keywords  Reading-to-write tasks • Multi-faceted Rasch model • Validation
An Analysis of Item and Person Fit to Rasch Model

Presenter
Guoxiong He¹, Huifeng Mu²
Faculty of Foreign Studies, University of Jiaxing
Jiaxing, Zhejiang Province
P. R. China
e-mail: 1. Hgx505@mail.zjxu.edu.cn
2. mhf775@yahoo.com.cn

Aims and Keywords
The study is to analyze item and person fit to the Rasch model, aiming to improve test quality and facilitate item banking.

Keywords Rasch • Fit analysis • Item difficulty

Sample
Five hundred first-year non-English major students from Jiaxing University, 18 years of age and up, male and female, are randomly sampled in the study. The scores were obtained from a reading subtest, which consists of 30 multiple-choice items.

Methods
The scores from 500 candidates on 30 items were arranged in successive order of difficulty and ability. GITEM III + SYSTEM was used to calculate item difficulty and candidate ability. For each item and each test taker, Rasch model fit statistics were calculated, including Wright’s squared standardized residuals (W) and associated t values. In the meantime, we split up the candidates into six groups on the basis of average ability. Bock’s chi-square (BCHI) was used in order to check whether the observed data can fit the IRT one parameter. We also take the natural logarithm of the likelihood function at the max. As this is dependent on the level of theta, so we need it standardized, which is Drasgow’s standardized appropriateness index ($l_z$).

Results
The results given by GITEM III + SYSTEM show that 6 items and 40 candidates were identified as misfitting by the conventional criterion of fit surpassing 2.00. But BCHI and $l_z$ produce the outcomes with different number of misfitting items and candidates, suggesting that several procedures are needed to check the degree of misfit.

Conclusion
No single fit statistic or statistical test can fully interpret item and person fit to the model because different methods produce different results. Some fit statistics are very sensitive to changes in sample size. We need several procedures for checking the degree of misfit, the unlikelihood of success or failure in responding to a test item.
The primary goal of objective language testing is to make an accurate and fair measurement of the language users. As a monitoring and evaluation mechanism, it maintains a positively interactive relationship with language teaching. In China, objective language testing has been widely used in some nation-wide tests, such as the Test for English Major (TEM-4), due to its quantitative design, convenient scoring and favoring fairness. Based on Alderson and Wall’s 15 washback hypotheses (1993) and empirical washback studies, the paper set out to investigate how effectively and efficiently TEM-4 can measure students’ competence in language learning and how it should be used appropriately in classroom teaching. Data were collected from over 525 teachers and students majoring in English of 10 sport universities and institutes in China by means of a questionnaire survey and in-depth interviews. Findings from this study indicated TEM-4 produced more positive washback effects than negative ones in that students improved their learning strategies and language proficiency; TEM-4 offered teachers feedback and helped them foster students’ comprehensive English abilities by using the test as a motivation tool. However, some discrepancies further supported that the washback effect was quite context-oriented and complicated. In sum, the paper achieved a breakthrough in carrying out a mixed (quantitative and qualitative) method from the perspective of teachers and students to investigate washback in the less explored area of Chinese sport universities and institutes, hoping the results could make some contribution to the improvement of English teaching in the sport universities and institutes in China.

Keywords Objective language testing • Washback effect • English teaching at sport universities and institutes • TEM-4
Educational assessment as well as studies on it has been playing a pivotal role in language teaching since the concept of “educational assessment” was put forward by R.W. Tyler in the 1930s. As more and more emphasis is put on the development of students’ ability and the process of teaching, formative assessment, which differs from the traditional result-oriented summative assessment, is being widely applied. Although the benefits brought by and importance of formative assessment have been confirmed, most of the researches are carried out only in college English classes, with little practice in English majors. This thesis intends to implement formative assessment in English majors’ translation course, aiming at finding out its effects on students’ translation ability as well as learning capabilities, and further providing suggestions on future studies in this area. Forty junior English majors from Beijing Sport University have been chosen for the research, during which self-assessment, peer-assessment and teacher assessment are used to implement formative assessment. Assessing tools including two achievement tests, journals, classroom observation, questionnaires and in-depth interviews have been used to collect both the quantitative and qualitative data. The research finds out that students’ role has changed from a listener to a participant, while the teacher from an authority to a guide. Students’ translation ability has been improved after the research and they have made progress in other learning capabilities like self-reflection, critical thinking and learning autonomy, which are crucial for future studies in other areas.

Keywords Formative assessment • Translation course • Translation ability • Learning capability
Measuring the Quality of Teacher-Child Interaction in China’s Preschools: A Rasch Measurement Approach

Presenters: August 6th (13:55–14:20)
Xiaoting Huang, Yingquan Song and Loyalka Prashant
China Institute for Educational Finance Research
Peking University
R404, College of Education
Peking University, Beijing, China
phone: 010–62758588
e-mail: xthuang@ciefr.pkue.du.cn

In recent years, Chinese government has drastically increased the national financial allocation on early childhood education (ECE), as a result of a growing consensus regarding the importance of ECE. Despite the rapid expansion, little attention has been brought on the quality of this sector. Previous research found that quality in ECE has many dimensions, including the “structural” and “process” aspects. Most of the studies in China focus on the structural aspect. Little has been said about the process quality. In this study, we constructed an instrument to measure teacher-child interaction-based process quality via Wilson’s four-building-blocks approach. Specifically, we designed an 11-item observation protocol. Trained raters observed each class for about 30 min before they scored the teachers. Eight hundred and eighteen teachers were observed in July 2011. The data were calibrated using Rasch-based partial credit model. The initial result showed that the overall reliability was only moderately high at 0.7. The scores were positively correlated with principal ratings, as well as teacher certification and professional training experience. The findings suggest that process quality can be scientifically measured using carefully designed instruments. The instrument needs to be improved in the future by adding more items and refining descriptors, etc. Finally, more elements of the process quality should be incorporated.
Multi-facet Rasch Model Applied to Investigating Rater’s Effects in PETS-SET

Presenter August 6th (14:20–14:45)
Laura Yang Hong*, Mingzhu Miao and Yan Zhao
Faculty of Foreign Studies, University of Jiaxing
56 Yuexiu South Rd, Jiaxing, Zhejiang Province, China
e-mail*: yanghong_158@yahoo.com.cn

Background
Public English Test System (PETS), one of the large-scale tests in China, is a proficiency test. The purpose of its Spoken English Test (SET) is to measure the ability of the test takers to communicate orally in English. It has a higher face validity. However, the raters’ subjective judgments and other problems of the raters’ instability and differences can affect the scoring, resulting in inconsistency of test scores, i.e. low reliability. Therefore, this study is to provide some implications for rater training so as to improve the rater reliability.

Aims and Keywords
Based on the modern Item Response Theory (IRT), the Multi-facet Rasch Model was used to investigate four types of rater effects: severity, rater instability, halo effects and extremism/central tendency to find out the rater reliability.

Keywords Rater effects • PETS-SET • Multi-Facet Rasch Model • Reliability

Sample
There were over 400 examination candidates took PETS-SET Band 2 in Jiaxing City in 2010. Eight pairs of raters and their ratings were chosen randomly, with each pair rating 30 candidates. The assessor adopted a 5-point analytic rating scale which consisted of three separate domains: Grammar and Vocabulary, Pronunciation, and interactive Communication. The interlocutor adopted a 5-point rating scale for Global Achievement.

Methods
The ratings were analyzed by using a Multi-facet Rasch Model (MFRM) conducted by the software package FACETS. The severity levels and the potential halo effects of the raters were investigated by using the primary analysis of MFRM and then the rater’s stability.
Results
Raters showed significant differences in their severity level; Raters behaved consistently in rating but some raters showed bias towards certain domains; and only 2 raters showed significant halo effects, i.e. they tended to give similar scores.

Conclusions
The results show statistically significant differences among all facets including severity, rater instability, halo effects and extremism/central tendency. MFRM is an effective means for measuring the rater effects.
Background
Foreign language aptitude is recognized as one of the most important individual variables in second language acquisition. Previous studies suggest that language aptitude is componential and different aptitude components may play different roles at different levels of second language proficiency. However, relatively little research to data has been conducted to test this hypothesis.

Aims
This study aims to explore the relationship between language aptitude components and different levels of English proficiency among Chinese English majors.

Sample
Sixty-two second-year English majors (9 males and 53 females, aged between 18 and 23) from a university in Beijing participated in the study.

Methods
An aptitude test composed of subtests of Pimsleur Language Aptitude Battery and subtests designed by the author was administered to the participants. The students’ TEM-4 and TEM-8 scores were used as measures of their English proficiency. Correlation analysis, multiple regression analysis and ANOVA were conducted.

Results
Results showed that different aptitude components had different relationship with the two language proficiency tests. Regression analysis showed that two aptitude components (Sound Discrimination and Memory for Text) were significant predictors of the students’ TEM-4 scores while three aptitude components (Language Analysis, Sound Discrimination and Memory for Text) were significant predictors of the students’ TEM-8 scores. Further analysis showed that students with higher and lower TEM-4 and TEM-8 scores also differed significantly in different language aptitude components.

Conclusion
The results lend support to the hypothesis that different language aptitude components may play different roles in second language acquisition when the learner is at different levels of proficiency, which implied that instructional methods could be tailored to suit learners’ language aptitude profiles according to their proficiency levels.
A Pilot Study Based on Rasch into the Appropriateness of the TOEIC Bridge Test for Chinese Students: Status Quo and Prospect

Presenter August 7th (16:00–16:25)
Quan Zhang
Member, Supervisory Committee of ELT in Vocational Higher Education, Ministry of Education, P.R. China and Faculty of Foreign Studies, University of Jiaxing, P.R. China
e-mail: qzhang141@yahoo.cn

Mingzhu Miao & Chunyan Zhu
Faculty of Foreign Studies, University of Jiaxing, P.R. China

Eng Han Tan
Head of ETS Beijing Rep Office
e-mail: etan@etsglobal.org

Background
Nowadays, the number of Chinese vocational students enrolled each year is increasing. Up to the present (2009), the total number nationwide amounts to almost 3.3 millions, of whom approximately 1.5 millions are English majors. However, over the past decades, there was no a well-accepted English language test for such a big number of students. TOEIC Bridge™ test (http://www.ets.org/toeicbridge) is being paid attention to by both Chinese government and university educators. It is the only English test officially introduced into China by Ministry of Labor and Social Security (MOLSS) under the Chinese government. To validate the test scores from the TOEIC Bridge as a measurement of English proficiency for and to eventually replace local tests of various kinds currently administered to Chinese vocational students annually of over 3.3 millions, the Supervisory Committee of ELT in Vocational Higher Education, Ministry of Education, P.R. China decided with much care to initialize the pilot study. The present study is part of this project.

Aims and Keywords
The purpose of conducting such a study is in an attempt to achieve at least four goals as follow in terms of (1) the appropriateness of TOEIC Bridge Test to Chinese vocational students and a research report to department of higher education for reference; (2) possible separation of teaching from testing; (3) improvement of test quality and (4) recognition of test certificate.
Sample
Students of Jiaxing University and other universities within Zhejiang Province.

Method
Comparison of TOEIC Bridge Test and other relevant English test in China using Gistest, Rasch-based software.

Results and Conclusions
The results thus obtained show that TOEIC Bridge test fits Chinese students better and the score interpretation better indicates Chinese students’ communicative competence in the real social and campus life. Evidences by comparison collected via questionnaires also indicate the possible reasons why the local parallel test is not very well accepted by educators and employers in China.
Investigating the Consequences of the Application of Formative Evaluation Reading-Writing Model

Presenter August 7th (15:35–16:00)
Laura Yang Hong*, Yan Zhao and Qiu Yulei
Faculty of Foreign Studies, University of Jiaxing
56 Yuexiu South Rd, Jiaxing, Zhejiang Province, China
e-mail*: yanghong_158@yahoo.com.cn

Background
Nowadays, we can see reading accounts for a large proportion in some large-scale English proficiency tests at home and abroad, such as CET-4, CET-6, TEM-4, GRE, IELTS, and TOEFL. Many studies show that a considerable number of Chinese students can get high marks but they have never developed true reading skill and one of the major reasons is that they lack adequate effective reading. How to combine reading and writing effectively in teaching has become an important research field in Second Language Acquisition (SLA) in China.

Aims and Keywords
The purpose of conducting such a study is in an attempt to answer the following questions: (1) Can formative evaluation be helpful in anchoring learners’ identity (attitudes and learning needs in reading)? (2) Can formative evaluation be helpful in students’ learning management (i.e. learning strategies)? (3) Can formative evaluation be helpful in promoting students’ writing and reading effectively?

Keywords Formative evaluation • Reading through writing model (RTWM)
• Consequence

Sample
There were 35 English Majors from Jiaxing University who had the course Extensive Reading. They were assigned an integrative task — choose to read some English newspapers and magazines two or three times a week and keep a notebook, and choose two original novels a semester and write book reports on storyline, theme, writing style, favorite hero(s)/heroines, comments, what they get, and so on.

