

KEYNOTES

13.30 –

15.00 PM

Chair: Professor Ron Hambleton, University of Massachusetts, USA**Friday, June 14***King Alfred Hall, 13.30 – 14.05 PM***KEY 6****OPTIMIZING QUALITY IN THE USE OF WEB / COMPUTER BASED TESTING (WBT / CBT) FOR PERSONNEL SELECTION****Keynote:** Professor Lutz F Hornke, University of Aachen, Germany

Computer or web based testing is used in several ways in personnel assessment within organisations: First, it makes scoring procedures easier, second, it uses technological features as in CAT to better estimate a person's ability, and, third, it may allow to design items that make use of certain multimedia aspects like simulation and dynamic presentation formats in order to better diagnose problem solving abilities. All three aspects are demonstrated by examples from the web, publishing companies, and psychological labs. In particular, special advantages of adaptive testing like item tailoring and saving of testing time are mentioned.

Another main advantage comes with computer based scenarios which present tasks that seem to be highly motivating to participants and do involve novel demands which are deemed to have a higher face validity than intelligence test. However, there are serious problems with so called complex problems, (1) in that the overall reliability is less than satisfactory, (2) the existence of a task-independent and thus generalisable problem solving ability has not yet been substantiated, (3) it is unclear whether a new measurement method asks for defining an independent new ability construct.

Above all construct, psychometric, and procedural problems which professional psychologists encounter while developing and applying their cbt/wbt test there are non-psychologists with nothing but programming competencies who begin to develop "tests" as well. While traditional tests were mainly developed out of sound knowledge of content / construct matters and in regard to a diagnostic questioning, with cbt/wbt there are professionals like economists or engineers without any substantive knowledge of psychological constructs and in psychometric methods.

So quality assurance is a major point that a society needs in order to arrive at sound personnel decisions. It asks for construct and psychometric sound item design, for proven qualities of modernistic "problem solving scenarios", and for state-of-the-art test and program design.

Friday, June 14

King Alfred Hall, 14.05 – 14.40 PM

KEY 7

COMPUTER-BASED TESTING (CBT) FOR PROFESSIONAL LICENSING AND CERTIFICATION OF HEALTH PROFESSIONALS

Keynote: Dr Donald E Melnick, National Board of Medical Examiners[®], USA

The National Board of Medical Examiners (NBME[®]) develops the United States Medical Licensing Examination[®] (USMLE[®]). In 1995, NBME implemented its first computer-delivered test; USMLE was converted in 1999. NBME adopted several strategies in implementing CBT programs:

1. Relied upon large item banks and many forms with random form assignment to assure examination security.
2. Emphasized examinee convenience.
3. Avoided major design changes and innovations such as adaptive item or testlet selection.

We introduced computer-based clinical simulations. Primum[®], the NBME's proprietary simulation system, presents realistic unfolding clinical scenarios which are "managed" by the examinee using an uncued entry system. Performance is scored by automatically evaluating examinee actions against expert-derived management approaches and modelling expert-rated examinee performances.

Examinees clearly prefer CBT, and other stakeholders seem pleased. CBT, particularly the move to continuous testing, required dedication of substantial resources in order to develop very large item banks, to generate and perform quality control on dozens of test forms, and to develop systems for managing the interaction with the test delivery vendor as well as continuous test administration, scoring, and reporting. Test security challenges have shifted from preventing pre-administration access to test materials to controlling item exposure during test administration. Documentation of the use of time by examinee groups raises new questions about examination timing. Changes in test format and content raise new questions about validity, score scaling, and standard setting. CBT promises to offer new opportunities for assessment as it matures.

Friday, June 14

King Alfred Hall, 14.40 – 15.00 PM

QUESTION & ANSWER

Discussant: Dr Charles Johnson, Chair of the British Psychological Steering Committee on Test Standards, UK

**K: SYMPOSIUM: ON THE INTERPLAY BETWEEN 15.30 AM–
17.30 PM
E-ASSESSMENT AND E-LEARNING: ISSUES AND PRACTICE
Convenor & Chair: Michal Beller, Educational Testing Service, USA**

Friday, June 14

King Alfred Hall, 15.30 – 15.55 PM

K1

DESIGN OF ASSESSMENT FOR E-LEARNING ENVIRONMENTS: TRADITIONAL AND NEW APPROACHES

Presenter: Michal Beller, Educational Testing Service, USA

E-learning environments are gaining popularity among distance learners as well as at universities, K-12 institutions, and corporations. This presentation will outline the new challenges and opportunities opened when both learning and assessment take place at the same time, in the same environment, using the learning material and the recorded students behaviours. An effort will be made to systematically identify factors that determine the quality of e-learning courses.

