

Thursday, June 13

Winchester Guildhall Foyer, 09.00 – 10.30 AM

MAIN CONFERENCE REGISTRATION

OPENING SESSION

10.30 – 12.30AM

Chair: Professor Bruce Bracken, The College of William and Mary, USA

Thursday, June 13

King Alfred Hall, 10.30 – 10.45 AM

WELCOME AND INTRODUCTION TO THE CONFERENCE

Presenter: Professor Dave Bartram, SHL Group plc, and ITC President

Thursday, June 13

King Alfred Hall, 10.45 – 11.15 AM

ITC HISTORY AND PREVIOUS CONFERENCES

Presenter: Professor Tom Oakland, University of Florida, USA, Past ITC President

Thursday, June 13

King Alfred Hall, 11.15 AM – 12.00 PM

KEY 1

THE IMPACT OF THE INTERNET ON TESTING: ISSUES THAT NEED TO BE ADDRESSED BY A CODE OF GOOD PRACTICE

Keynote: Professor Dave Bartram, SHL Group plc, UK

The presentation will review some of the ways in which testing is being carried out on the Internet. From this a range of issues relating to technology, security, privacy and fairness will be set out. A framework for considering how each of these issues can be addressed will be described as a potential basis for developing guidelines in this area. This framework is based around four different modes of test administration. I will discuss the implications each has for the degree of control that can be exercised over testing. In particular, I will consider the extent to which they pose challenges and opportunities for high versus low impact assessments and for measures of typical (e.g. personality inventories) or maximum (e.g. ability and achievement testing) performance. The Internet presents a number of challenges and opportunities for test developers, test users and test takers. Ways in which these might change the nature of testing and assessment procedures will be outlined.

Thursday, June 13

King Alfred Hall, 12.00 – 12.30 PM

THE CONFERENCE PROGRAMME

Presenter: Professor Ron Hambleton, University of Massachusetts, USA

KEYNOTES

13.30 – 15.00 PM

*Chair: Professor Thomas Oakland, University of Florida, USA***Thursday, June 13***King Alfred Hall, 13.30 – 14.05 PM***KEY 2****EXCITING NEW TECHNOLOGIES IN COMPUTER-BASED TESTING****Keynote:** Professor Wim J van der Linden, University of Twente, Netherlands

In hindsight, the history of test theory has long been hampered by the lack of computational power at the testing site. Already at its inception, it was known among test theorists that the precision of test scores would improve much if we could adapt the difficulty level of the test to the ability of the examinee. However, such adaptation required the estimation of the ability of the examinee during the test, and we had to wait until the advent of the computer revolution before real-time ability estimation became possible. In the first part of this presentation, an overview of the modes of CBT currently being explored in the research community and/or already applied in the testing industry will be given. We will highlight the important role played by item response theory (IRT) models as well as such new developments in statistics as Bayesian methods, multilevel modelling, and mathematical programming. In the second part of this presentation, we will show how these developments have made possible a variety of technological innovations in CBT. The cases discussed will include: adaptive testing with item cloning, item-exposure control, uniform speededness, equating of observed-scores, and multidimensional abilities.

Thursday, June 13*King Alfred Hall, 14.05 – 14.40 PM***KEY 3****NEW ITEMS AND NEW TESTS: OPPORTUNITIES AND ISSUES****Keynote:** Professor Fritz Drasgow, University of Illinois, USA

Advances in computer hardware and software and the growth of the Internet have created enormous opportunities for innovations and improvements in assessment. Although the development of paper-and-pencil tests lagged behind Gutenberg's introduction of movable type, that technological innovation enabled the growth of twentieth century testing. Computerization and the Internet will similarly enable twenty-first century assessment.

One way to organize a discussion of opportunities and issues related to the new assessments is to consider item types, test deployment, and scoring. Each offers tremendous opportunity for innovation albeit at the cost of solving thorny problems.

During the past twenty years, by far the most common innovation in item type for computerized assessment has been no innovation at all. Computerized versions of tests such as the Graduate Record Exam have simply delivered traditional paper-and-pencil items via computer. In contrast, the certified public accountant licensing exam will use a number of computer enhancements in simulation exercises. Even more novel are items that use multimedia capabilities to assess interpersonal skills such as leadership and teamwork.

The Internet provides a mechanism for a paradigm shift in test deployment. In paper-and-pencil testing, a small number of test forms with infrequent revisions were used due to printing and shipping requirements. With the Internet, item banks can be revised daily, provided that security can be maintained.

Some of the most difficult issues involve scoring. In contrast to multiple-choice items with simple right/wrong scoring, drag-and-drop and related formats may benefit from scoring algorithms that evaluate the degree of correctness. Response time and the process used to answer a question may also contribute to improved scoring.