Methods
This study adopts a comparative way, in which two freshman English major classes with similar motivation, strategy and proficiency, sex distribution and taught by the same teacher will be engaged. Class A, where RTWM is practiced, is the experimental group and Class B with the conventional assessment is the control group. The data elicited from interview, observation and students’ notebook are used for qualitative analysis and the data from questionnaires and reading and writing tests are used for quantitative analysis by using the software SPSS 15.0.
Results and Conclusions
The results thus obtained show that the application of formative evaluation to RTWM is of great significance in anchoring learners’ identity and improving students’ learning management in that it is the students who can tell the teacher what they have learned and what they are learning. Evidences by comparison collected via questionnaires and tests also indicate that formative evaluation be helpful in promoting students’ writing and reading effectively.
A Rasch Analysis on the Development and Validation of Mathematics Test for Use by Primary Five Student in Hong Kong

Presenter August 7th (15:10–15:35)
Jingjing Yao¹, and Magdalena Mo Ching Mok¹,²
¹Assessment Research Centre, Hong Kong Institute of Education
e-mail: jingjing@ied.edu.hk
²Psychological Studies, Hong Kong Institute of Education
e-mail: mcmok@ied.edu.hk

Background
This study was part of a large-scale quasi-experiment on the causal relations among feedback, self-directed learning, and mathematics achievement of primary students in Hong Kong. The larger study was sponsored by a General Research Grant awarded by Research Grants Council of UGC, Hong Kong (Project Number 844011).

Aims and Keywords
The study aimed to develop and validate, using a Rasch approach, of an instrument for measuring mathematics achievement of Primary five students.

Keywords Rasch model • Validity • Mathematics achievement scale • Primary student

Sample
The sample comprised 1,368 Primary 5 students from 16 Hong Kong schools who participated voluntarily.

Methods
A pre-test/post-test design was used for the larger study. Data for this study were collected at the pre-test in form of a mathematics test completed during class-time. A 35-item mathematics test was developed after careful analysis of the local curriculum, and in consultation with teachers on the suitability of the test for students by the end of Semester One of Primary 5. Rasch analysis with the Winsteps software (version 3.70) was then used to validate the instrument.

Results
The results showed that: (1) item-fit was between 0.5 and 1.5 for all items except one; (2) Eigen value of the first contrast of the Rasch PCA of Residuals was below 2.0; (3) Rasch person and item reliabilities were 0.80 and 1.00 respectively; (4) there was good alignment between item difficulty and student ability; (5) no gender-DIF was found in the items.
Conclusions
Measurement validity is one of the most crucial factors for the overall validity of any research study. This study developed a valid and reliable instrument for measuring mathematics achievement of Primary 5 students.
The Mixture Facets Model for Differential Rater Functioning

Presenter: Jin Kuan-Yu, Wang Wen-Chung
The Hong Kong Institute of Education
Hong Kong

August 6th (16:20–16:45)

Background
Essay items have been widely used in educational tests. Often, raters need to be recruited to mark essay items and raters may have very different degrees of severity. Furthermore, a rater may show different degrees of severity for different groups of examinees, which is referred to as differential rater functioning (DRF). In most DRF studies, the group memberships of examinees are known, for example, gender or ethnicity. However, DRF may occur when the group memberships of examinees are unknown.

Aims and keywords
This study aims to develop a new mixture facets model to assess DRF when the group membership is unknown.

Sample
Simulated datasets were generated with two latent classes by manipulating six conditions: (a) sample size; (b) group mean difference; (c) number of items; (d) number of raters; (e) magnitude of DRF; and (f) tendency of DRF.

Methods
The generating model was fit to the simulated datasets. The dependence variables were the bias and root mean square error of parameter estimates, and the accuracy of identification of latent classes. Twenty replications were made under each condition. A Bayesian method was used to estimate the parameters by using the WinBUGS freeware.

Results
The parameter estimation and the classification of latent classes were more accurate when the dataset was larger (i.e., larger sample sizes, longer tests, and more raters), the group mean difference existed, the differences in the rater parameters between latent classes were larger, and the pattern of DRF was balanced between groups.

Conclusions
Fitting the proposed mixture facets model is useful to explore the inconsistency of rater severity with respect to different latent classes.
Multiple Regression of Civic Knowledge Using Plausible Values

**Presenter**
Kuang Xiaoxue
PhD student in Hong Kong Institute of Education
e-mail: Kuangxiaoxue2006@126.com

**Professor KENNEDY, Kerry John**
Chair Professor of Curriculum Studies, Associate Vice President (Quality Assurance),
Dean of Faculty of Education Studies, Co-Director of Centre for Governance and Citizenship
e-mail: kerryk@ied.edu.hk

**Aims and Keywords**
The purpose of our study is to find out whether students’ views on values and attitudes underlying citizenship issue can be used as the variable to predict their civic knowledge.

**Keywords** ICCS 2009 • Plausible value • Multiple regressions

**Sample**
The data used for this study were retrieved from the ICCS database. The samples used in the International Study were also sued for the Asian Regional Module (ARM). Sample sizes for each participating society are shown in Table 1. The sample was drawn from Grade 8 students between the ages of thirteen and fourteen. The average age of the sample was 14.4 years with student ages ranging from 14.2 in Chinese Taipei to 14.7 in Korea.

**Table 1** Sample size for the five Asian Societies participating in the ARM

<table>
<thead>
<tr>
<th></th>
<th>Chinese Taipei</th>
<th>Hong Kong SAR</th>
<th>Indonesia</th>
<th>Korea, Republic</th>
<th>Thailand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>5,152</td>
<td>2,739</td>
<td>5,048</td>
<td>5,252</td>
<td>5,263</td>
</tr>
<tr>
<td>Valid</td>
<td>5,152</td>
<td>2,737</td>
<td>5,048</td>
<td>5,252</td>
<td>5,263</td>
</tr>
</tbody>
</table>

**Methods**
Multiple regression analysis has been used to interpret the relationship among the Asian Regional Module (ARM) and civic knowledge. The predicted variables are the three latent construct of the ARM—students’ perceptions of government and law in Asia, students’ perceptions of identity, citizenship, and culture in Asia students’ perceptions of public service for student’s civic knowledge. In order to do the analysis, the plausible value of the three predicted variables of five Asian countries are computed using ACER Conquest software (von Davier, Gonzalez, and Mislevy 2009). Since there are five plausible values, the multiple regression are
repeated five item. The final result is the average of the five regression coefficients and the variances are also computed using its formulation.

**Conclusions**
Except for Hong Kong, the three latent construct of the ARM account for approximately 20% of the variance of student’s civic knowledge. The students’ perceptions of government and law and students’ perceptions of public service in Asia have positive prediction effect for student’s civic knowledge. While the interesting thing is that the second latent variable—students’ perceptions of identity, citizenship, and culture in Asia has negative effect for student’s civic knowledge.
Implementing the DIF-Free-Then-DIF Strategy on DIF Assessment

Presenter Kun Xu*, Wen-Chung Wang and Magdalena Mo Ching Mok
The Hong Kong Institute of Education
e-mail*: xukunjacob@gmail.com

Background
A prerequisite of DIF assessment is a pure scale on which the performances of reference and focal groups can be compared. In practice, all items in a test have to be assessed for DIF, such that a pure scale is not feasible before DIF assessment. To tackle this question, the DIF-free-then-DIF strategy was proposed, where a set of DIF-free items are selected first and then all other items are assessed for DIF using the DIF-free items as anchored.

Aims and Keywords
This study aims to implement the strategy on three popular DIF detection methods: the Mantel-Haenszel method, the logistic regression method, and Raju’s area method, and compared their performances with traditional approaches.

Keywords Differential item functioning • Rasch measurement • Anchor item • Mantel-Haenszel method • Logistic regression method • Area method

Methods
Monte Carlo simulations were conducted. The five independent variables were: (a) method (Mantel-Haenszel, logistic regression, and area methods); (b) scale purification procedures: yes and no; (c) difference in mean ability between groups: 0 and 1; (d) percentage of DIF items in the test: 10, 20, 30 and 40 %; and (e) number of anchored items: 4 and 8. A total of 100 replications were made in each condition. The data were generated from the Rasch model and the DIF was uniform. The Type I error rate and power were computed.

Results
The traditional methods without scale purification procedures worked fairly well only when the percentage of DIF items was as small as 10 %. When scale purification procedures were implemented, the Type I error rates were well-controlled when the percentage was no greater than 30 %. The selection of a set of DIF-free items was very accurate, but not perfect. The DIF-free-then-DIF strategy was effective only when the percentage of DIF items was as high as 30 or 40 %.

Conclusion
Scale purification procedures should be implemented. The new strategy is helpful when tests have a high percentage of DIF items.
Differences in the Perception of Feedback Between Hong Kong Primary School Students and Their Teachers

Presenter August 6th (13:30–13:55)
Michael Ying Wah Wong¹, and Magdalena Mo Ching Mok¹,²
¹Assessment Research Centre, The Hong Kong Institute of Education
e-mail: mywwong@ied.edu.hk
²Psychological Studies, The Hong Kong Institute of Education
e-mail: mmcmok@ied.edu.hk

Background
Abundant research has been undertaken on the importance of teachers’ feedback to the school achievement of students. However, focus has rarely been placed on the differences between the perception on feedback of students and teachers. Being a part of the study which was sponsored by a General Research Grant awarded by Research Grants Council of UGC, Hong Kong (Project Number 844011), we attempted to find out the differences between students’ and teachers’ perception on feedback.

Aims and Keywords
The study aimed to find out the differences between Hong Kong primary school students and teachers in their views to feedback.

Keywords Rasch model • Effect size • Feedback • Primary student • Primary school teacher

Sample
Four thousand five hundred and seven students between primary 3 and primary 5, 132 teachers from 26 Hong Kong primary schools were sampled in the study.

Method
A 42-item instrument with 4-Likert point response scale was constructed by the research team to collect the responses from Hong Kong students and teachers of their perception on 6 dimensions in feedback. Multidimensional Rasch analysis was conducted on the collected data with CONQUEST 2.0.

Results
The results showed that: (1) item-fit was between 0.5 and 1.5 for all items except one; (2) Rasch person reliabilities in the 6 dimensions ranged from 0.72 to 0.86 and item reliabilities was 1.00; (3) the effect size of the respondents’ status (teacher or student) in the 6 dimensions ranged from 0.32 to 0.64.
Conclusions
In this study, there are moderate differences between the students and teachers in the perception on the source of feedback, the effectiveness of feedback and the function of feedback. On the other hand, there are little differences among the two types of respondents in the perception on the goodness of feedback, the focus/target of feedback and the expectation of feedback.
Background
Sensitivity and threshold are two important elements in making judgment (Jackson 1972). When raters are recruited to mark responses to constructed-response items, the effect induced by raters should be fully considered. However, the standard facets model (Linacre 1989) or generalized facets model (Wang and Liu 2007) accounts for the threshold element (rater severity) and does not consider the sensitivity element.

Aims and Keywords
Rater sensitivity can be defined as the effectiveness of a rater in differentiating ratees with varies degree of proficiency. By treating the combination of an item and a rater as a pseudo-item, we intend to decompose the attached slope (discrimination) parameter on the pseudo-item into two parts: item discrimination and rater sensitivity.

Sample
A dataset gathered by Congdon and McQueen (1997) was analyzed, in which each of the 8,296 students’ writing scripts was graded by two raters randomly chosen from a set of 16 raters on two criteria (items) against a six-point scale.

Methods
Different formulations of facets models were fitted by using the freeware WinBUGS.

Results
The results supported our expectation: the two rating criteria had degrees of discrimination power; the 16 raters had different sensitivities; and the slope of the pseudo-items was approximately the product of item discrimination and rater sensitivity.

Conclusions
This resulting generalized facets model can account for item characteristics (difficulty and discrimination) and rater properties (severity and sensitivity) simultaneously.
Construction of the Structural Q-Matrix of the Tactics Used in Badminton Singles Games

Presenter: Henry Hoi-Wai Wong
The Hong Kong Institute of Education

Background
Mr Wong Hoi Wai is interested in the researches of cognitive measurement in physical education, by implementing various cognitive diagnosis model such as DINA model, Fusion model and etc. Mr Wong is now a PE teacher in Hong Kong C.C.C. Kei Wai Primary School and concurrently a doctoral student in the Hong Kong Institute of Education.

Aims and Keywords
The main objective of this research is to explore the underlying attributes of the tactics that are frequently used by the world-class badminton singles players. This study is going to construct a set of structured attribute patterns (Q-matrix) associated with different components from a cognitive model of task performance (Linear Logistic Latent Trait Model).

Sample
This research is based on the videotaped badminton singles games held in BWF World Championship 2011. There are in total 16 games, including Quarter-Final, Semi-Final and Final Men’s and Women’s Single games.

Methods
Though the notational analysis on all badminton singles games and the extensive literature reviews on badminton tactics, all possible tactics will be tagged with specific attributes. For each test item, all the prerequisite information, including the positions of the opponent and the player, the velocity of the shuttlecock and the flight height of the shuttlecock will be noted. The skills that the player used will be recorded with labeling “1” for presence or “0” for absence. It is hypothesized that a particular response pattern is correspond to a particular underlying tactics. The structural matrix will be entered for analysis by the software R with extended Rasch modeling (eRm). Refining the structural matrix is needed until the goodness-of-fit is obtained.