In considering new methods of assessment and feedback for new learning environments there is a need to refer back to the students' and teachers' needs and to ask in what way e-learning is different from traditional modes of teaching and learning and how technology allows us to rethink what we are doing. For example, a key to success in learning online is sustaining the motivation for students to learn. Maintaining quality interaction with peers, with mentors and with the material is an essential component in the design of online learning environments. Part of an effective interaction has to do with providing ongoing feedback to each student. Such feedback requires sophisticated, diagnostic assessment and provision of relevant information to the learner and the mentor. Thus, both formative and summative assessments can and should become integral parts of the learning process online. Collaborative work can be performed and assessed at both the group and the individual levels. Prior knowledge and skills can be assessed, enabling the learning environment to become more student centered and more adapted to learning styles.

Friday, June 14

King Alfred Hall, 15.55 – 16.20 PM

K2

THE NAEP PROBLEM SOLVING IN TECHNOLOGY RICH ENVIRONMENTS PROJECT (TRE)

Presenter: Randy Bennett, Educational Testing Service, USA

This paper describes the Problem Solving in Technology-Rich Environments (TRE) study. The TRE study is one of several studies associated with the Technology-Based Assessment (TBA) project. The purpose of the TBA project is to lay the groundwork for incorporating new technology in NAEP, the National Assessment of Educational Progress. The TRE study is producing a set of example modules to assess problem solving with technology, which can be used to explore issues related to the development, delivery, and scoring of electronic performance assessments in NAEP. The TRE modules are built around simulation and electronic information search. Among other things, the modules are designed to incorporate incidental learning as a goal of good assessment, capture the multidimensional nature of problem solving in technology environments, take advantage of the unique capabilities of the computer, and disentangle component skills to describe student characteristics more

meaningfully. In operational NAEP assessments, many such modules might be randomly spiralled among groups of students to provide evidence of problem solving with technology generally. Alternatively, a few such modules might be combined with a traditional subject-matter survey as a means of adding depth to the picture of what students know and can do.

Friday, June 14

King Alfred Hall, 16.20 – 16.45 PM

K3

EMBEDDED ASSESSMENT WITHIN A SCAFFOLDED LEARNING ENVIRONMENT FOR STATISTICS

Presenter: Malcolm Bauer, Educational Testing Service, USA

The goal of this project is to explore the ways in which embedded assessments can be used in the service of learning. We have developed a prototype assessment module for hypothesis testing within the context of an AP statistics curriculum. We have worked through several of versions of the prototype with substantial involvement from AP teachers and students. Feedback has taken the form of interviews, demonstrations, pilot studies, usability tests, and field trials. Our current version is a web-delivered system that can be run on several different platforms, and has the following components:

1. A scaffolded learning environment that supports learners of a domain. Learners solve problems and seek help (scaffolding) as needed. As learners improve, they naturally rely less on the scaffolding until they can solve problems on their own.
2. Metacognitive prompts that encourage learners to reflect on strategies that they use to solve problems and explain their reasons for using those strategies.
3. Self-assessment activities that allow learners to compare their solutions and explanations with those of experts, providing additional opportunities for self-explanation. This assessment activity encourages learners to internalize evaluation criteria (Lavigne & Lajoie, 1996). We expect this activity will help teachers to better diagnose and remediate student work via these assessments.

There are many forms of assessment within the prototype, and the talk will highlight the ways in which they are used to support learning. Systems like the one we are developing can be used by students in either a classroom or distance-learning environment or as a teacher training aid for preservice and beginning teachers.