Thursday, June 13

King Alfred Hall, 14.40 – 15.00 PM

QUESTION & ANSWER

Discussant: Professor Ron Hambleton, University of Massachusetts, USA

A: **SYMPOSIUM: PSYCHOMETRIC ISSUES** **15.30 – 17.30 PM**
Convenor & Chair: José Muñiz, University of Oviedo, Spain

Thursday, June 13

Wintonian, 15.30 – 15.50 PM

A1

UTILITY OF THE MANTEL-HAENSZEL PROCEDURE FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING WITH SMALL SAMPLES

Presenters: Ángel M Fidalgo, University of Oviedo, Spain
Doris Ferreres, Universidad de Valencia, Spain
José Muñiz, University of Oviedo, Spain

Sample-size restrictions limit the contingency table approaches based on asymptotic distributions, as the Mantel-Haenszel (MH) procedure, for detecting differential item functioning (DIF) in many practical applications. Within this framework, the present study investigated, using simulated data, the power and Type I error performance of MH procedure with small samples under a variety of conditions (item parameters, ability distributions and amount of DIF). A second goal of this study is to evaluate the relationships between the MH estimation of DIF (MH common odds ratio estimator) and the amount of DIF in relation to the number of examinees and item factors. The final purpose is to provide practical rules to assessment professionals when using the Mantel-Haenszel statistics with small samples, which is a very frequent situation in applied psychological evaluation. The practical rules will be designed to be applied to both type of evaluation settings: distant/direct and computer based/conventional.

Keywords: Differential Item Functioning, Small Samples, Mantel-Haenszel

Thursday, June 13

Wintonian, 15.50 – 16.10 PM

A2

A COMPARISON AMONG DIFFERENT COMPUTERIZED TESTS AND REVIEW CONDITIONS

Presenters: Julio Olea, Javier Revuelta, Carmen Ximénez, Vicente Ponsoda, Universidad Autónoma de Madrid, Spain and Pedro Hontangas, Universidad de Valencia, Spain

Three versions of a computerized English vocabulary test for Spanish speakers (adaptive, easy-adaptive and fixed) and four review conditions (no review, review at the end, review by blocks of 5 items and item-by-item review) were compared. Statistically significant effects were found in test precision among the different types of test. Response review improved ability estimates and increased testing time. Contrary to a previous study results, no psychological effects on anxiety were found. Examinees participating in the review conditions considered more important the possibility of review than those who were not allowed to review. These results concur with previous findings on examinees' preference for item review, and raise some issues that should be addressed in the field of computerized tests with item review. The link of this research with our previous studies on the psychometric and psychological consequences of different types of computerized tests will also be outlined.

Keywords: Item Selection, Item Review, Computerized Adaptive Tests, Computerized Fixed-Item Tests.

Thursday, June 13

Wintonian, 16.10 – 16.30 PM

A3

RASCH MODEL GOODNESS OF FIT THROUGH THE ANALYSIS OF THE PARAMETER INVARIANCE PROPERTY

Presenters: Pedro Prieto, University of La Laguna, Spain
Maribel Barbero & Juan C Suárez, UNED, Spain
Emilio Rodríguez, University of La Laguna, Spain
V Ponsoda, Universidad Autónoma de Madrid, Spain

The property of invariance of item and ability parameters is the cornerstone of Item Response Theory and its major distinction from classical test theory. However, since not always this advantage is found in practical research (Barbero, Prieto, Suárez, and San Luís, 2001; Fan, 1998; Gil, Suárez, and Martínez-Árias, 1999; Stage, 1999), it seems necessary to investigate this question in order to find new ways to establish whether this property is met in empirical works. In a recent research by Barbero et al. (2001) the use of Lord's χ^2 test, usually applied in Differential Item Functioning studies, is proposed for the analysis of the invariance property in the two-parameter logistic model. In the present research several simulation studies are conducted in order to estimate the fitting of the Rasch model under different conditions using Lord's χ^2 test applied to the study of parameter invariance. First results seem to confirm this method as a good way to ensure the goodness of fit of the Rasch model to a set of data.

Keywords: Estimates Invariance, Rasch Model Fit, Item Response Theory

Thursday, June 13

Wintonian, 16.30 – 16.50 PM

A4

FULL INFORMATION ESTIMATION OF ABILITY WITH MULTIPLE-CHOICE ITEMS

Presenters: Javier Revuelta & Francisco Abad, Universidad Autónoma de Madrid, Spain

Tests composed of multiple-choice items are usually scored using binary data (right or wrong responses). However, it is widely recognized that responses to incorrect alternatives contain information about the ability of the examinee. The psychometric models for scoring the test from polytomous (complete) responses present several drawbacks that keep them at a research stage and impede their general application. One of them is that the estimated ability may be reduced after the selection of the correct alternative instead of a distractor. This problem is reflected in the possibility to find one non-monotone observed response function for the correct response (NMRFC). Two solutions for this problem are presented in this communication. Thissen and Steinberg's multiple-choice model is the best known polytomous model for multiple-choice data and it suffers from the problem described above. However, there is frequently not enough information in the sample to establish the NMRFC as a reliable item property. Our first purpose consists of a modification in the E-step of the EM estimation algorithm of the Thissen and Steinberg multiple-choice model in order to avoid the consequences of the NMRFC undesirable characteristic. The second solution is a new psychometric model based on the criterion of rising selection ratios. When this criterion is met the problem of the reduction of ability is avoided. The model satisfies the criteria and permits the scoring of the individuals for any combination of the item parameters. On the counterpart, it is not as flexible as TS but it is still general enough to fit an ample variety of data from multiple-choice items.