Results
Analyses are underway. The initial result will be obtained by the end of April. The final result will be presented at PROMS conference. The initial result will be sent to PROMS once obtained.
Conclusions
Physical Education has become emphasized the tactics teaching in lessons. However, there are no substantial assessment methods for tactic-based teaching methodology. Once the ideal response patterns of tactics used by the badminton singles players are constructed, PE teachers and Sport coaches can make the diagnosis on the used tactics by using the DINA model to compare the ideal response patterns and the observed response patterns. Coaches can then monitor the competition process of player and provide the specific feedbacks to players with full substantial reasons. Players can learn from this to self-regulate their skill uses during the badminton singles games.
The Facets Graded Response Model

Presenter: Joseph Chow; Kuan-Yu, JIN; Wen-Chung, Wang  
The Hong Kong Institute of Education,  
HKSAR, China  
e-mail: chowkf@ied.edu.hk

Background
Raters are often recruited to mark constructed-response items. For example, student essays are graded by teachers; or teachers’ performances are judged by students. Linacre (1989) has developed the facets model to account for the joint effects of item, person, and rater on item responses using the conditional-probability formulation. In the statistical literature, the latent-response formulation is used more often. The latent-response formulation is in line with the cumulative-logit models, which incorporate the graded response model (Samejima 1969) and the 2-parameter logistic model (Birnbaum 1968) as special cases.

Aims and Keywords
In this study, we aim to extend the two-facet graded response model to more than two facets so that rater effects can be evaluated. Simulations were conducted to evaluate its parameter recovery using the Mplus program.

Keywords: Graded response model • Facets model • Cumulative-logit models • Parameter recovery • Mplus

Methods
The item responses were generated under the three-facet graded response model, where 5 raters judged 200 examinees on 5 tasks on a five-point rating scale. A total of 100 replications were made. The true model was used to analyze the data using the Mplus. The bias and root mean square error (RMSE) were computed to evaluate parameter recovery.

Results
The bias values were (−0.154 ~ 0.186), and the RMSE values were (0.043 ~ 0.215). Apparently, the parameter recovery of the three-facet graded response model was satisfactory.

Conclusion
The two-facet graded response model is extended to facets graded response model to examine rater effects on open-ended items. The parameters can be recovered fairly well using the Mplus. Applications of the new model are welcome.
Using Combining Software to Investigate the Relationship Between Rated and Counted Features of Spoken and Written Performance in English Language

**Presenter**  
Dr Cheung Kwai Mun Amy  
Manager of Hong Kong Examinations and Assessment Authority (HKEAA)

Ms LEUNG Fung Yin Flora  
Senior Curriculum Development Officer of Education Bureau(EDB), Hong Kong

**Background**  
This study attempted to analyse students’ performance in speaking and writing at Secondary 3 (Grade 9) in Hong Kong.

**Aims**  
This study aimed at using combining software to investigate the relationship between rated and counted features of spoken and written performance in English Language.

**Sample**  
A stratified sample of 180 students.

**Methods**
- students’ performances rated by teacher raters and also subjected to verifiable quantitative measures (VQM) (Cheung 2010)
- obtained ‘fair average scores from rater’ ratings using FACETS (Linacre 1991–2008)
- calculated the VQM using *vocd* (Malvern et al. 2004) to measure ‘lexical diversity (D)’, and *RANGE* (Heatley et al. 2002) to measure ‘types’, ‘families’, and ‘tokens’
- compared ratings against each other and against VQM to ascertain the reliability and validity of ratings
- key questions: (i) To what extent do the sub-constructs within a skill and across two skills correlate? (ii) Do students with strong written vocabulary exhibit strong spoken vocabulary, and vice versa?

**Results**  
The main findings showed ‘high’ correlation levels in ‘vocabulary and language patterns’ across spoken and written skills.
**Conclusion**
Transference from writing to speaking seems more likely given the prioritization of the written mode in the Hong Kong school system and the fact that Hong Kong students usually do not have much chance to practise oral English outside school. The next phase of the study would be to develop corpus-based human assisted error identification and classification system to address learning needs in productive skills.
Effect of Academic Aims, Goal Setting and Planning on Academic Achievement of Secondary Students in Hong Kong

Presenter August 8th (9:25–9:50)
Mr. Jinxin Zhu
Assessment Research Centre, Hong Kong Institute of Education
Hong Kong
e-mail: jxzhu@s.ied.edu.hk

Prof. Magdalena Mo Ching MOK
Chair Professor, Psychological Studies Department
Hong Kong Institute of Education
Hong Kong
e-mail: mmcmok@ied.edu.hk

Background
This study is part of a larger study entitled “Secondary Students’ Independent Learning” funded by a Competitive Earmarked Research Grant (Number HKIED 8005/03H) of the Research Grants Committee, Hong Kong SAR Government. The original project aimed to establish and validate with empirical data, a conceptual framework of self-directed learning, the characteristics and processes involved and their relations with academic achievement of secondary students.

Aims and Keywords
This study aimed to explore the effect of academic aims, goal setting and planning on academic achievement of secondary students.

Keywords Rasch measurement • Academic aims • Academic goal setting • Academic planning • Academic achievement • Secondary students

Sample
The sample comprised 14,846 students currently enrolled at Secondary 1 to Secondary 6 (except Secondary 5) from 23 secondary schools in Hong Kong.

Methods
Students completed a self-administered questionnaire comprising 4-point Likert items (1: Strongly Disagree, 2: Partly Disagree; 3: Partly Agree, 4: Strongly Agree) on academic aims, goal setting, planning, and other aspects of self-directed learning not included in this study. Each construct was measured by five items. Measurement quality of the scales according to the framework was established using the
Winsteps software (version 3.72.3). Based on Rasch scores of the latent variables and standardized scores of Chinese, English and mathematics, structural equation modeling was undertaken using Mplus (version 6).

Results
Analyses indicated that academic aims and academic planning both directly predicted academic achievement, with educational aims being the stronger predictor. Academic goal setting indirectly predicted academic achievement via academic planning. Consistency was found both in the structure and in the effects of the predictor across gender and year levels.

Conclusion
The study shows that self-directed learning has positive effect on academic achievement of secondary students of both gender and all year levels.
Rasch Validation of a Combined Measure of Basic and Extended Daily Life Functioning After Stroke

**Presenter**
Hui-Fang Chen  
Post-doctoral Fellow, Assessment Research Centre  
Hong Kong Institute of Education, Hong Kong  
e-mail: hfchen@ied.edu.hk

Ching-yi WU  
Professor, Chang Gung University, Taoyuan, Taiwan  
e-mail: cywu@mail.cgu.edu.tw

Keh-chung LIN  
Professor, National Taiwan University, Taipei, Taiwan  
e-mail: Kehchunglin@ntu.edu.tw

**Background**
Tools used to measure poststroke functional status must include basic and instrumental activities of daily living (ADL) and reflect both the patient’s and the clinician’s perspectives.

**Aims and Keywords**
This study combined the Functional Independence Measure (FIM) and the Nottingham Extended Activities of Daily Living (NEADL) to create a scale providing objective and subjective evaluations of ADL function after stroke.

**Keywords**  Activities of daily living • Psychometrics • Rasch analysis

**Samples**
A total of 188 participants, who completed the NEADL and the FIM, was included in the present study. Those patients met the inclusion criteria: (1) first-ever stroke, (2) Brunnstrom stage II or above for the proximal and distal upper extremity, (3) no severe physical conditions and medical problems, (4) no cognitive impairment, and (5) no excessive spasticity at any joint of the arm.

**Methods**
Rasch analysis was conducted to investigate the psychometric properties of the new scale.
**Results**
A 3-point and a dichotomous scale were suggested for use in the FIM and the NEADL, respectively. The combined 40 items worked consistently to reflect a single construct, and “bladder management” and “bowel management” were highly related. After “bowel management” was removed, all but 3 items fit the model’s expectations and showed reasonable item difficulty hierarchy with high reliability. However, the 3 misfit items were removed, and no differences were found between the 36-item and 39-item scales.

**Conclusions**
The combined measure provides a comprehensive picture of ADL from patients’ and clinicians’ perspectives. It extends the utility of the FIM and the NEADL and is recommended for use to measure independence of stroke patients.
Developing the Content of Certification Programme for Infection Control Nurses (ICNs) in Hong Kong

Presenter: August 6th (16:20–16:45)
Chan WF
Advanced Practice Nurse, Infection Control Team
Tung Wah Eastern Hospital, Hong Kong SAR
e-mail: chanwf3@ha.org.hk

Adamson B
Head, Department of International Education and Lifelong Learning,
The Hong Kong Institute of Education, Hong Kong SAR
Bond T, Adjunct Professor, School of Education, James Cook University, Australia
CHUNG JWY, Peter TC Lee Chair Professor of Health Studies,
Department of Health and Physical Education, The Hong Kong Institute of Education, Hong Kong SAR
CHOW MCM, Assistant Professor, School of Nursing,
The Hong Kong Polytechnic University, Hong Kong SAR

Background
Nursing specialization is the world trend in healthcare development. Certification is one of the means to recognise the nurse specialists. There is no certification system in nursing field in Hong Kong. Participating in the certification examination in the overseas does not address the problem because the practices of nurse specialists are local-context specific.

Aims and Keywords
The aim of this research is to develop the content of certification programme for ICNs in Hong Kong.

Keywords: Infection control nurse • Certification • Hong Kong

Sample
Expert panels, 18 ex-ICNs and 117 ICNs in Hong Kong

Methods
This is a 3-phase research. Phase One proposed the core competency of ICNs by literature review and Delphi survey. Validity and reliability were established employing content validity survey by experts and repeated surveys on ex-ICNs. Phase Two confirmed the core competency of ICNs through an opinion survey on practicing ICNs. Data were analysed using Winsteps programme. A qualitative questionnaire survey on experts was conducted to define the critical competency of
ICNs in Phase Three. With Rasch-based concept, safety margin was added as resulted in true critical competency. The items were rescaled and importance levels (in logit) were transferred to content weight to develop the content blueprint of certification programme for ICNs of Hong Kong.

**Results**
The core competency of ICNs identified in Phase Two consisted of 76 items with importance levels (in logit) attached in individual items. The competency scale was unidimensional with very good reliability estimates. The critical competency was the most important portion of the core competency. The true critical competency comprised 35 items, including 25-item expert-defined critical competency and 10-item safety margin.

**Conclusions**
The Rasch analysis concluded the core competency of ICNs in Hong Kong by identifying the items objectively and deciding the respective importance levels. Based on the expert-defined critical competency, Rasch-based concept was employed to ensure the inclusion of essential items in the critical competency that serving as the content blueprint of certification programme for ICNs. This is the first reported work on developing the content of certification programme for nurse specialists in Hong Kong. This study can be replicated in other nursing specialties for developing the content of respective certification programmes. It also served as the ground work for nursing professional development in future.
Applying the Rasch Model to Diagnostic English Language Testing

Presenter August 7th (16:25–16:50)
Jerry Gray¹, Gwendoline Guan², (email: ygguan@cityu.edu.hk)
Michelle Raquel³, Winnie Shum¹, Carrie Tsang³, Alan Urmston³, Roxanne Wong²
¹Lingnan University
²City University of Hong Kong
³The Hong Kong Polytechnic University

Background
In 2009, three Hong Kong institutions shared their expertise and resources for the purpose of developing and establishing a diagnostic assessment to enhance students’ English language learning. They aimed to develop an instrument of the highest quality in terms of reliability, validity and usefulness. Based on Rasch measurement principles, they developed the Diagnostic English Language Tracking Assessment (DELTA). Through the DELTA, students can diagnose their strengths and weaknesses in reading, listening, grammar and vocabulary; track their English language gains; and plan their English language learning.

Aims and Keywords
This research aims to investigate the application of Rasch measurement to the diagnosis of English language proficiency.

Sample
Two thousand and five hundred Year one undergraduates, the DELTA.

Methods
Rasch measurement, Winsteps software package.

Results
- Calibrated person measures to track growth over time;
- DIF analysis to demonstrate students’ strengths and weaknesses on reading, listening, vocabulary and grammar;
- An adapted KIDMAP to show individual performance on a set of items on each component of the test

Conclusions
The DELTA test is a valid and reliable instrument that measures students’ English language proficiency, and the Rasch Model can be applied to diagnostic language testing.
Constructing Rasch-Based Measures for Assessing Academic Attainment for Students with Intellectual Disabilities

Presenter August 7th (13:25–13:50)
Vicky TSang; Trevor G. Bond; Joseph Chow
The Hong Kong Institute of Education, HKSAR, China
James Cook University, Australia
e-mail: chowkf@ied.edu.hk

Keywords Rasch measurement • Scale construction • Students with intellectual disabilities • Curriculum-based assessment • Evaluation tool

Background
Since 2005, professionals in nine special schools collectively had developed an assessment tool called SCALES for all students regardless they are in special or mainstream schools in Hong Kong. Based on the concept of “one common curriculum for all”, SCALES is classified into the four Key Learning Area (KLA) strands of learning as in the mainstream curriculum framework, namely, Chinese, Science, Personal, Social and Humanities Education (PSHE) and Mathematics. This means that for the first time all of the students are being measured by exactly the same criteria irrespective of the ability or disability.

Aims
In this study, we aim to apply Rasch measurement approach to analyzing a this innovative evaluation tool for children with special needs due to intellectual disabilities. Since Rasch measurement can convert ordinal measures into interval measures, the latter of which is fundamental for objective measurement in human sciences (Bond and Fox 2007).

Methods
By using Winsteps, statistics such as infit and outfit mean square, person separation index, reliability, and dimensionality of the measurement scales are supported empirically.