Friday, June 14

King Alfred Hall, 16.45 – 17.10 PM

K4

ASSESSMENT AND TECHNOLOGY

Presenter: Henry Braun, Educational Testing Service, USA

Formal assessment has always been a critical component of the educational process, determining entrance, promotion, and graduation. In many countries over the last two decades, assessment has assumed an even more salient place in governmental policy. Attention has focused on assessment because of its role in monitoring system functioning at different levels (students, teachers, schools, and districts) for purposes of accountability and, potentially, spearheading school change efforts. At the same time, new information technologies have exploded on the world scene with enormous impact on all sectors of the economy, education not excepted. In the case of education, however, change has come mostly at the margins. The core functions of most educational systems have not been much affected. Nonetheless, many hold the belief that the convergence of powerful computers, multimedia, and communication networks will eventually leave their mark on the world of education, and on assessment in particular.

In this paper, we provide a structure for exploring the relationship between assessment and technology and argue that it is important for educators to appreciate the different ways in which technology increasingly influences assessment. Without a deeper understanding of this relationship, educators will be hard pressed to harness technology in ways that are educationally productive.

In our analysis, we distinguish between direct and indirect effects of technology. By the former, we refer to the tools and affordances that can, or will, change the practice of assessment and are the principal focus of attention in the education literature. Excellent examples are provided by Bunderson et al. (1989) and Bennett (1999). In both studies, the authors project how the exponential increase in available computing power and the advent of affordable high-speed data networks will affect the design and delivery of tests, lead to novel features, and, ultimately, to powerful new assessment systems.

There is noticeably less attention to what we might term the indirect effects of technology; that is, how technology helps to shape the context in which decisions about assessment take place. These decisions, concerning priorities and resource allocation, exert considerable influence on the evolution of assessment. Indeed, one can argue that science and technology give rise to an infinite variety of possible assessment futures, while forces at play in the larger environment determine which of these futures is actually realized.

L: SYMPOSIUM: PRACTICAL ISSUES AND ADVANCES 15.30 AM – 17.10 P
M
IN COMPUTER-AIDED ASSESSMENT
Convenor & Chair: Joerg A Prieler, Dr G Schuhfried GmbH, Austria

Friday, June 14

Saxon Suite, 15.30 – 15.50 PM

L1

GAIN SCORES INTERPRETED FROM AN IRT PERSPECTIVE: A COMPARISON OF ASYMPTOTIC AND EXACT TEST STATISTICS

Presenter: Gerhard H Fischer, University of Vienna, Austria

Within the world of classical test theory, many authors discourage the use of 'gain' scores (i.e., of differences between posttest and pretest scores) as indicators of change because they often seem to suffer from low reliability. In the present paper, it is shown that this is a problem of the classical concept of reliability rather than of the gain scores themselves. First an IRT framework is defined by assuming that (a) a unidimensional pool of items is available to which a Partial Credit Model (PCM; or some of its special cases, namely, the Rating Scale Model or the Rasch Model) has been found to fit, that (b) the item \times category parameters of that PCM are known from a previous calibration study, and (c) that a pretest and a posttest are defined as any subsets of items from this pool. Then it is shown that change on the latent dimension can very conveniently be measured by estimating a certain change parameter η and that the true amount of change can be assessed statistically by computing both a confidence interval for η and significance probabilities under the null-hypothesis of no change. To that end, two techniques are considered, one employing the asymptotic normal distributions of the ML estimators of the person parameters under the PCM, the other based on the exact conditional distribution of the gain score, given the sum of the pretest and posttest scores. In either case, a detailed assessment of the precision of the change measurement results which - logically - depends on the composition of the two tests in terms of their item \times category parameters. For an illustration of this methodology, the results for three concrete test scales are shown. It is demonstrated that its practical application mainly requires the use of certain significance tables for raw score combinations, which is easy even for an empirical researcher without specific training in psychometrics. The examples moreover show that in many typical situations the gain score as such, in spite of its possibly low reliability, is not a bad indicator of change after all, except when one of the scores lies near to the boundary of the score range. Another result is that the exact conditional approach is superior to the asymptotic method, except in long tests where both methods are practically equivalent. Finally it is remarked that the described methods are also applicable to testing the null-hypothesis that two testees are equally able against the one-sided or two-sided alternative hypothesis of different abilities.

Friday, June 14

Saxon Suite, 15.50 – 16.10 PM

L2

DECISION BASED ADAPTIVE TESTING

Presenters: Lutz F Hornke & Martin Kersting, Aachen University of Technology, Germany

To rank order candidates from high to low on some psychological meaningful dimension has been the main emphasis of psychological measurement. And the latter is preoccupied with specifying, considering, and reducing errors of measurement and its determinants. However, from a utilitarian point of view the error of measurement may be less important per se unless it does not alter any personnel decision. This was convincingly voiced by the late Lee J. Cronbach and Goldine Gleser's famous book on "Personnel Decisions".