Keywords: Multiple-Choice Items, Polytomous ITR Models, Thissen And Steinberg's Model

Thursday, June 13

Wintonian, 16.50 – 17.10 PM

A5

AN EVALUATION OF THE MULTIPLE-GROUP MEAN AND COVARIANCE STRUCTURE MODEL FOR DETECTING DIFFERENTIAL ITEM FUNCTIONING IN GRADED RESPONSE ITEMS

Presenters: Vicente González-Romá & Ana Hernández, Universidad de Valencia, Spain
Juana Gómez-Benito, Universidad de Barcelona, Spain

Items with ordered response categories such as Likert-type items are frequently used in Psychology. Based on the assumption that responses to these items approximate responses on a continuous line, the Multiple-Group (MG) extension of the Confirmatory Factor Analysis (CFA) with Mean and Covariance Structure (MACS) model has been used for evaluating differential item functioning (DIF) of polytomous graded response items (e. g., Chan, 2000). In contrast to evaluating DIF by fitting IRT models designed for the analysis of these type of items, the MG-MACS is more parsimonious and requires smaller sample sizes. In addition, the software employed to fit this model offers practical indices of overall fit and modification indices that are useful for detecting DIF. These advantages play a role when researchers have to decide which method they will use for detecting both uniform and non-uniform DIF. However, accuracy of DIF detection is one of the most important aspects to consider. There is a lack of studies analyzing the accuracy of the MG-MACS model as a method for DIF detection in polytomous graded response items. Thus, this is a question that must be clarified. The present study addresses this question using simulated data. The Generalized Partial Credit Model (Muraki, 1992) is used to generate responses to 10 items with five response options. Three factors are varied to generate the DIF conditions: sample size (400 and 800), type of DIF (uniform and non-uniform) and size of DIF (low, medium and large). For each condition 100 samples are generated. The data are analyzed by means of LISREL 8. Two indicators are calculated to determine the accuracy of DIF detection: the number of correct DIF identifications (true positives) and the number of DIF incorrect identifications (false positives).

Thursday, June 13

Wintonian, 17.10 – 17.30 PM

A6

DO ITEMS THAT MEASURE SELF-PERCEIVED PHYSICAL APPEARANCE FUNCTION DIFFERENTIALLY ACROSS GENDER GROUPS?

Presenters: Vicente González-Romá, Inés Tomás, Doris Ferreres & Ana Hernández,
Universidad de Valencia, Spain

The aim of this study is to test whether the six items of the Physical Appearance Scale (PAS), a scale included in the Physical-Self Description Questionnaire (PSDQ; Marsh et al. 1994), show Differential Item Functioning (DIF) across gender groups of adolescents. Roskam (1985) suggested that in personality items the discrimination parameter (a) is related to the extent to which the item is formulated in concrete, non-ambiguous terms. Gender groups may differ in the way they interpret the self-descriptions involved in physical appearance items. Girls are more sensitized than boys regarding physical appearance, and this may cause that girls interpret physical appearance items in a less ambiguous, more concrete way than boys. Therefore, we hypothesized that PAS items will show non-uniform DIF (i.e., differences in the a parameter estimates) across gender groups: the a parameter estimates will be larger in the girls group than in the boys group. This hypothesis is tested in two groups of Spanish adolescents (boys, N=499; girls, N=531). The results obtained yield partial support for our hypothesis: only one out of the six items analysed show DIF, and it was in the expected direction. Two additional analyses carried out to assess the “practical significance” of the DIF found revealed that it was trivial. The theoretical and practical implications of the results obtained are discussed.

Thursday, June 13

King Alfred Hall, 15.30 – 15.50 PM

B1

ITEM GENERATION AND THE INTERNET: EXPECTATIONS? NO, PRESCRIPTIONS FOR PROGRESS

Presenter: Sidney H Irvine, University of Plymouth, UK

Item generation theory (IGT) has been in use for 15 years or more in the construction of items for cognitive tests. A first round table attempt to gauge its potential in the task of modern test development was made in an invited seminar at ETS in 1998. The result of that is a collected volume (Irvine & Kyllonen, 2002) of papers, commentary and discussion. What might be learned from it in the delivery of tests at a distance?

Perhaps the fundamental lesson is the knowledge that computer delivery of tests compromises items faster than they can be written and standardised by conventional means, including scaling by item-response theory methods. Without a theory, for the mass production of generic item types to produce isomorphs to a prescribed level of difficulty, Internet testing renders most conventional tests useless in a very short time. The items and answers of tests having only one or two forms become widely available through repeated exposure internationally, with catastrophic effects on test security. To a similar extent, CAT (computerised adaptive testing) without item-generation will suffer the same fate, but in a slightly more subtle form. Wainer (2002) identifies the weakness of the CAT movement when he points out that only the most difficult items need be 'stolen' through memorisation and distributed once a solution is found.

The availability of algorithms for item production hand in hand with well-verified mental models of problem solving is one way forward out of the chains of traditional methods of item production. Different examples of the types of successes using IGT are shown.

Nevertheless, technical problems of delivery on the Internet remain the second critical issue. Where time is of no importance in the production of item responses, then the Internet is a perfect partner for Item-generation methods. If a test is timed, then interactive testing from a single source becomes more problematic. Some unresolved issues in test delivery where timing is critical are reported for discussion.

References

Irvine, S.H. and Kyllonen, P.C. (Eds.), (2002). *Item generation for test development*. Lawrence Erlbaum Associates, Mahwah, NJ, USA.