Results
The uni-dimensionality of the four major subject areas is supported with good psychometric properties such as high reliability and person separations that help differentiation of subjects. Selective applications of some items of these scales could be used by practitioners for quick screening and better intervention for care of children with special leaning needs.
Conclusion
The scales are supported with both subject experts’ observation and quantitative evidence from sophisticated psychometric tests. Successful establishment of the instrument by subject experts is evident and quality-assured administration of the Rasch-based items in the future is recommended as the SCALES has proved to be a valid and reliable assessment measure for diverse learners.
Identification of the Patterns of Chinese Character Recognition in Students with Learning Disabilities Requiring Tier-2 Support: A Rasch Analysis

Presenter
Fuk-chuen HO and Zi, Yan
Hong Kong Institute of Education

August 7th (15:35–16:00)

Background
Dr Fuk-chuen Ho is Assistant Professor in the Department of Special Education and Counselling and the associate director in the Centre for Special Needs and Studies in Inclusive Education at the Hong Kong Institute of Education. He was formerly an inspector in the Special Education Inspectorate of the Hong Kong Education Bureau. He is now the project leader of three external funded projects in the areas of dyslexia, Theory of Mind and collaborative mode of professional development for teachers in special schools respectively. (fcho@ied.edu.hk)

Dr Zi Yan is Assistant Professor in the Department of Curriculum and Instruction at the Hong Kong Institute of Education. He has a research interest in educational measurement and inclusive education. He is now participating in a commissioned project on the evaluation of the inclusive education in Macau. (zyan@ied.edu.hk)

Aims and Keywords
This study investigates the Chinese reading patterns of students with learning disabilities (LD). The performances of students with LD in reading the three categories of Chinese characters were particularly analyzed: regular, irregular, and pseudo-characters.

Keywords Learning disabilities • Chinese character recognition • Regularity

Sample
Fifty-three students with LD in reading and 44 average students at age 9 were selected from five Hong Kong primary schools.

Method
Their abilities for reading Chinese characters were measured using Rasch analysis.

Results
Both types of students found regular characters as the easiest to read. Average students showed better performance in reading irregular characters than pseudo-
characters, whereas students with LD exhibited no significant performance difference in reading these two categories.

**Conclusions**

Students with learning disabilities and students without disabilities have different preferences for the phonological and orthographic routes to read Chinese characters.
Applying the Many-Facet Rasch Model to an Exit English Speaking Proficiency Assessment

Presenter August 6th (13:55–14:20)
Felicia Fang, Alan Urmston
The Hong Kong Polytechnic University
e-mail: Alan.Urmston@inet.polyu.edu.hk

Aims and Keywords
This research aims to investigate the application of Many-facet Rasch measurement in the reporting of test scores to students and the university, the analysis of the performance of the tasks and criteria and the performance of raters in a university-level exit English assessment of speaking in Hong Kong.

Sample
Approx. 3,000 students who sat the assessment in academic year 2011–2012, 30+ raters, 5 tasks.

Methods
Many-facet Rasch analysis was conducted using the FACETS Programme (Linacre and Wright 1993).

Results
The major findings of the study:

- FACETS proved to be a reliable instrument in providing a clear picture of how students performed on this test, how well the tasks performed, and how reliable the raters were when rating the performances
- The all facet vertical rulers
- Candidate measurement report: candidates at different levels were clearly distinguished
- Rater measurement report: misfitting (inconsistent, and/or overly harsh/lenient) raters were identified
- Task measurement report: how well the tasks performed
- Item measurement report: how well each criteria on each task performed
- The separate rater report: provided information on rating consistency, the rater’s individual use of the levels (e.g. whether clear discrimination between levels was achieved, whether any particular level was under representation, etc.)

Conclusions
The results informed future rater accreditation and training, and the writing of new test tasks.
Implication of Differential Item Functioning in Investigating the Possible Effect of learners’ Identity on the Underlying Structure of Listening Comprehension Ability

Presenter
Parisa Daftarifard
A PhD candidate at Azad University (Science and Research Branch)
Tehran, Iran
e-mail: pdaftaryfard@azad.ac.ir

Minoo Alemi
A PhD candidate of Applied Linguistics at Allameh Tabataba’i University and a faculty member of Languages and Linguistics Department at Sharif University of Technology, Iran
e-mail: alemi@sharif.ir

Background
Second Language (L2) Self identity gains a focal attention in research enquiries nowadays (for example, LoCastro 2001) though filtering hypothesis (Krashen 1981) and research on good language learner (GLL) specifically when culture is the main focus (Finkbeiner 2008) and when belief is attributed to GLL (White C 2008). We hypothesized that items of listening comprehension might function differently in terms of learners’ level of identity. To test this, one way is to use the method of differential item functioning (DIF). DIF studies mostly focus on age, gender (Maller 2001) different background knowledge (Pae 2004), different first languages (Ryan and Bachman 1922) but to our knowledge, there is not much research done on investigating the effect of different identity type on test performance. To this end, the following questions were posited.

- What is the relationship between L2 self identity and different levels of listening ability?
- Does L2 self identity can predict learners’ score on different listening sub-skills?
- Are there some different target populations in terms of identity in the focal group?

Sample
About 60 Iranian students of Engineering at Sharif University participated in this study.

Method
The method of this research is ex post facto because no treatment is provided and no manipulation is done. The data is gathered through two instruments: the listening
section of TOEFL and a questionnaire of self identity adapted from LoCasro (2001). Data is analyzed through both Classical True score theory and Rasch model using Winstep.

**Result**
The result revealed that not all sub skills of listening comprehension are affected by Iranian learners’ identity.
Learning by Assessing in Second Language Writing

Presenter: Trevor A. Holster  
Fukuoka Women’s University  
e-mail: trevholster@gmail.com

Bill Pellowe  
Kinki University  
e-mail: pellowe@fuk.kindai.ac.jp

J. Lake  
Fukuoka Women’s University  
e-mail: jlake@temple.edu

Aaron Hahn  
Fukuoka High-school  
e-mail: qwyrxian@hotmail.com

Background
Recent studies have used many-faceted-Rasch-measurement (MFRM) to investigate the potential of peer assessment (PA) in second language classrooms. However, “learning by assessing” (LBA), where the assessment process results in learning by raters, has not been widely investigated.

Aims
This study aimed to demonstrate LBA in second language writing by testing the research hypothesis that gains measured on a PA rubric would exceed gains on a secondary rubric measuring general academic writing.

Keywords  Peer assessment • Learning by assessing • Writing proficiency • Second language assessment

Sample
A convenience sampling of 26 first year university students in an academic English program participated as part of a compulsory academic writing course.

Method
Pre and post treatment writing was rated by teachers on the rubric used for PA and a general academic writing rubric, using the Facets software package. Students were exposed to the PA rubric, but not to the secondary rubric. The MOARS audience response system allowed students to rate each others’ work and to access their own ratings instantaneously using internet connected mobile devices.
Results
Substantive gains were observed on the PA rubric but not on the secondary rubric, providing evidence of LBA. Response patterns suggested holistic rating by peer raters, casting doubt on the potential of PA to provide formative feedback and supporting the view that the gains observed arose from LBA.

Conclusions
The holistic response patterns displayed by peer raters discounted PA as a source of diagnostic feedback. Instead, the results suggest that the major benefit of PA lies in improving peer raters’ knowledge of the target construct through repeated engagement with the rubric during the rating process.
Metalinguistic Knowledge of Low-Proficiency EFL Learners at a Japanese University

Presenter
Miki Tokunaga
Fukuoka University
e-mail: tokunagamiki@gmail.com

Background
Recent studies have shown correlations between learners’ L2 proficiency and their L2 metalinguistic knowledge. Although many of the studies were conducted on learners with intermediate or higher levels of proficiency, a much simpler test with only high frequency words is necessary to measure metalinguistic knowledge of low proficiency EFL learners.

Aims and Keywords
This research aimed to design a simple metalinguistic test to find out what metalinguistic features are recognized by low proficiency EFL students and whether there is a correlation between their English proficiency and metalinguistic knowledge.

Keywords Metalinguistic knowledge • Explicit knowledge • Remedial education

Sample
The participants were 708 mixed major students taking required English classes at a private university in Japan.

Methods
A metalinguistic test was designed using simple L2 vocabulary to minimize the effect of learners’ vocabulary knowledge. The items aimed to test basic metalinguistic knowledge by identifying parts of speech, parts of sentence patterns, tenses and word structures. Winsteps was used to analyze the learners and items and results were compared with proficiency test (TOEIC Bridge) results.

Results
Moderate correlations were found between the Rasch measures of the metalinguistic knowledge and all sections of the TOEIC Bridge scores. The strongest correlations was with the Reading section ($r = .66$). Many of the participants had difficulty identifying simple metalinguistic features.
Conclusions
The results suggest that many of the subjects have insufficient metalinguistic knowledge to understand explanations by teachers and textbooks. Many universities in Japan now offer remedial level English classes teaching basic grammar and vocabulary. However, simplified content alone does not solve the problem when learners cannot understand instructions. Understanding students’ metalinguistic knowledge and providing targeted remedial instruction is thus necessary.
Mobile Audience Response System for Peer Assessment

Presenter
Bill Pellowe
Kinki University
e-mail: pellowe@fuk.kindai.ac.jp

Trevor A. Holster
Fukuoka Women’s University
e-mail: trevholster@gmail.com

Background
Previous studies have investigated peer assessment (PA) in second language classrooms, but the very large datasets generated by PA limit its practicality for regular classroom use. However, internet-connected mobile devices allow students to enter PA ratings directly into an online database, allowing for immediate analysis and feedback in the classroom.

Aims
This project aimed to develop a peer assessment module for the open source Mobile Audience Response System (MOARS), allowing users to enter ratings using any mobile device with a wireless internet connection. The MOARS PA module provides output in the form of a specification file and data file formatted for immediate analysis using the Facets or Minifacs software packages, allowing users with minimal technical skills to conduct MFRM analysis.

Sample
This pilot study comprised approximately 170 first year students enrolled in presentation skills classes in an academic English program at a Japanese university.

Method
The MOARS PA module was tested operationally. Students and teachers rated training videos, providing linking data to equate disjoint class groups. Students then rated each others’ live presentations in class. Data was collected using MOARS and analyzed using Facets, providing equated ability measures for all students in the 12 class groups and diagnostic analysis of students and teacher raters.

Results
No major problems were encountered with the MOARS system. Communicating diagnostic results to teachers and students was the major difficulty, requiring development of graphic outputs to supplement those available from Facets.
Conclusions
The combination of MOARS and Facets resolved problems of the practicality of collection and analysis of PA data in the classroom. The next stage of development will focus on simple graphical outputs to simplify diagnostic analysis for classroom teachers.
Can Difficulty of Items Be Guessed Intelligently Without Degrading CAT Results?

Presenter
Tetsuo Kimura  
Niigata Seiryo University  
e-mail: tetsuo.kmr@gmail.com

Keizo Nagaoka  
Waseda University  
e-mail: k.nagaoka@waseda.jp

Background
The authors constructed item banks for in-house computer adaptive tests (CATs) for EFL learners (Kimura 2009; Kimura and Nagaoka 2010). After several pretesting and equating by common item method, the number of item reached 258 for vocabulary and grammar and 307 for listening comprehension. The Rasch-based computer adaptive test program named UCAT written in BASIC (Linacre 1987) was converted into PHP so that CATs can be administered on a major open source learning management system (Kimura, Ohnishi and Nagaoka 2012).

Aims, Sample & Method
Linacre (2000) argued that the difficulty level of the new items can be guessed intelligently without degrading the resulting ability estimates in CAT. The aim of this study is to confirm this argument by comparing the results of two CATs administered to the same test takers (59 Japanese freshmen of engineering department). Both CATs used the same 258 items for vocabulary and grammar. The difficulty level of items of the first CAT was determined by pretesting ($M = 0.19, SD = 1.37$). That of the second CAT was guessed intelligently either one of four levels ($-1.5, -0.5, 0.5, $ or $1.5$). In either test, the initial estimate ability was set as 0.0 and 16 items were selected adaptively. The first item was selected randomly between $-0.5$ and $0.5$. The next item was selected randomly between lower limit (LL: ability estimate when the answer would be wrong) and upper limit (UL: ability estimate when the answer would be correct). If no item was found between LL & UL, the closest one was selected.

Keywords  Computer adaptive test • Item difficulty • Rasch

Results & Conclusions
The ability estimate averages of the two CATs were not significantly different (0.81 and 0.66). The standard errors were almost identical (0.53 and 0.54). The correlation coefficient between the two measurements was high (0.83). Consequently the current study suggests that the difficulty of items can be guessed intelligently without degrading the CAT results.
Abstract Proposal for PROMS 2012

Presenter
Aaron Olaf Batty
A visiting assistant professor at Keio University’s Shonan Fujisawa Campus, in the Department of Environment and Information Studies.

Background
English verbs of utterance (e.g. “speak,” “talk,” “say,” and “tell”), despite being basic vocabulary items, exhibit a wide range of uses by native speakers which are not predictable based on dictionary word meaning alone (e.g. “talk politics”). However, it is possible that as learners have more exposure to the language that they become more accurate and/or more confident with these special uses.

Aims and Keywords
The present research investigates the interactions between level, accuracy, and confidence with regards to special uses of the utterance verbs “speak,” “talk,” “say,” and “tell.”

Keywords  MFRM • English linguistics • Accuracy vs. confidence

Sample
Japanese high school students (n = 22), university students (n = 140), and native speakers living in the USA, the UK, and Japan (n = 15; N = 177).