With adaptive testing it became possible to tailor the amount and the psychometric quality of items to any participant of an assessment program. Like with Binet and Simon's testing procedure there are two individualized decisions made after observing the response on an item and its subsequent evaluation: (1) Continue or stop testing, and, (2) the selection of the next best suited item. Nowadays, testing may be short or long depending on the behaviour of candidates. But it is assured that each one receives just those items which, in a psychometric sense, yield the most information about her/him. With the advent and wide use of PC's or the Web such tailored testing programs become more and more possible. The psychologist who immediately evaluated each response to an item of Binet's days is no longer necessary. Computer programs control the mock-up and the course of test taking.

However, the decisions during testing so far consider item difficulties as well as the intermediate candidate score. They lead to stop testing after some level of measurement was attained. This is nothing but an investigatory decision. Going a step further calls for a terminal decision. In some instances this is to accept or reject someone for a job, in other instances it might mean to assign candidates to different training programs. Seen from this perspective a terminal decision operates with one or several cut off points on a given dimension / scale.

A terminal decision may be made how wide the confidence interval for some intermediate score may be, as long as its upper or lower bound does not cross the prespecified cut off point. If that holds then testing may be stopped (investigatory decision) and the candidate may be assigned (terminal decision) to some arbitrary training / treatment. However, for two adjacent cut offs the candidate's confidence interval has to fit deliberately into the gap between cut offs. This means that adaptive testing must be continued as long until the prespecified risk is guaranteed and the upper and lower confidence bounds just touch but do not cross the cut offs. In general, the individual confidence interval as a function of a candidate's ability, the items taken, and the information embedded therein determines how reliable a test ought be to arrive at a terminal decision with a prespecified risk: Reliability is variable and not fixed! Then, cut offs and prespecified risks are the only fixed determinants for a psychometric based personnel decision program.

The paper will report on prerequisites and results from simulated decision programs to be used in practice.

Friday, June 14

Saxon Suite, 16.10 – 16.30 PM

L3

ISSUES AND MODELS IN COMPUTERISED TESTING

Presenter: John Raven, University of Edinburgh, UK

Probably the most important defects in the forms of assessment most widely used today are (1) their failure to provide a sufficiently comprehensive picture of the talents of an individual or the effects (both positive and negative) of an educational programme, and (2) their failure to adequately register the unique high-level motives and talents of the individual.

These deficiencies result in assessments which cannot legitimately claim to be objective, competent, or ethical. They seriously harm the lives and careers of many of those assessed and the organisations and societies in which people work.

Both defects stem from the hegemony of an inappropriate psychometric modal and failure to adopt an appropriate technology.

A way forward can perhaps be indicated by suggesting that, instead of thinking in terms of variables, we need to develop a descriptive framework to think about the variance between individuals. The kind of framework required can perhaps be indicated by drawing an analogy with biology or chemistry.

Clearly, it would be possible, using information technology to introduce the kind of branched questioning and pattern making process characteristic of a chemical analysis or classification of an unfamiliar animal or plant.

The paper will be illustrated with examples of the difficulties encountered when using traditional approaches and when trying to develop an alternative.

Friday, June 14

Saxon Suite, 16.30 – 16.50 PM

L4

COMPUTER-AIDED TESTING, NEW TRENDS AND DEVELOPMENTS

Presenter: Joerg A Prieler, Dr G Schuhfried GmbH, Austria

The advantages of computer-aided diagnosis will be presented. A short demonstration of a test system is done to show the possibility of a modern test system. The emphasis of the paper will be adaptive testing (the optimal relation of test length and measurement precision), multimedia tests (videos and animations), and the usage of the Mixed Rasch Model for test construction (here the answer vector is analyzed for person comparison, not the raw score). Furthermore, the use of peripheral equipment (the possibility to measure psychological dimensions, which are not optimally measurable with a PC-equipment, like fine motoric skills). In addition, a presentation of the program of Fischer's new method (measurement of change of individuals) will be presented.