Wainer, H, (2002). *On the automatic generation of test items: Some whens, whys and hows*. In Irvine, S.H. and Kyllonen, P.C. (Eds.). *Item generation for test development*. Lawrence Erlbaum Associates, Mahwah, NJ, USA.

Thursday, June 13

King Alfred Hall, 15.50 – 16.10 PM

B2

THE PSYCHOMETRIC IMPACT OF USING AUTOMATICALLY GENERATED ITEMS

Presenter: Catherine M Hombo, Educational Testing Service, USA

Automatic item generation allows a computer to create new assessment items for examinees. These items are generated from a “shell”, a parent item that has some fixed aspects and some variable aspects. These new items are intended to assess the same construct as the parent item. In addition, the new items are presumed to be isomorphic to the parent item (i.e., they are expected to have the same statistical item characteristics). If this assumption holds, then the “family” of items can be treated as a single item in calibration and estimation. This assumption of isomorphism is a strong one, and requires scrutiny. Failure of this assumption in operational assessments potentially could lead to incorrect decisions about examinees in high-stake situations.

Results from simulation studies have indicated that isomorphism does hold to an acceptable extent for operational conditions, with limited bias in either the item parameters or the examinee ability estimates above that inherent to the estimation procedure (Dresher & Hombo, 2001; Glas & van der Linden, 2001; Hombo & Dresher, 2001; Matthews-López & Hombo). Results from a special study done as part of the US National Assessment of Educational Progress (NAEP) will be presented. In this study, it was demonstrated that items can be generated on the fly and delivered across the Internet to examinees in remote locations. In addition, pre-generated fixed forms were administered with sufficient sample sizes to examine the assumption of isomorphism in this data.

References

- Dresher, A. R. & Hombo, C. M. (2001). *A simulation study of the impact of automatic item generation on item and ability parameter estimation*. Paper presented in the session “Items: The next (automatic) generation” at the annual meeting of the National Council on Measurement in Education, Seattle, WA, April 2001.
- Glas, C. A. W. & van der Linden, W. J. (2001). *Modelling variability in item parameters in CAT*. Paper presented at the Annual Meeting of the Psychometric Society, King of Prussia, PA.
- Hombo, C. M. & Dresher, A. R. (2001). *A simulation study of the impact of automatic item generation under NAEP-like data conditions*. Paper presented in the session “Items: The next (automatic) generation” at the annual meeting of the National Council on Measurement in Education, Seattle, WA, April 2001.
- Matthews-López, J. L. & Hombo, C. M. (2001). *Modelling the hyerdistribution of item parameters to improve the accuracy of recovery in estimation procedures*. Paper presented in the session “Items: The next (automatic) generation” at the annual meeting of the National Council on Measurement in Education, Seattle, WA, April 2001.

Thursday, June 13

King Alfred Hall, 16.10 – 16.30 PM

B3

PRINCIPLES OF ITEM GENERATION

Presenter: Professor Patrick Kyllonen, Educational Testing Service, USA

Automatic item generation methods are useful for creating items at a targeted difficulty level, developing parallel forms of a test, ensuring item security, and focusing item development on the construct being measured. Devising an item generation scheme for a particular test forces the researcher to differentiate the difficulty from the non-difficulty controlling factors (“radicals” vs. “incidentals”) underlying test items. Further, differentiating construct-related versus construct-unrelated radicals puts construct validity at the forefront of test design. Three principles are important for item generation: (a) The Item-Generation-Design Principle (the design of the radicals defines the item templates; the design of the incidentals populates the templates and enables the creation of parallel forms), (b) The Item-Variability Principle (to minimize carryover effects, incidentals should minimize the similarity of parallel items but not affect item difficulty); and (c) The Not-All-Tests-Are-Alike Principle (for complex tests, such as working-memory and procedural knowledge tests, radical variability is important; for processing speed and learning tests, incidental variability is important). These principles are illustrated with a cognitive test battery designed to measure changes in performance over time. Expanding item generation to new tests and new applications will be discussed. Knowledge tests are prime candidates for future work, and applications such as medical (e.g., neurological), and readiness (e.g., “cognitive sobriety tests”) batteries could benefit from the item generation approach. Further developments in item generation will move us closer to a science of test design centered on construct validity.

Thursday, June 13

King Alfred Hall, 16.30 – 16.50 PM

B4

**MULTILINGUAL ITEM MODELLING AS A MECHANISM FOR TEST ADAPTATION:
APPLICATIONS TO OPEN ENDED AND DISCRETE ITEMS**

Presenters: Isaac I Bejar, Cedrick Fairon & David M Williamson
Educational Testing Service, USA

The ubiquity of the computer as a test delivery tool has fuelled an increasing trend toward greater use of automated techniques for test development and scoring of both discrete and open-ended tasks. The increased automation of every aspect of assessment design, administration and scoring has implications for test adaptation. In this presentation we offer the concept of task/item modelling as an approach to addressing the problem test adaptation for open ended and discrete item formats.

A task model is an abstract, and therefore potentially language independent, representation of the assessment task seen by the examinees. The abstract representation allows the instantiation of fundamentally equivalent tasks facilitating scoring by an identical by means of automated procedure. That is, the invariance of the deep structure representation of the item, and therefore the automated scoring procedure for the item, can be leveraged to facilitate multilingual implementation. Similarly, an item model is an abstract representation from which a large number of psychometrically equivalent items can be generated from a deep structure or radical. By decoupling the instantiation of natural language from the generation of the deep structure it becomes possible to generate fully formed items in other languages.