Method
A vocabulary test and vocabulary questionnaire was administered on paper to the non-native speakers of English during normal classtime or online at their leisure. The questionnaire required participants to select the correct utterance verb to complete a sentence, and then indicate their degree of confidence for their answer. These data were scaled in Facets (Linacre, Facets. Beaverton: Winsteps.com. Retrieved from http://www.winsteps.com/facets.htm, 2011) and compared via the methods developed by Paek et al. (A study of confidence and accuracy using the Rasch modeling procedures (Research Report No. RR-08-42). Princeton: Educational Testing Service, 2008).

Results
Overall, even high-level learners of English demonstrated ignorance of many of the common constructions tested. The verb “talk” proved especially difficult, even for non-native speakers who had spent years living in Anglophone countries. Generally, confidence tracked accuracy weakly.
Conclusions
Some core meanings of verbs in English, and the “exceptional” uses they cause, may need to be explicitly taught, as it seems that learners do not naturally acquire these, even with a high degree of exposure to native speakers.

References
Using Student Experience for Class Composition

Presenter
Jeffrey Durand and Forrest Nelson
Associate Professors at Tokai University
Japan
e-mail: kandajeffd@gmail.com

At this large Japanese university, third-year students were streamed into an English class according to weighted test scores, with some difficulty from missing scores. For the most part, this method produced reasonably homogeneous classrooms according to ability, but two to five students sometimes appeared to belong in a different level. The purpose of this research is to develop a method of better separating students for purposes of class placement. Accordingly, information on students’ study experiences prior to university was used. The goal is to quickly obtain relevant, useful information to better construct classes. All students from the third-year course were used. A small number of questionnaires were not returned, mainly from students who never attended the class. Students were placed into classes according to test scores. On the first day of class, they completed a questionnaire on their experiences studying English. Winsteps was used to fine-tune the questionnaire and obtain experience scores for each student. Winsteps was also used to both compare and combine the experience items with the test scores. In addition, each teacher responded to a questionnaire during the first class and later in the year on whether students were appropriately placed. In fine-tuning the questionnaire, some kinds of experience “did not matter” in connection with student ability. More importantly, the experience scores improved student separation for streaming purposes. However, questions remain about the unidimensionality of this ‘ability to succeed’ construct. Using information on student experience in a Rasch analysis can greatly improve student placement. The unidimensionality issue has brought up a broader issue, however, of what the goals of class streaming should be: what kind of homogeneity is desired.
Using Rasch Model Analysis to Investigate and Compare Teachers’ Conception of Formative Assessment

Presenter Sharifah Norsana Bt Syed Abdullah
e-mail: eshalbiha@gmail.com

Mohamed Najib B Abdul Ghaffar
e-mail: p-najib@utm.my

Malaysia has been hard at work to transform its education system. In line with the Education Transformation Programme (ETP) to empower the education system, government had introduced a new assessment system which is more focus on outcome-based education and not be too examination oriented. A movement in assessment paradigms from measuring the amount of learning to enhancing learning which focused on more contextualized, communicative, performance-based as well as authentic assessment. Therefore, in this new assessment system more emphasis will give to classroom assessment rather than standardize test. Teachers must be prepared to integrated assessment with instruction in classroom. Teacher assessment plays a useful role in any assessment system because it will provide specific feedback on the progress of individual pupils to feed into teaching and learning and informs about individual strengths and weaknesses.

A descriptive study was conducted to identify the conception of formative assessment among teachers in Gombak district, Selangor. The main objectives of this study were to investigate teachers’ conception of formative assessment and to compare teachers’ conception in different teaching levels and subject areas. The Rasch rating scale measurement model was applied to the responses of primary, secondary and high school teachers (N = 150) from 6 schools in Gombak district, Selangor. Analysis of Differential Item Functioning was performed to compare the teachers’ conception of formative assessment according to subject areas and teaching levels. The results show that conception of formative assessment of social science and science teachers, primary, secondary and high school teachers are different. Analysis of the data revealed that the TFAC is a valid and reliable instrument to measure teachers’ formative assessment practices. The findings provided the Ministry of Education with useful information for training of school teachers in formative assessment practices in classroom.
Rasch Analysis of Goal Orientation Test in a Heterogeneous Setting

Presenter
Monsurat Olusola Mosaku, and Mohd Najib Ghafar
Department of Educational Foundation
Faculty of Education
Universiti Teknologi, Malaysia
e-mail: mabaqo@ymail.com, p-najib@utm.my

Higher education providers of the knowledge based economies contend with the challenges of internalization and diversification in order to be globally relevant and competitive. Assessment in contemporary higher education is expected to measure and evaluate cognitive and non-cognitive learning outcomes as both make up the Key Performance Index used to appraise these institutions. Students consequently formulate a mechanism in order to achieve both learning outcomes. Goal orientation implying students’ disposition at allocating varying amount of effort and time in order to achieve their aims is one of such mechanisms. Abundant literature asserts the implication of goal orientation in the successful learning experience and outcome of higher education. However, can it be generalized to a heterogeneous setting? This study aims to develop a valid instrument interpreting goal orientations among undergraduates in a heterogeneous setting. For this purpose, an exploratory quantitative research design using survey technique will be implemented among 254 undergraduate students in a multi stratified sampling procedure to represent the population of a university. Content, construct, predictive validities will be determined using Partial Credit Model of WINSTEP. Analysis resulted in variation based on discipline and religious beliefs while no significant effect was based on gender. This study will help to understand and appreciate the heterogeneous nature among Asian undergraduates.

Keywords  Goal orientation • Higher education • Rasch analysis
Validation of Fractions Skills Test for Primary Students

Presenter
Shafiza Mohamed
Malaysian Examinations Council

Kamisah Osman
National University Of Malaysia

Background
Most of the students struggle to understand fractions as well as individual adults. Proficiency in fractions is crucial and useful in everyday life especially in the measurement context.

Aims and Keywords
The purposes of this study are to validate and revalidate the fractions skills test amongst the primary students.

Keywords Fractions skills • Primary students • Conceptual knowledge • Procedural knowledge • Rasch model

Sample
Total of sample was 160 students (78 boys and 82 girls) from Year Five (aged 11 years) and Year Six (aged 12 years) in Malaysia.

Methods
A fractions skills test (83 items) was administered using paper and pencil. The test comprises of conceptual knowledge construct (Part-Whole, Proportion and Number subskills) and procedural knowledge construct (Simplification, Problem Solving and Computation subskills). Rasch dichotomous model approach was applied using Winsteps (version 3.71.0.1).

Results
Result shows the high internal reliability and demonstrated a goodness-of-fit of items except for 6 items. The sample shows high internal reliability of students. Items-students mapping shows on-targeted but with a small ceiling’s effect is detected.

Conclusions
This study is part of the ongoing investigation process to verify a valid test in classroom assessment for fractions’ teaching and learning. Students who did not perform well on fractions skills will lead to the difficulties to learn decimals, percentages and algebra.
Validation of the Internet Addiction Scale in Context of Higher Education: Applying Rasch Model

Presenter August 7th (13:25–13:50)
A.Y.M. Atiquil Islam
Institute of Education
International Islamic University Malaysia (IIUM)
PhD Student in Education
e-mail: skyiium@yahoo.com

Muhammad Mehedi Masud
Department of Economics
University of Malaya (UM)
PhD Candidate
e-mail: mehedi_rajapur@yahoo.com

In this age of exponential knowledge growth, where internet is playing a dominant role, the authorities of Higher Education concerned have to ensure that this tool remains within the reach of the students. However, despite a decade of existence, the internet was discovered to be addicted, especially by students. In so doing, the objective of this study is to determine the extent of students’ addiction in using internet and examine the validity and reliability of the Internet Addiction Scale. A total of 200 students from four faculties (Economics, Human Science, ICT and Engineering) were selected using quota sampling procedure. A questionnaire consisting of items validated from prior studies was put together and modified to suit the current study. A five-point Likert scale asking the respondents of the extent of their agreement/disagreement to the items constituting the construct in the questionnaire was used. The questionnaire’s validity and reliability were established through a Rasch model using Winsteps version 3.94. The results exhibited that (i) the items reliability was found to be at 0.94 (SD = 55.7), while the persons reliability was 0.88 (SD = 12.9); (ii) the items and persons separation were 4.02 and 2.76 respectively; (iii) all the items measured in the same direction (ptmea. corr. >0.36); (iv) all items showed good item fit and constructed a continuum of increasing intensity. The findings of this study foster support for the internal consistency reliability, unidimensionality, and measurement properties of the Internet Addiction Scale which is valid.

Keywords Internet addiction • Rasch model • Higher education
Development and Validation of the Corporate Citizenship Scale: Applying Rasch Model

Presenter August 7th (15:10–15:35)

Kamala Vainy Kanapathi Pillai
Centre of Graduate Studies
Open University Malaysia (OUM)
PhD Student in Business Administration
e-mail: kamalavarny@yahoo.com

A.Y.M. Atiquil Islam
Institute of Education
International Islamic University Malaysia (IIUM)
PhD Student in Education
e-mail: skyiium@yahoo.com

Corporate citizenship practices continue to accelerate as more global social and environmental issues arise; hence leading to increased pressure on authorities, leaders and corporations in Malaysia to re-evaluate their roles and impacts on society and environment. However, critical analyses to date indicate that there has been a lack of study on corporate conduct as well as firm-stakeholder collaboration within the Malaysian context, therefore, creating a gap for investigation. In so doing, the objective of this study is to develop and validate the corporate citizen scale to measure the corporate conduct of companies in Malaysia. The data for this study was collected through an online survey administered on senior managers involved in corporate citizenship efforts. A total of 31 managers from various industries participated in this study. The instruments’ reliability and validity were conducted by Rasch Model using Winsteps version 3.49. The results of Rasch Analysis demonstrate that (i) items and persons measured reliably \( r = .79 \), and \( r = .96 \), respectively), (ii) all valid items measured in the same direction (ptmea. corr. >0.30), (iii) most items depict good item fit and construct a continuum of increasing intensity. The findings reveal that this study’s significance stands on its contribution towards developing and validating the corporate citizenship scale that could be applied by future researchers in diverse educational and industrial settings.

**Keywords** Rasch model • Corporate citizenship • Corporate conduct • Stakeholder coll
The purpose of this study is to examine the degree to which academic self-handicapping behaviour exhibited in young adults are influenced by various degrees of student engagement. One’s behaviour is in accordance to one’s belief about oneself (Woods 1998), thus academic self-concept is critical in the academic growth of the student because it has a direct effect on the students’ performance, parents’ and community expectations, student’s future career as well as his/her lifestyle and successes – in this and in the after-life. Survey has been conducted to 1,032 students and the return was 832, whereby only 790 were correctly filled, thus the analysis is based on 790 respondents who are undergraduate students of International Islamic University Malaysia. Self-concept of each student was identified by utilizing self assessment tools with an aim of assessing the presence and/or the degree of students’ academic self-handicapping behavior as well as their stand, commitment and engagement to the university requirements and activities. Hence, this study investigates whether self-handicapping behavior and student engagement (emotional, cognitive, behavioral and religious engagements) represents five conceptually and empirically distinct psychological constructs when studied within the same domain; the nature of relationship existing between the four inter-related constructs of student engagement; the predicting quality of the four student engagement constructs on SHB construct; the fitness of the model of predictions of self-handicapping behaviour and the moderating quality of gender and nationality on the model of predictors of self-handicapping behaviour. Results proved the four factor model of predictors of SHB as empirically fit and reliable, while the SHB tool is being studied through Rash Model so as to formulate a partial disaggregation SEM model which would consist of one latent variable and four manifest variables.

**Keywords** Academic self-handicapping behavior • Student engagement • Counseling psychology
Reading ability in English as second or a foreign language is highly demanded as English has been extensively used in all fields of human knowledge. Thus, much research has been conducted to identify the nature of reading skill in L2. Two major views could be figured out: reading as a ‘unitary’ skill and reading as ‘multi-divisible’ skill. Despite this, the literature shows a lack of consensus in determining the number of the skills/sub skills that reading includes, and whether they are hierarchically ordered. Some studies have found that item difficulty is influenced by a number of factors other than person ability such as item test features i.e., the interaction between item characteristics and item difficulty. These features include, among others, text type, context type, test length, item format etc. As such, it has become problematic to ascertain the hierarchical assumption of reading skills. To resolve this, the multifaceted approach has been recommended to examine item difficulty taking into consideration factors which affect item difficulty level. This study employs the Many-faceted Rasch Model to ascertain the hierarchical assumption of reading skills for support of a developmental of reading ability. A 42-MCQ item test was administered to 944 ESL secondary students in Malaysia. The test items were identified according to the skill areas associated with them (interpreting information, making inference, understanding figurative language, drawing conclusions, scanning for details, and finding out word meanings); context type (linear and non-linear); and text type (ads, notices, a chart, a story extract, short messages, a poem, a brochure, and a formal letter). As connectivity is established for skill areas and context type only, the results for these two aspects are reported as they can be compared directly and replicated. The results show that item context types are not equally difficult; linear context types tended to be more difficult than non-linear context types. And the skill categories do not have the same difficulty level. The most difficult skill category was interpreting information and the easiest one was finding out word meanings. This supports that notion that different reading skills exert differential cognitive demands and those that require higher order thinking skills such as analysis are more difficult than those requiring lower order skills such as finding out word meaning and scanning for information. To conclude, the results of the FACETS analyses have provided the much needed evidence that there is a strong possibility of such reading hierarchy.

**Keywords** Reading hierarchy • Reading skills/sub skills • Multifaceted Rasch model
Applying Rasch Model to Assess Interactive Whiteboard Scale

Presenter
Essam Saleh Ahmed Bakadam
University Technology Malaysia (UTM)
PhD Student in Education
e-mail: essam-12@hotmail.com

A.Y.M. Atiquil Islam
Institute of Education
International Islamic University Malaysia (IIUM)
PhD Student in Education
e-mail: skyiium@yahoo.com

An interactive whiteboard (IWB) is a large, touch-sensitive board which is connected to a digital projector and a computer. The projector displays the image from the computer screen on the board. The computer can then be controlled by touching the board, either directly or with a special pen. The IWB is a useful interactive device that holds the students' attention, and accommodates different learning styles due to its interactive component, it also improves the way a teacher presents information, by utilizing it the Teacher will have less classroom management problems because interactive whiteboards keep students stimulated and interested. However, since its implementation at the Prince Sultan Intermediate School (PSIS) 2007, there has been no research conducted to analyze the use of the IWB by teachers, what teachers perceive to be the benefits and the problems encountered. In so doing, the aim of this study was to develop and validate the Interactive Whiteboard scale to assess the perception of the teachers on benefits in using IWB for their teaching and learning. The survey respondents consisted of 50 male teachers were selected using purposive sampling technique those who were utilizing the IWB in their instructions at the PSIS. A five-point Likert scale that indicated degrees of agreement/disagreement was used to capture the teachers’ views about the benefits of the board. The questionnaire’s reliability and validity were conducted through a Rasch model using Winsteps version 3.94. The results demonstrated that (i) items and persons measured reliably (r = .73, and r = .75, respectively); (ii) all items measured in the same direction (ptmea. corr. >0.49); (iii) most items showed good item fit and construct a continuum of increasing intensity. The findings also discovered that the variance explained by the measures was 47.5% which indicated that the items were able to endorse the PSIS teachers’ benefits in using IWB.

Keywords Interactive whiteboard • Rasch model • Perceived benefits
Motivation and Arabic Learning Achievement: A Comparative Study Between Two Types of Gansu Islamic Schools in China

Presenter August 7th (15:10–15:35)
Juping Qiao
e-mail: qiaojuping@gmail.com
Prof. Noor Lide Abu Kassim
e-mail: dr.noorlide@gmail.com
Dr. Kamal Badrasawi
International Islamic University Malaysia, Malaysia
e-mail: kamalbadrasawi@gmail.com

Background
Motivation is one of the most highly studied issues within the field of L2 learning. A number of theories of motivation have been used to explain the effects of motivation on L2 learning. The Self-determination theory (SDT), a popular motivational theory developed by Ryan and Deci (1985), has been applied by many L2 researchers to explicate the relationship between motivation and L2 learning outcomes from social and psychological perspectives. In addition, SDT distinguishes intrinsic motivation from extrinsic motivation on L2 learning achievement. Bakar et al. (2010) in their study extended Ryan and Deci’s self-determination theory to investigate the role of religious motivation in Arabic (L2) language learning achievement. This study sought to extend previous findings by examining the relationship between motivation aspects: Religious Motivation (RM), Internal Motivation (IM), External Motivation (EM), and Amotivation (AM) and Arabic language learning achievement.

Aims and Keywords
Specifically, it aims to (a) explore factors that influence Arabic learning achievement and examine the relative contribution of the different aspects of motivation (RM, IM, EM and AM) on Arabic language learning achievement; (b) investigate the relationship between those aspects; and (c) compare the level of motivation across two subgroups of Arabic language learners.

Keywords Self-determination theory • Arabic learning achievement • Religious motivation • Rasch analysis

Methods
To achieve these purposes, 348 students were randomly selected from two types of Gansu Islamic schools in China to complete a 36-item questionnaire.
Interviews were also conducted with four respondents to get more in-depth information as regards Arabic language learning. Qualitative analysis was conducted to answer the first research question, while the Rasch Measurement Model, independent sample t-test, Person correlations, and multiple regression analysis were utilized to answer the other research questions.

Results
The qualitative analyses indicated that there are some positive and negative factors (internal and external) that affect student learning of Arabic language either positively or negatively. Among the positive factors are religion, positive attitudes towards learning Arabic, and finding a more prestigious job. On the other hand, the negative factors include, lack of motivation to learn Arabic and the teaching methodology. The quantitative data showed significant correlations among all the motivation aspects except Amotivation. The multiple regression analyses indicated that AM and RM were significant predictors of Arabic language learning.

Conclusions
All the motivation aspects influence students’ Arabic learning achievement either positively or negatively. Religion and a positive attitude towards Arabic language learning motivate students to do better in learning the language, while Amotivation and inappropriate teaching methodology deter students from learning Arabic effectively. Direct intervention and new learning and teaching strategies should be formulated to promote effective learning of Arabic language.
Can Mindfulness Be Measured Using Self-Report Questionnaires? A Critical Examination of the Five Factor Mindfulness Questionnaire by Rasch Model

Presenter August 6th (14:20–14:45)
Ying Zhang
e-mail: Ying.Zhang@umcg.nl

Adelita V. Ranchor
Robbert Sanderman
Joke Fleer, and Maya J. Schroevers
Graduate School for Health Research
University of Groningen, the Netherlands

Background
The current study examined construct validity of a psychological scale. The authors are psychologists.

Aims
Several self-report questionnaires exist to measure mindfulness, with the 39-item Five Factor Mindfulness Questionnaire (FFMQ) being one of the most frequently used scales. The aim of the current study was to examine whether the scale items are understandable and whether the total score can be used to identify people with a different level of mindfulness.

Sample
The study was carried out in 152 undergraduate students.

Methods
Students were randomized in two conditions: 71 students filled out the original scale with five response categories and 81 students were given an extra response option (6 “never paid attention”). When students chose response option 6 at least once, they were asked to give reasons for choosing this option, and then to rate the items again, using the original 5-point rating scale. Rasch model was used to examine items’ ability to discriminate different levels of mindfulness.

Results
Option 6 was indicated by more than 5% of the students in 7 of the total 39 items. Rasch analysis showed items in four of the five subscales were less discriminative: the responses were either less accurate than expected or falling out of the predictable range.

Conclusions
The construct validity of the FFMQ is in doubt. Half of the items are either not understandable or not able to identify people with a different level of mindfulness.
Rasch Calibration of General Science Test at Grade-VIII in Pakistan

Presenter August 7th (16:25–16:50)
Muhammad Javid Qadir*, Abdul Hameed
School of Social Sciences and Humanities,
University of Management and Technology
Lahore, Pakistan
e-mail: javid_qadir123@yahoo.com
   abdul.hameed@umt.edu.pk

Iram Gul Gilani,
Department of Education
Bahauddin Zakarya University Multan
Pakistan
e-mail: iramgul@bzu.edu.pk

The present study aimed at calibration of General Science achievement test for grade (VIII) through Rasch Model. For this purpose a General Science achievement test comprising 45 items was constructed from the text book of General Science for class VIII. Finally the test was administered to 300 students (M/F) in different high schools for boys and girls in Multan District. The answer sheets were scored and results were tabulated. Eleven (11) items were rejected on the basis of F, D and $\phi$. Fifteen (15) items were to be improved on the basis of F, D and $\phi$. Remaining all items were good items. Reliability value of the test was (0.82) and (0.85) by using Kuder Richardson # 20 and Kuder Richardson # 21 formula, which is very close to standardized value. Rasch Model indicates that overall test is good to measure the achievement of the students class (VIII) in the subject of General Science.

Keywords Calibration • General science • Achievement test • Rasch • Item analysis
Validation Study of Cut-Scores in School Achievements’ Monitoring (SAM) Toolkit

Presenter
Наталья Гапонова
Center of International Cooperation in Education Measurement
e-mail: g.natalia999@gmail.com

Data collection during mass testing always leads to presentation of performance results. Establishment of cut-scores and division of examinees into groups due to their test performance is one of the most common ways to solve the problem of interpreting the results. However standard-setting procedure and its outcome are one of the most controversial which is difficult to validate empirically. Within each project that implements education assessment there is a professional and ethical responsibility for the results presented. That is why it is crucial to investigate the problem of how to provide examinees with valid and reasonable interpretation of test results. The first step in validation of benchmarks is development of standard-setting methodology for SAM, which results in establishment of cut-scores. The next step is to investigate internal, procedural and external validity of benchmarks. We investigated consistency between theoretical judgments and empirical estimates. Judges needed to set hypothetical p-values for each item as if examinees answered. Then we investigated the correlation between theoretically expected p-values and empirical within three different levels of competency. Secondly, validity evidence was investigated with the help of comparisons with other sources of evidence, such as consistency with results from, widely used method of standard-setting, Angoff. Thirdly, in order to prove procedural validity every step of standard-setting procedure is articulated clearly, including the purposes and processes. Moreover, the extent to which repeated applications of benchmarks through times get the same distribution of examinees on different levels was searched to prove internal validity. To conclude, we expect to prove the validity of the SAM’s benchmarks, so that examinees could be sure that the interpretation of test results is rather valid, reasonable and precise.
Exploring the Features of Adaptive Neuro-Fuzzy Inference System and Path Analysis in Determining Factors Influencing Item Difficulty: A Study of While-Listening Performance Tests

Presenter August 7th (14:15–14:40)
Vahid Aryadoust
Centre for English Language Communication,
National University of Singapore,
10 Architecture Drive, Singapore 117511

Language researchers have adapted a variety of statistical tools to explore the variables that influence task difficulty in listening comprehension. Three notable examples are regression models, rule space methodology, and the fusion model. Although these studies have informed the field of language assessment, the focus of the majority is limited to post-listening performance (PLP) tests and their methodologies have limitations. PLP tests demand answering test items after the auditory experience. By contrast, a group of tests are while-listening performance (WLP) in which test takers are exposed to oral texts and engaged in simultaneous reading and answering.

This study reports a novel application of a class of neuro-fuzzy models (NFMs) called Adaptive Neuro-Fuzzy Inference System (ANFIS) into the International English Language Testing System (IELTS) listening test, a WLP test, and compares the findings with path analysis. NFMs are powerful artificial intelligence technologies which embrace the tenets of Artificial Neural Networks (ANN) and fuzzy set theory. The synergy of ANNs and fuzzy sets in NFMs provides promise for language assessment. In this study, seven variables influencing task difficulty were flagged during the item coding stage, a finding which was further supported by the ANFIS and partly by the path analysis. Results show that the NFM technique is promising in the context of language assessment. Further applications of the model in language assessment are discussed.
Synthesizing Language Measurement Advances

Presenter August 6th (16:45–17:10)
Vahid Aryadoust
Centre for English Language Communication
National University of Singapore
10 Architecture Drive, Singapore 117511
65–6601 2504(DID):: 65–6777 9152 (Fax):: elcsva@nus.edu.sg (E):: www.nus.edu.sg (W):: Company Registration No: 200604346E

This paper seeks to review the contribution of three major measurement trends to language assessment: classical testing theory, items response theory, and structural equation modelling. It is argued that classical testing theory and latent trait approach have propelled forward language (and educational) assessment since the 1970s. The paper further discusses the advantages of the latent trait models, specifically the Rasch model, differential item functioning, and cognitive diagnostic assessment models such as the fusion model and rule space methodology. Subsequently, it elaborates the construct modeling approach (Wilson, Constructing measures: An item response modeling approach. Mahwah: Erlbaum, 2005), as a useful method to develop language tests. The author discusses how Rasch-based construct modeling approach would assist language testers to develop second language listening comprehension tests. It finishes off by offering recommendations for future research in the field of language assessment and stressing the need to adapt novel measurement methods into language testing.

Evaluating the Quality of Ratings in Writing Assessment: Indices from Rasch Measurement Theory

Presenter August 6th (15:30–15:55)
Susan Tan and Chew Moh Leen
National University of Singapore
Stefanie A. Wind and George Engelhard, Jr.
Emory University

Background
The quality of ratings assigned in performance assessments must be systematically examined to ensure that ratings are valid, reliable, and fair for all students. Research on rating quality in performance assessment typically evaluates raters in terms of agreement, error, and systematic bias as indirect indicators of accuracy (Murphy and Cleveland 1991; Johnson, Penny, and Gordon 2009). Rasch Measurement Theory (RMT) provides a variety of indices for monitoring the quality of ratings in writing assessment, including indices of rater error and direct measures of rater accuracy (Engelhard 1994, 1996; Wind and Engelhard 2011).

Aim
The goal of this study is to examine the quality of ratings over two administrations of a placement test and to determine how such ratings affect student placement in a large-scale writing assessment.

Sample
This study uses essay scores from two English Placement Tests in Singapore. Fifty scripts were used from the 2010 test, scored by 37 operational raters and 30 scripts were used from the 2011 test scored by 32 operational raters. In addition, for each data set there is a team of three benchmark raters.

Methods
Differences in rating quality were considered in terms of rater experience, types of course taught and behavior over the two administrations. Many-Facet Rasch Models were used to examine raters in terms of error and accuracy. Indices of rater accuracy include rater severity calibrations and model-data fit statistics, direct measures of a match between operational and “benchmark” ratings. All analyses were performed using Facets (Linacre 2010).