The first example we discuss concerns a licensing examination for architects. The English version of the assessment was done as a series of task models from which isomorphic instances of the task were prepared. The approach to scoring will be described and shown to be invariant over instances. Indeed, a French-language version of the assessment was subsequently developed which uses the same scoring engine as the English-language version. The second example we discuss is concerned with the assessment of deductive reasoning. We discuss an approach that separates the reasoning and linguistic component. The linguistic component allows the instantiation of items in any language by means of a suitable "grammar" for a specific language. The feasibility, of producing test content in several languages through item/task modelling in no way avoids the problem of equivalence of score across languages or other issues of score meaning. Rather, modelling provides a mechanism for facilitating the process and making it more efficient to implements assessments in other languages, especially in computer-based contexts. At the same time, the more careful analysis required by a modelling approach should lead to better articulated test designs, which in turn should be instrumental in better test adaptation.

Thursday, June 13

King Alfred Hall, 16.50 – 17.10 PM

B5

SCHEMA BASED GENERATION: THE ONLINE NUMERICAL REASONING TESTING SYSTEM

Presenters: Helen Baron, SHL Group plc, UK
Anthony Miles, Psychometric Services Ltd, UK

This paper is based on a project to create a non-proctored numerical ability test which applicants to a financial services organisation could complete together with an online application form. The organisation faced the difficulty of processing a large number of candidates for a job requiring high levels of numerical reasoning quickly and effectively.

The item design was based on commercially relevant questions to be answered on the basis of information in a data table or graph with five answer options. A set of 100-item prototypes was generated by item writers based on 15 different tables. Each item prototype formed the basis for 2 variants by repopulating the table with different data, or altering the item or response options slightly. Items were pilot tested on a sample of over 500 students and graduate bankers. The selected items were calibrated initially with a one-parameter IRT model and the operational test consisted of 18 items chosen in a constrained random fashion from the item bank.

An internet based deliver system was designed to allow candidates to access the test. In validity studies with three different samples it correlated 0.47 with a longer, proctored paper and pencil test of numerical reasoning and proved a very effective sifting measure.

Following operational use with over 9,000 candidates the item bank was recalibrated using a two parameter model and comparisons made between item variants. The typical variation in parameters across variants and the factors which promote and detract from parameter invariance were then examined.

C: VALIDITY ISSUES

15.30 – 17.30 PM

Chair: Professor Jacques Gregoire, Université Catholique de Louvain, Belgium

Thursday, June 13

Winchester Conference Chamber, 15.30 – 15.55 PM

C1

A PROPOSED STRATEGY FOR VALIDATING COMPUTER AUTOMATED SCORING METHODS

Presenters: Chad W Buckendahl, Barbara S Plake & James Impara
University of Nebraska-Lincoln, USA
Yongwei Yang, The Gallup Organization, USA

Computer automated scoring (CAS) is emerging as a popular tool in both traditional and computerized test delivery programs. Various studies have evaluated the quality of the CAS generated scores. However, research has generally highlighted the level of inter-rater agreement between human raters and the CAS procedure. Furthermore, the research has typically been conducted by the organizations that have the most to gain by favorable performance of their methodologies. The underlying question of the validity of the scores generated from this scoring process still lingers. This paper presents a conceptual framework for designing validation studies for CAS procedures. An additional goal is to review the current practice of validating automated scoring and synthesize these studies into a general description. An empirical comparison of multiple CAS procedures is offered as an illustration of the framework. A discussion of the implications both theoretical and empirical that arise from using this new technology is also included.

Thursday, June 13

Winchester Conference Chamber, 15.55 – 16.20 PM

C2

STANDARDS FOR TEST QUALITY: APPLICATIONS TO COMPUTERIZED-BASED TESTS

Presenters: Barbara S Plake, James C Impara & Chad W Buckendahl,
University of Nebraska-Lincoln, USA

Standards for the technical quality of educational and psychological tests have been available since the mid 60's. With the introduction of computer as a medium for test delivery, the applicability of these technical quality standards has been drawn into question. Several efforts have been directed at the development of special standards for computer-based test. The first, sponsored by the American Council on Education, was published in 1994 and focused on computerized-adaptive test development and use in education. More recently, the Association of Test Publishers has been working on set of standards more broadly aimed at computer-based testing. The purpose of this presentation will be to highlight these efforts to articulate technical quality criteria for computer-based tests. In addition, this presentation will discuss special challenges to meeting current expectation for technical quality with computer-based tests. In particular, the presentation will focus on issues related to the gathering content validity evidence with computerized-adaptive tests and also special issues in setting passing scores on tests that are delivered by computer, especially related to tailored computerized testing. Finally, the presentation will identify areas where additional attention is needed to ensure that tests delivered by computer meet established standards for technical quality.