Results
Preliminary analyses show differences in rater severity, model-data fit, and direct measures of accuracy which suggest that rater characteristics may influence the quality of student scores.
Conclusions

Differences in rating quality hold implications for research, theory, and practice. Inconsistent rating quality implies that the quality of inferences informed by rater-assigned scores may not be invariant across the raters.
Effects of Three Dimensional Visualizations on Learning Molecular Biology

Presenter

Sandra TAN,
Hwa Chong Institution, Singapore

Russell WAUGH
Graduate School of Education
University of Western Australia
e-mail: sandra@hci.edu.sg

This paper describes the design considerations, implementation and results of formulating an approach to help students “see” DNA, proteins and cellular structures. An initial sample size of 249 Secondary three biology students was drawn from seven schools in Singapore. Purposive sampling was done in the selection of schools with different academic ability levels and yet fairly representative of biology students in Singapore. The experimental study leverages on novel immersive virtual reality technologies to help students understand the three-dimensional structures and the molecular interactions between them that enable function. Participants were taught topics in molecular biology either by traditional classroom “chalk-and-talk” or by a series of three-dimensional visualisation exercises using ICT. The students were tested for visual spatial ability and molecular biology achievement both before and after the intervention and the data analysed using a Rasch measurement model. Results indicate increases in visual spatial ability and molecular biology achievement. These effects were particularly pronounced in male students. Focus group interviews reveal that, prior to this intervention, students relied heavily on memorisation. This observation corroborates well with the analysis that revealed that students were well-trained in memorising specific biology phenomenon, but could not link these to general concepts and theory. The visualisation exercises helped clarify understanding while increasing interest and engagement. The results of this study recommend the use of technology in the teaching and learning of concepts in molecular biology, especially for male students in Singapore.
Rating Scale Model; Attitudes Toward Lesbians and Gay Men Scale; Dimensionality; Acquiescent Response Bias

Presenter
Prof. Dr. Gideon P. de Bruin
University of Johannesburg
Box 524, Auckland Park, Johannesburg, South Africa
e-mail: deondb@uj.ac.za

Dr. Marlene Arndt
University of Johannesburg
Box 524, Auckland Park, Johannesburg, South Africa

Prof./Dr. Trevor G. Bond
James Cook University

Aims and Keywords
We examined measurement disturbances in the Attitudes toward Lesbians and Gay Men Scale (ATLGS).

Sample
Participants were 1,192 students (516 men and 671 women) who represented six faculties at a university in Johannesburg, South Africa. Participants’ ages ranged from 17 to 42 years. Students represented four ethnic groups: Black, White, Mixed, and Asian.

Methods
Responses to the 20-item ATLG were fitted to the Rasch rating scale model. There are two subscales (Attitudes towards Lesbians and Attitudes towards Gay men). Items have a 9-point response scale.

Results
Four potentially problematic disturbances were detected: (a) a response scale for the items that calls for too fine a distinction in different attitudes, (b) a small but non-negligible group of persons who respond inconsistently across the Attitudes Toward Lesbians subscale and Attitudes Toward Gay Men subscale, respectively, (c) a small but non-negligible group of persons who respond inconsistently across favourable and unfavourable items, respectively, and (d) persons and items that do not fit the requirements of the Rasch model.
Conclusions
Better measurement with the ATLG can be achieved by using a response scale with fewer response categories. The use of favourable and unfavourable items introduces unwanted multidimensionality (possibly due to acquiescent response bias) and poor fit. Finally, the results show that for most persons the ATLG measures a unidimensional attribute. However, for a small group of persons the total score is an invalid indicator of their attitudes. Rasch analysis can be used to identify these persons and other persons who respond in unexpected ways.
Intraclass Coefficient to Report the Strength of Dimension Using Rasch Principal Component Analysis on Standardized Residuals

Presenter Tsair-Wei CHIEN
Chi Mei Medical Center, Taiwan
e-mail: smile@mail.chimei.org.tw

August 6th (15:30–15:55)

Background
Dimensionality is an important assumption in item response theory (IRT). Principal component analysis (PCA) on standardized residuals has been used to check dimensionality, especially under the family of Rasch models. Smith (2002) proposed the percent of two sets of person ability with significant difference beyond 5% that is deemed any potential multidimensionality occurred and can be supplemental to the traditional criteria such as a cutoff of 60% of the variance explained by the Rasch factor, eigenvalues smaller than 3 and the percentage variance explained by the first contrast of less than 5%. However, the message regarding dimension strength is required to report in a study.

Aims and Keywords
Following Smith (2002) rule on two sets of person ability, we proposed another way using intraclass coefficient (ICC) to verify dimension strength of a scale.

Keywords Intraclass coefficient • Rasch model • Unidimensionality • PCA • Cronbach’s alpha

Sample
The Rasch rating scale model was used to analyze the 2009 English inpatient questionnaire data regarding patient satisfactory perception, which were collected from 162 hospitals, examined unidimensionality.

Methods
We developed a visual plot in Excel incorporated with WINSTEPS to automatically compute ICC, Smith’s percent of two sets of person insignificant ability and its correlation coefficient that were compared with the results yielded by the responding simulation data.

Results
ICC could be as an indicator to measure the strength toward unidimensionality for a scale, through which 0.80 can be a cut off to discriminate whether a scale is unidimensional.
Conclusions
It is required to report the dimension strength of a scale when one factor is extracted from data that is like Cronbach’s alpha reported as usual. In this study, we demonstrated dimensional validity for a scale with ICCs that might respond to the question about the standard format for reporting Rasch analysis in publications (Linacre 2010).
A Simple Screening Tool for Dengue Fever in Children

Presenter August 6th (13:30–13:55)
Wen-Pin Lai, Tsair-Wei Chien
Chi Mei Medical Center, Taiwan
e-mail: smile@mail.chimei.org.tw

Background
Dengue fever (DF) is a significant public health issue in Asia. Some studies using univariate approach report that the presumptive diagnosis of DF is so imprecise because the signs and symptoms are not useful for detecting DF. Other multivariable regression analysis was attempted to distinguish patients with dengue from those with other fibrile illness; however, none had significant statistical validity and none considered changes in clinical features over the course of illness.

Aims and Keywords
We aimed to utilize clinical and laboratory data to derive a rapid and accurate DF case-finding tool for children.

Keywords Dengue fever • Dengue serologic test • Rasch analysis • Receiver operating characteristic (ROC) curve

Sample
This retrospective study used 24 DF-related characteristics and clinical features (17 clinical; 7 laboratory) from 177 pediatric patients (69 with DF diagnosis). Data were psychometrically evaluated and their effectiveness and accuracy in predicting DF risk were also evaluated.

Methods
Guided by the DF literature, a total of 24 DF-related clinical feature (7 laboratory and 17 clinical) capturing clinical, historical, and laboratory indicators were selected from patients clinically suspected of DV infection at the emergency department for the construction of a scale to screen for DV infection. According literature regarding symptoms of DF, data were obtained from the patients’ medical records. Rasch principal component analysis (PCA) on standardized residuals were analyzed with WINSTEPS to search an acceptable combination of items to distinguish DF occurred.

Results
The 14-item scale (DF-14) was found to fit the measurement model in assessing DF tendency. The sensitivity (specificity) of the DF-14 measure was 0.759 (0.855) with a cut-off point of −1.15 (in logit), and the area under the ROC curve was
0.932 (95% CI: 0.885–0.965). The DF-2 scale, comprised of white blood cell and platelet counts data, was brief but less comprehensive.

**Conclusions**

Simple laboratory data, such as those in the DF-2 and DF-14, can be useful for the early detection of DF risk in children. The DF-14 scale can be helpful in discriminating DF from other febrile illnesses before conducting a costly and time-consuming dengue confirmation test.
Rasch Analysis of the Relationship Between Physical Therapy and Occupational Therapy Manpower in Taiwan

Presenter August 6th (15:55–16:20)
Hing-Man Wu, Tsair-Wei Chien
Chi Mei Medical Center, Taiwan
e-mail: smile@mail.chimei.org.tw

Background
Manpower supply is an important issue in healthcare service industry. Whether the physical therapists in Taiwan are sufficient is required to study.

Aims and Keywords
The aim of this study is to investigate (1) the difference between physical and occupational therapy manpower in various medical settings and different insurance bureaus of geographical location, (2) the correlation coefficient between physical and occupational therapy manpower, (3) the performance for physical and occupational therapy manpower using a graphical representation in order to fulfill the requirement of the new version of 2011 hospital accreditation.

Keywords Physical therapy • Occupational therapy • Correlation coefficient • Graphical representations • structure equation modeling.

Sample
All the physical and occupational therapists registered in healthcare industry were recruited in this study.

Methods
Web-query system for medical practitioners in Taiwan is retrieved to obtain the numbers of physical therapist, occupational therapist, and hospital bed. Using the ratio of manpower supply to hospital beds, we examined whether it is adequate to meet the patient needs in hospital. The maximum of hospital capacity reaches to 2,918 beds and the minimum is 10 beds. The Rasch analysis with WINSTEPS and structure equation modeling are used to inspect the association between physical and occupational therapists in manpower supply.

Results
The results showed that the ratio of physical and occupational therapists in manpower supply was 2.54. The largest number of physical and occupational therapists serving for medical settings was in clinics (34.49 %), the least was in medical centers (12.64 %). A significant association was found in types of institutes related to rehabilitation professionals (Chi-square = 892.06; p < .0001), indicating interdependent with each other.
**Conclusions**
The data of medical practitioners and hospital beds retrieved from the website of Department of Health could be useful to compare the manpower of rehabilitation professionals using graphical representation. In addition, the method we introduced could be helpful for researchers to consider and explore in future studies.
Construction and Evaluation of an Item Bank for an Introductory Statistics Class: A Pilot Study

Presenter Sieh-Hwa Lin
National Taiwan Normal University
e-mail: linsh@ntnu.edu.tw

Pei-Jung Hsieh, (corresponding author)
National Academy for Educational Research
e-mail: peijung2@gmail.com

Background
Introductory Statistics is a required course for psychology and education related fields’ students. However, many students are uneasy about the learning materials.

Aims
The purpose of the present study is to construct and evaluate the applicability of an item bank for an Introductory Statistics class.

Sample
The participants of the study were 54 college students enrolled in Statistics in Psychology and Education in the spring and fall semesters of 2010.

Methods
To establish the test bank, the authors first adopted and re-wrote the questions from the midterm and final tests of Statistics in Psychology and Education from the previous school year (2009). Students could practice the test in advance of the class on the instructional platform, Moodle. A total of 15 units were prepared with 45 questions. The collected data were analyzed by WINSTEPS 3.70.

Results
The results revealed that (1) the point-biserial correlation of 34 items (75.6 %) reached .25, meaning the test items constructed for this study had enough discriminatory power; (2) 80 % of the Rasch item difficulty values ranging from −1 to 1, indicating an appropriate difficulty of the test bank, which not only met the students level, but was also appropriate to help students with preparation for the unit.

Conclusions
Students’ performance on the course score correlated positively with the number of times they took the preparation tests, suggesting that the implementation of the lesson preparation activity enhanced the learning effectiveness of statistics. In the future, the test bank will serve as the supplementary material for Introductory Statistics in psychology- and education-related fields.
The Aesthetic Intention Scale: Using the Health Belief Model to Predict Aesthetic Intention and Action Taking Among Hospital Employees

Presenter
Shih-Bin Su, Tsair-Wei Chien
Chi Mei Medical Center, Taiwan
e-mail: smile@mail.chimei.org.tw

August 7th (13:50–14:15)

Background
An aesthetic intention scale should be examined and applied to know the mostly possible potential customer would like to receive aesthetic treatment for a marketing plan.

Aims and Keywords
To develop an aesthetic intention scale which can predict aesthetic intention and action among hospital employees is incorporated with a report card disclosing individual examinees’ aberrant responses on items.

Keywords  Perceived susceptibility subscale • Improvement essentiality subscale • Rasch model • Item response theory • Aesthetic intention scale

Sample
The setting was a 900-bed hospital in southern Taiwan. A total of 1,800 full-time workers in the studied hospital participated in an aesthetic intention survey in May of 2009. The effective sample size was 1,124 with a return rate of 62.64 %.

Methods
Item response theory based Rasch model with WINSTEPS software was performed to examine an aesthetic intention scale whether forms a single construct. Multiply regress analyses were used to seek factors in association with aesthetic intention. Receiver Operating Characteristic (ROC) was applied to determine a cut-off point to looking for examinee’s aesthetic action taking.

Results
The 12-item aesthetic intention scale has been detected to have two characteristic factors (i.e., subscales of perceived susceptibility and improvement essentiality). The cut-off point is set at 2.975 after combining those two scaling scores.
It accounts for 87.1% of accuracy with sensitivity. The report card for investigating examinee’s aberrant responses on items using standardized residual scores can provide institutes of aesthetic service with a diagnostic tool to inspect any response answered unusually in comparison to model’s expectation.

**Conclusions**
The aesthetic intention scale examined by Rasch model can be used to predict aesthetic intention and action taking among hospital employees.
The Measurement of Creative Teaching Competency:
Development of an IRT-Based Scale

Presenter August 8th (9:25–9:50)
Yu-Shu Chen
An associate professor at National Chung Cheng University in Taiwan ROC
e-mail: xiaoshu700@gmail.com

Yuan-Chi Lai
An assistant professor at Wu Feng University in Taiwan ROC
e-mail: yclai@wfu.edu.tw

Background and Purpose
Traditionally, Asian countries emphasize rote learning and passing examinations. The lack of creative teaching performance has had an adverse impact on human resource education. Teachers are required to produce competent graduates, ensure that the necessary technical knowledge is acquired so that graduates can effectively contribute to the workforce. How to improve human resource education hinges on teachers’ creative teaching competency. However, there is no instrument to measure teachers’ creative teaching competency. The major purpose of this study is to develop a reliable and valid scale to assess teachers’ creative teaching competency.