Thursday, June 13

Winchester Conference Chamber, 16.20 – 16.45 PM

C3

EQUIVALENCE OF COMPUTERISED AND TRADITIONAL ABILITY TESTS: TEST FAIRNESS AND CONSTRUCT VALIDITY

Presenter: Dorothea Klinck, Bundesanstalt für Arbeit, Germany

The target of the study was to test the equivalence of paper-and-pencil and computerised versions of cognitive ability tests which are administered to adults who are assessed by the Psychological Service of the German Federal Employment Services (sample size: about 6700 persons). The study covered the question of psychometric equivalence (means, dispersions and shapes of score distributions, equivalence on item level by using IRT modelling, invariance of construct validity), equivalence across groups (e.g. age, sex, educational level) and individual variables (anxiety, computer experience, attitude towards computers), and equivalence of perception of the test administration by the respondents. The presentations will focus on two aspects of the study:

1. Is there a test bias of the computerised versions against people with a lack of computer experience?
2. How can equivalence concerning construct validity be examined using a between subjects design?

Results indicate that the relation of computer experience to test results is invariant across administration modes (positive correlation independent of the mode of administration). The question of construct validity was addressed by using structural equation modelling (multi-sample analysis), which indicated that the tests measure the same construct no matter which mode the test are administered.

Thursday, June 13

Winchester Conference Chamber, 16.45 – 17.10 PM

C4

USABILITY OF PSYCHOMETRIC ADMEASUREMENTS

Presenter: J M Müller, Tubingen, Germany

Test users too often neglect the psychometric properties of a test as indicators of several aspects of quality (Archer, Maruish, Imhof & Piotrowski, 1991; Piotrowski & Keller, 1992; Wade & Baker, 1977). This may also hold in computer-based testing. For this reason, a theory-based analysis of the weakness of psychometric coefficients is recommended. A list of about 20 criteria for evaluating the usability of psychometric admeasurements with respect to the foundation, scale and interpretation of a coefficient is presented. Based on these criteria, several coefficients of reliability are evaluated. Because of certain weaknesses of these coefficients, alternatives are proposed, especially a coefficient for displaying the performance on separate test scores 'Personenunterscheidungsvermögen' (Müller, 2001) and a rescaled reliability coefficient ('Differenziertheit', Müller, 2001). Both coefficients are introduced and discussed. Additionally, SAS-Marcos to calculate the coefficients are given.

References

- Archer, R. P., Maruish, M., Imhof, E. A. & Piotrowski, C. (1991). Psychological test usage with adolescent clients: 1990 survey findings. *Professional Psychology: Research and Practice*, 22, 247-252.
- Müller, J. M. (2001). Kennwerte psychologischer Testverfahren. Dissertation. http://elib.suub.uni-bremen.de/publications/dissertations/E-Diss165_M%FCller.pdf
- Piotrowski, C. & Keller, J. W. (1992). Psychological testing in applied settings: A literature review from 1982-1992. *Journal of Training & Practice in Professional Psychology*, 6, 74-82.
- Wade, T. C. & Baker, T. B. (1977). Opinions and use of psychological test. *American Psychologist*, 32, 874-882.

D: CBT APPLICATIONS

15.30 – 17.30 PM

Chair: Professor Ron Hambleton, University of Massachusetts, USA

Thursday, June 13

Walton Room, 15.30 – 15.50 PM

D1

SIMTEST: A FOREIGN LANGUAGE PROFICIENCY CAT

Presenters: M Sumbling & P Sanz, Servei d'Idiomes Moderns, Universitat Autònoma de Barcelona, (SIM/UAB) Spain
M C Viladrich & E Doval, Laboratori d'Estadística Aplicada, Universitat Autònoma de Barcelona, Spain

SIMTEST is a computer-adaptive test of foreign language proficiency that classifies examinees according to 6 levels as defined by the Council of Europe. It was developed for placement and certification purposes and has been edited in CD-ROM format for individual or collective use. Initial estimation of a candidate's level is provided by 4 C-Tests, followed by a series of multiple choice questions [MCQs] to give a refined estimation. An optional self-assessment component is included. It takes around 30 minutes to complete the whole test. All SIMTEST components were initially selected by expert judgement prior to piloting. C-Test and MCQs were calibrated by Classical Test Theory before item-banking. Using the C-Test result as an entry point, MCQs are presented on a CAT administration based on the algorithm described by Henning (1987).

750 new enrolments on SIM/UAB English courses in the Barcelona area were tested for placement purposes in September 2001 using the C-Test and multiple choice components of SIMTEST. Their scores were compared to the placement decision made by an expert. C-Tests presented difficulties for certain examinees with lower levels of language proficiency, causing some to abandon, perhaps because the format is relatively unfamiliar in Spain. The number of MCQs delivered ranged from 6 to 23 items. Scores assigned by SIMTEST covered the full range of possible proficiency levels. The acceptable degree of agreement found between the SIMTEST placement result and the level assigned by experts encourages further investigation of its psychometric properties in order to improve the test.

Thursday, June 13

Walton Room, 15.50 – 16.10 PM

D2

COMPUTER-BASED SIMULATIONS TO ASSESS PHYSICIANS' PATIENT-MANAGEMENT SKILLS

Presenters: Brian E Clauser & Stephen G Clyman
National Board of Medical Examiners®, USA

The National Board of Medical Examiners® develops the USMLE® Step Examinations, which are the tests used for physician licensure in the United States. Since 1999, the USMLE® Step 3 Examination has included an innovative computerized item format designed to assess physicians' patient management skills. The format, known as Primum® computer-based case simulations, provides a dynamic simulation of the patient care environment in which the simulated patient's condition changes based on both the examinee's actions and the patient's underlying problem. The examinee manages the case by making free-text entries to order tests, treatments, and consultations. As the examinee advances the case through simulated time, realistic feedback on changes in the patient's condition is provided. The simulation is scored using a computer algorithm designed to yield a result approximating that which would have been produced by expert clinicians rating the performance. The presentation will provide an explanation and demonstration of the simulation format as well as an overview of the research on the format.

Thursday, June 13

Walton Room, 16.10 – 16.30 PM

D3

THE PAPERLESS EXAMINATIONS PROJECT - STEPS TOWARDS INTRODUCING COMPUTER-BASED EXAMINATIONS

Presenters: Dorit Reppert, Alastair Walker & Martyn Roads
CCEA, Northern Ireland

This phased Project, undertaken jointly by two Awarding bodies (CCEA in Northern Ireland and Edexcel in England), set out in October 2000 to investigate the implications and technical feasibility of introducing an ICT dimension to high-stakes examinations such as GCSE. The increasing use of ICT in teaching and learning requires that consideration be given to the use of computer technology that allows.

- an increased flexibility in exam delivery;
- the enhancement of the current assessment methodologies;
- the delivery of new forms of assessment, for example by use of multimedia facilities; and
- that either the computer or, where professional marking is required, an examiner working on-screen, can mark examinees' responses to questions.

In the first phase of the Project 120 Students took on-screen pilot tests based on GCSE Science. The tests, taken in April 2001, were partially marked by computer and partially marked on-screen by examiners. The outcomes of the first phase of the project showed that the majority of the participating students enjoyed the experience and that there are no insurmountable technological difficulties for setting up the infrastructure necessary for computer-based exams.

The second phase of the Project started in October 2001. It is focusing on issues of electronic test item development, training for examiners, design and format of on-screen exam 'papers', and again the combination of electronic and human marking of responses to questions. It is envisaged that electronic tests, based on GCSE Science and GCSE Geography, will be piloted in approximately 10 centres in Northern Ireland and 10 centres in England. Furthermore, this phase will look at the trial of processing online Basic and Key Skills tests.

Thursday, June 13

Walton Room, 16.30 – 16.50 PM

D4

DYNAMIC ASSESSMENT OF LEARNING POTENTIAL BY MEANS OF LINKED COMPUTERISED ADAPTIVE TESTS

Presenter: Marie De Beer, UNISA, South Africa

Learning potential assessment emphasises development and allows for improvement in cognitive performance if relevant training is provided. The accurate measurement of improvement in performance in the dynamic test-teach-retest strategy typically used to assess learning potential is more readily achieved by means of IRT-based computerised adaptive testing.

While the use of classical test theory leads to measurement problems regarding difference scores, the use of IRT latent trait models and CAT provide a means of accurately equating scores and providing a solution to many of the problems that have been associated with dynamic assessment and the measurement of learning potential.

Separate item banks are used for the pretest and the post-test. Items are therefore not repeated and memory does not confound test performance. By linking the pre- and post-tests, the exit level attained in the pretest can be used as entry level in the post-test, thereby further improving the accuracy of measurement. Because measures are obtained on the IRT latent ability scale, the difference between post-test and pretest scores reflects change in performance (latent ability) due to training and not change due to different difficulty levels of the two tests.

The answering procedure is very simple and use of only the space bar and the enter key allows for the assessment of learning potential of from adult illiterate to tertiary level examinees. The use of IRT and CAT procedures can provide the psychometric base needed for a technically sound dynamic assessment instrument for the measurement of learning potential.

Thursday, June 13

Walton Room, 16.50 – 17.10 PM

D5

AUTOMATIC ITEM GENERATION METHODOLOGY: THEORY, CURRENT PRACTICES, AND FUTURE DIRECTIONS

Presenter: Mary Pitoniak, University of Massachusetts, at Amherst, USA

Interest in automatic item generation (AIG) procedures has grown as computer-based tests, which require larger pools of items for security purposes, have become more common. In AIG, the computer is used as an item creation adjunct in order to produce items more efficiently—and eventually, after initial cost and effort outlays, more economically. In this paper, a review of the theory and practice of AIG is presented. Early methods that grew out of the criterion-referenced testing movement and more advanced approaches tied to cognitive psychology are described. Their operational implementation and important directions for future research will be highlighted in the paper.

Thursday, June 13*Keats Room, 13.00 – 17.30 PM***P1****ASP = Assessment Service Provider?****Presenter:** Dr Rainer Kurz, SHL Group plc

The face of psychometric assessment is changing rapidly in the Internet age. All functionality of an "Intelligent Testing System" (ITS) as described by Bartram & Bayliss (1984), namely test choice, instrument administration, scoring, analysis, interpretation, feedback and decision making, can be automated and delivered via the Internet. The poster explores potential problems and opportunities afforded by the new medium with specific reference to the Application Service Provider (ASP) model in occupational assessment.

ASP delivery redefines the roles of assessment professionals, candidates and managers, enabling a high degree of self-service and self-sufficiency, but also increased sophistication of assessment practice. This paradigm change needs to be accompanied by increased clarity and availability of information resources that should be integral to the ASP offering, and changes to the training certification regimes.