Methods
The study was administered in two stages. In the first stage, based on Spencer and Spencer’s (1993) competency developing model, the researchers developed a creative teaching competency scale which including seven dimensions: applying team learning, creating safe environment, deferring judgment, tolerance for ambiguity, accepting frustration, playfulness and humor, and risk-taking. The scale consists of 33 indexes. Then, the researchers converted the indexes into Likert-type items. For each of the 33 items, a response was sought in two perspective—importance and possession. In the second stage, we recruited 400 teachers from elementary schools and junior high schools. Two hundred of them are normal teachers and the other 200 are creative teaching award winners. Data were analyzed with a between-item multidimensional model because the scale contains seven dimensions and each measuring related but distinct latent dimensions. In this study, each item belongs to only one particular dimension, and there are no items in common across the dimension.

Results and Conclusions
The results yielded good psychometric property for the scale and revealed that the scale is appropriate for assessing teachers’ creative teaching competency. The
seven-dimensional creative teaching competency scale has good reliability and discrimination. The scale showed substantial power for the explanation of variable in creative teaching performance. Educational implications of the current findings and suggestions for future studies are also addressed.

**Keywords** Creative teaching competency • Item response theory • Multidimensional between-item model • Rash model
Exploring and Modeling the Multidimensionality of Metalinguistic Knowledge

Presenter: Wen-Ta Tseng, PhD
Associate Professor, English Department
National Taiwan Normal University
e-mail: wttseng@ntnu.edu.tw

Hsiao-Ling HSU, Doctoral Student, English Department
National Taiwan Normal University

Background
With many studies focusing on distinguishing between implicit and explicit knowledge, there is little research using the Rasch Modeling and the structural equation modeling to examine the explicit knowledge in particular.

Aims and Keywords
This study was aimed at examining the trait space of and confirming the structural relationships underlying the metalinguistic knowledge.

Keywords: Metalinguistic knowledge • Rasch Modeling • CFA • MIMIC

Sample
Two hundred Taiwanese university students were recruited as the participants of the present study.

Methods
To measure metalinguistic knowledge, two types of test were adopted—untimed GJT (grammatical judgment test) and MKT (metacognitive knowledge test). Untimed GJT items were adapted from Loewen (2009) and MKT items were adopted from Lin (2009). Both GJT and MKT contain 20 items respectively, and each test was further sub-divided into three groups based on their feature of grammaticality— that of vocabulary, basic English grammar, and functional words. The trait space of metalinguistic knowledge was checked by the Rasch Modeling, and the structural relationships underlying it was examined by the Confirmatory Factor Analysis and MIMIC model analysis. Winsteps 3.72 was utilized to execute the Rasch Modeling, and AMOS 18 was then further implemented to perform a series of CFA model comparisons and MIMIC analysis.

Results
The results from the Rasch Modeling showed that the latent trait under investigation could be considered a bi-factor model. Based on this initial but fundamental finding,
the CFA analysis with this confirmed bi-factor model could also be well-supported by a number of model fit indices. More importantly, the results of the MIMIC model revealed that from a developmental perspective metacognitive knowledge has moderate predictive power over the functioning of learners’ grammaticality judgment in English.

**Conclusion**
Both the theoretical and pedagogical implications were duly proposed in response to the research findings.
Is Analysis Under the 3PL Model of Practical Value to Test Makers?

Presenter
Jeffrey Stewart
Kyushu Sangyo University, Cardiff University
e-mail: jeffjrstewart@gmail.com

Background
Although the Rasch model assumes minimal guessing, it is commonly applied to multiple-choice test formats, for which some argue the 3-parameter logistic (3PL) model is more appropriate. Theoretical debate aside, it is necessary to empirically test if analysis under the 3PL model can result in practical improvements of test forms.

Aims
To answer the following question: If a test which exhibits high person separation and an ideal TIF under the Rasch model is redesigned in accordance to implications of the 3PL model, does the revised form demonstrate higher reliability under the criteria of Rasch measurement?

Sample
Two samples each of approximately 1,500 first and second year English conversation students at a Japanese university.

Methods
Using a bank of 300 items, a 100-item test was designed which, under subsequent analysis in WINSTEPS, produced an ideal TIF under the Rasch model and had a person separation of 3.18. However, in a model comparison in the R package ltm, the 3PL model attained superior fit, and analysis under it indicated the test was somewhat difficult for the target population. A new 100-item test was designed to produce an optimal TIF under the 3PL model, and it was also examined under WINSTEPS in order to compare its Rasch reliability to the previous form. Although the TIF under the Rasch model was somewhat skewed, Person separation increased to 3.65, higher than under the Rasch-optimized test form.

Conclusion
Even if the goal of a test maker is to produce a test calculated under raw score, if multiple-choice formats are used, analysis under the 3PL model can result in tests of substantially higher person reliability, even under the criteria of Rasch measurement.
Measuring Cumulative Learning of Energy Topics

Presenter
Ou Lydia Liu
660 Rosedale Rd, Mailstop R7
Princeton, NJ, 08541, United States
e-mail: lliu@ets.org

Background
This research is about measuring cumulative understanding of energy topics among middle school students. Author is Dr. Ou Lydia Liu, Research Scientist at the Educational Testing Service at Princeton, New Jersey, United States.

Aims and Keywords
The aims of the research are to develop curricula and assessment to help advance the learning and assessment of energy topics among middle school students.

We ask three research questions: (a) How valid, equitable, and reliable are the knowledge integration energy items developed by the Cumulative Learning using Embedded Assessment Results project? (b) Do the items provide evidence for cumulative learning, cross-sectionally and longitudinally? And (c) What is the impact of unit learning on cumulative learning.

Keywords  Science assessment • Energy topics • Rasch partial credit model

Sample
Participants are 4,160 middle school students from four schools in California in the United States. The sample consisted of 1,186 6th graders, 1,807 7th graders, and 1,167 8th graders. The sample included 2,108 males and 2,181 females. Among the participants, 3,087 speak English as their first language and 1,202 speak English as a second language.

Methods
Five energy curriculum units were designed and taught to the students. After learning the units, students took the end of the year assessment in both 2010 and 2011. The psychometric quality of the assessment items was evaluated using both classical testing theory and the Rasch partial credit model. In addition, the performance was compared among students who didn’t receive the instruction, and students with different exposure to the instruction.

Results
The assessment items showed satisfactory psychometric quality. Students who had more exposure to the energy instruction made bigger progress on the assessment than students who had less exposure to the instruction.
Conclusions
The assessment designed to measure cumulative learning of energy topics demonstrated satisfactory reliability and validity. The energy curricula were effective in terms of cultivating students’ cumulative understanding of energy.
Construct Mapping, Theory Building, and Validity Testing: An Examination of the Student Moral Character Scale

Presenter
Jade Caines, Ph.D.
University of Pennsylvania
Graduate School of Education
3700 Walnut Street, Room 335
Philadelphia, PA 19104–6216
e-mail: jcaines@gse.upenn.edu

Background
I was an educator for over 10 years, teaching Kindergarten through university students. In 2011 I received my Ph.D. in Educational Research, Measurement and Evaluation. Currently, I am a postdoctoral fellow at the University of Pennsylvania.

Aims and Keywords
In recent years, there has been a greater focus on the development of student character and how it influences performance in the classroom (Davidson, Khmelkov, and Lickona 2010). Limited research, however, exists on the validity of instruments used to measure character in students. To address this gap in the literature, this study uses the Rasch Rating Scale model to investigate the convergence of item fit and theoretical expectations of a scale that measures Student Moral Character (SMC).

Sample
Data were collected on 239 middle-school students. The student demographic data are as follows: 92 males (39 %), 134 females (56 %), and 13 Unknown (5 %); 180 Black students (75 %), 40 Hispanic/Latino students (17 %), 1 Native American student, 1 White student, and 17 students “Unknown” (7 %); 163 sixth graders (68 %), 56 seventh graders (23 %), 9 eighth graders (4 %), and 11 students “Missing” (5 %). Also data were collected from four different public schools that range in size and geographical location.

Method
The Rasch rating scale model was used. WINSTEPS was used to produce item level statistics and Wright maps.

Results
Results suggest that the SMC scale does not follow many of the theoretical expectations. Although all of the SMC items show mean square error statistics (OUTFIT values) that fall within the acceptable, restrictive range, some item disordering exists.
Conclusions
First, the Rasch model provides a useful framework for examining item fit statistics and theoretical expectations. Second, scale revisions should be made before inferences from this scale can be considered valid. Finally, implications include theoretical advancement for the educational measurement and character education fields.
Metaphor as a Basis for Unifying the Conception of Measurement Across Physics and Psychometrics

Presenter
William P. Fisher, Jr.
University of California, Berkeley
e-mail: wfisher@berkeley.edu

A. Jackson Stenner
MetaMetrics, Inc.

Background
New things lacking concepts come into words and take their places in a language’s network of significations via a process through which unclear images or figures not yet signified are identified and located relative to everything else. In physics as in psychology, that process is metaphor.

Aims and Keywords
Does metaphor calibrate a virtual measure of meaning? A study of the metaphor “love is a rose” explored this question.

Sample
Thirty-six residents (18 men, 18 women) of two locations in the US state of Illinois rated 68 entailments of the metaphor (“love is a plant,” “love is beautiful,” etc.) on a six-point agree-disagree rating scale.

Method
A theory of the construct implied by historical study of the metaphor’s entailments was devised. The explanatory capacity of this theory was experimentally tested. Items representing the major theoretical aspects of the construct were separated onto three survey forms. Items tapping the same theoretical aspect of the construct should calibrate to the same locations, within a 95% confidence interval. Data were fit to a polytomous Rasch model, and mean item locations were compared using ANOVA.

Results
Model fit was satisfactory, person measure separation reliability was 0.89, and item calibration reliability was 0.92. The three groups of items calibrated in the predicted order, with the low and medium, and medium and high, group means both differing by about a logit (respectively, t-statistics were 3.4 and 5.2, with 38 and 36 d.f. and p ≤ .002). Average measures by sex and location were statistically identical, as were the average calibrations by form.
Conclusions
Measurement is a complex array of interdependent metaphors involving an invariant, portable unit defined via a mathematical law, predicted from theory to a useful degree of precision, and efficiently deployed in practical applications in a universally uniform metric traceable to a reference standard.
Contrasting Physics and Psychometrics: Units, Laws, Theory, and Metrology

Presenter
William P. Fisher, Jr.
University of California, Berkeley
e-mail: wfisher@berkeley.edu

A. Jackson Stenner
MetaMetrics, Inc.

Background
The ongoing cultural and economic successes of the physical sciences stand as a model to be emulated by all fields. Similarly rigorous sciences of psychological and social constructs require close attention to qualitative issues of unit definitions, lawful data regularities, predictive theories, and metrology’s social networks.

Aims
Illustrated contrasts of the different ways psychometrics and physics approach these issues may be instructive.

Results
For instance, the identification and maintenance of a unit has long been of central importance in physics, but has become of interest in psychometrics only in recent decades. A second difference involves the definition of measures in terms of nonarbitrary lawful regularities; qualitative relationships between magnitudes of different attributes, such as mass, force, and acceleration, are embodied in physical measures but usually not in psychometric ones. A third difference involves theoretical understanding of the constructs measured, where explanatory power informs a capacity to accurately predict the causal trade-offs resulting from interventions on one attribute while holding another attribute constant. In psychometrics, predictive theories facilitating control over constructs have been obtained, and are of increasing interest, but are not commonly pursued. A fourth difference involves the convening of international standards groups responsible for maintaining traceability to theory-informed constant unit values essential to scientific and commercial activities globally. No such groups or traceability mechanisms exist in psychometrics, where most instruments measuring a given construct do so in their own unique units, and there is little awareness that traceability of this kind might be viable and valuable in human, social, and economic terms.

Significance
Issues concerning units, laws, theories, and metrological networks may represent opportunities for advancing the state of the art in psychometrics.
Measuring Nurse’s Attitude Toward Communication Between Nurses and Physicians Using Construct Modeling and Rasch Analysis

Presenter: Mary P. Bourke RN, MSN, PhD
Indiana University School of Nursing Kokomo
Office Location KE 312
e-mail: mbourke@iuk.edu

Aims and Keywords
Understand how Construct Modeling is used to identify dimensions of a construct under study—a guide for instrument development. Synthesize how Rasch Model diagnostics are used to evaluate the effectiveness of instruments developed for measurement.

Keywords
Construct modeling • Rasch model

Sample
The population used for this study was a convenience sample of 89 RN’s currently working at a regional hospital.

Method
A partnership was formed between a large regional hospital and Indiana University for the purpose of research. The purpose of the research was to ultimately improve patient outcomes by understanding the dimensions of physician/nurse communication as perceived by the nurse. The research team developed an instrument based on Construct Modeling, thus identifying dimensions of communication within the context of medical care. The identified dimensions were used as a guide in the creation of items for the instrument. Rasch diagnostics were performed using Winsteps software. Several iterations of the instrument were developed after consultation with peers and analysis of usability test results. A final version of the survey was distributed to RN’s who volunteered to participate. IRB approval was obtained.

Results
Tool diagnostics were performed as follows: category frequencies and average measures, as well as, threshold estimates, probability curves, and category fit statistics. Rasch analysis provided detailed information about the psychometric properties of the instrument.
Conclusions
Validity was articulated within the context of the Rasch model. Person and item reliability indexes were clarified in relation to defining reliability of the instrument and interpretation of the data.