In the Internet ASP model information can be shared widely across geographical areas, and applications, thus changing the landscape of assessment practices. It enables better management of assessments; solution-oriented integration of test results and empowers the assessment process stakeholders. In the occupational assessment field this is accompanied by a gradual shift from trait to competency oriented measurement, from individual to organisation measures, and from clinical to actuarial basis of predictions.

This paper argues that a lot of assessment tasks can be automated and delivered via the Internet. However a lot of fine professional judgement is required in realising the potential of the web - the ultimate responsibility for setting up and using Assessment Service Provider offerings rests with the human.

Thursday, June 13*Keats Room, 13.00 – 17.30 PM***P2****COMPUTER-BASED ASSESSMENT IN THE BELGIUM ARMED FORCES****Presenter:** Yves Devriendt, Belgium Armed Forces, Belgium

Military, technological and methodological evolutions have influenced a lot the military selection procedures. During World War I and World War II solid testing programs were developed in order to select and allocate large numbers of recruits. Traditionally the armed forces have always used large-scale selection programs. Nowadays the armed forces are facing all over the world recruiting problems. One of the possible reasons is the use of non-automated, time-consuming selection methods that are not client-centred. Other disadvantages of traditional paper & pencil test batteries are elaborated. Some possible solutions to make the military selection more attractive in the eyes of the candidates, are item generative and computer-adaptive testing systems. Psychometric advantages in using these systems are: better reliability and validity, and a more effective and efficient decision-making process. The Belgian Armed Forces have made a blue print for the validation of an automated testing procedure and the outlines of the project design are discussed. This poster ends with some remarks concerning the positive and negative sides of automation in the field of selection and a discussion on future developments concerning computer-based testing.

Thursday, June 13

Keats Room, 13.00 – 17.30 PM

P3

COMPUTER-BASED OBJECTIVE PERSONALITY ASSESSMENT: A RULE TRANSGRESSION TEST

Presenters: V Rubio, P Shi & J Santacreu, Universidad Autónoma de Madrid, Spain

Personality assessment has been traditionally based on paper-and-pencil (P/P) questionnaires and self-reports in which people have to respond about the appropriateness to them of several sentences. However, questionnaires are vulnerable to motivational distortion. Moreover, there is a strong controversy about the relationships between what people say about their behaviour and how they behave. Alternative to traditional P/P personality assessment, objective personality tests assess personality dimensions using a set of tasks to which people do not know what is really measured. Thus, they are not able to change their responses according to social desirability or expectations about the profile the assessors are looking for. That was what Cattell & Warburton (1967) called T-data.

The present paper presents a new objective test for the assessment of a personality dimension: rule transgression. Rule transgression is commonly associated with fairness or responsibility traits from the traditional perspective. It is usually requested for personnel recruitment contexts. However, the assessment biases mentioned above are particularly severe in the assessment of a dimension such as rule transgression.

The Rule Transgression task consists of 5 one-minute trials in which people have to guide mobiles to the end of a maze **only** when the rules allow. The test assesses competency, efficiency, and rule violation, in order to discriminate the rule transgression tendency from other variables.

Results in terms of internal consistency, correlation between rules violation, competency and efficiency indexes, correlation between other dimensions and differences between violators and non-violators are shown.

Thursday, June 13

Keats Room, 13.00 – 17.30 PM

P4

COMPUTER-BASED ASSESSMENT TRAINING WITH INTERNET-APPLICATIONS IN AN EDUCATIONAL SETTING

Presenters: Mark Schittekatte, Marc Covents, Griet Vermeir & Paul Verhaeghe
Ghent University, Belgium

At the Ghent University a training programme for psychology students in computer-based assessment is operational. Since October 2000, several case studies were developed to train student skills regarding psychological testing and in particular decision-making in clinical assessment. Recently streaming video and Internet-applications were added as new dimensions in these case studies. Problems, pro's, contra's and future plans for diagnostic training of psychology students through computer-based training with Internet-applications are discussed.

Thursday, June 13

Keats Room, 13.00 – 17.30 PM

P5

FACULTY & STUDENT PERCEPTIONS OF SECURITY AND CONFIDENTIALITY OF AN ONLINE STUDENT RATING SYSTEM

Presenter: Cheryl Davis Bullock, University of Illinois at Urbana-Champaign, USA

The Division of Measurement and Evaluation at the University of Illinois has developed an online faculty evaluation system entitled EON. This system uses the University's secure verification system to validate instructor, student, and administrator access. It is co-ordinated and housed on a secure server within the Office of Instructional Resources. Faculty log onto the system (choosing from an item pool of questions) and create mid-semester and end-of-semester course and instructor evaluations for their students to complete. Once instructors have submitted semester grades they can immediately access the evaluation results via the web. Additionally, instructors can analyse results by demographic variables such as class rank and gender.

Students access the EON evaluation forms over a two-week period via the web. Since open-ended comments are typed, there is no chance for instructors to recognize handwriting, thus strongly preserving anonymity. Once faculty authorize the release of this information, authorized administrators may view results of their faculty individually or as a group by semester via the web. These results are used in tenure and promotion decisions.

The poster session proposed will focus on the ethical issues involved with confidentiality and security of the EON system. The safeguards built into the system will be presented. But of most interest, our office has conducted interviews with students, faculty, and administrators to understand the trust that they have in the system's ability to protect their privacy. Results from this study will be compared and contrasted with the traditional paper and pencil student ratings system.