Friday, June 14

Saxon Suite, 16.50 – 17.10 PM

L5

IMPROVEMENTS IN THE FIELD OF PERSONNEL SELECTION THROUGH NEURONAL NETWORKS

Presenter: Markus Sommer, Dr G Schuhfried GmbH, Austria

In the field of personnel selection and traffic psychology tests became more and more important to select candidates for a position, or identify individual at risk to cause accidents. In general, validation studies provide the basis for the selection of the specific tests and a rational to derive a prognosis about the testee's future performance in the domain of interest. In such validation studies significant and highly significant correlations between individual aspects of the performance in the domain of interest and the results of different tests can be found. However these correlation coefficients are often considerable low and in most cases practical irrelevant. One thus needs to combine the available information about a candidate to derive a valid diagnosis or prognosis. One can choose to use the clinical or the statistical judgment to do so. In the clinical judgment the diagnostician combines the existing data on the basis of his or her knowledge and experience. However past studies have shown that the clinical judgment features several problems such as strategic differences and interindividual differences in the weights assigned to the various information about the candidate even among experts. The diagnosis derived are thus presumable less objective. In contrast to the clinical judgment, the statistical judgment uses mathematical algorithms. Research has shown that prominent methods of this approach such as the discriminant analysis arrive at equal good or even better prognosis than the clinical judgment. However this approach is seldom used which is in part due to the vulnerability of the classic statistical procedures to violations of their assumptions and their difficulties to take patterns and non-linear relations into account. Neuronal networks on the other hand have been designed to perform such a pattern recognition and make little assumptions about the data used. In two pilot studies the author explores the potential of this method to combine data and compares this new approach with established means to integrate data such as the discriminant analysis. The results of these two studies show, that this method was able to classify 81 to 75 percent (traffic psychology) and 85 to 95 percent (aviation psychology) of the subjects correct while the discriminant analysis only classified 58 to 68 percent (traffic psychology) and 57 to 70 percent (aviation psychology) of the subjects correct. The results reported above reflect the initial estimate of the classification rate as well as the classification rate obtained in a cross validation using the "leave one out" method. This new method even proved to be well suited in the case of a violation of the assumption of the discriminant analysis like in the aviation study. Even though the sample sizes are small we can thus conclude that neuronal networks can provide an improvement in the field of personal selection.

M: **MEASURING ATTITUDES AND STYLE**
PM

15.30 AM – 17.10

Chair: Professor Emeritus Barbara Byrne, University of Ottawa, Canada

Friday, June 14

Winchester Conference Chamber, 15.30 – 15.55 PM

M1

CURRENT PRACTICE ISSUES IN RELATION TO INTERNET-DELIVERED PERSONALITY TESTS

Presenters: Iain Coyne, University of Hull, UK
 Dave Bartram, SHL Group Plc, UK
 Penny Smith-Lee Chong, University of Hull, UK

As the market for Internet-delivered computer-based testing develops, and as the technological sophistication of the products increases, so the issue of ensuring those using such assessment tools follow good practice will increase in importance. Given this, and following on from the successful design, development and formal adoption of the International Guidelines for Test Use (ITC, 2000), the ITC decided to examine the issue of developing guidelines for computer-based/Internet-based testing (CBT). Initial research on this project identified that there was a need for a more systematic survey to be undertaken of Internet-based testing sites in relation to good practice issues. This current paper outlines the work in progress on such a systematic review concentrating for the moment on personality-based Internet-delivered testing sites. Personality-based testing sites were identified from an initial search of the Web and questions were created, which directly related to good practice issues in Internet-based and computer-based testing previously identified in a review of the literature. Next, questions were incorporated into an interview schedule and a series of telephone interviews were carried out with those organisations willing to take part. The results will be analysed in relation to examining current practice and good practice issues within Internet-based personality testing.

Friday, June 14

Winchester Conference Chamber, 15.55 – 16.20 PM

M2

COMPUTER GAME-EMBEDDED TESTING: ASSESSING SOCIAL BEHAVIOUR IN A CONFLICT/CO-OPERATION SIMULATION

Presenters: Eugene Aidman & Armando Vozzo,
 Defence Science and Technology Organisation, Australia

Computer game technology offers an appealing paradigm for the development of new models for psychological testing (Case, 1995; Porter, 1995). Both off-the-shelf games and purpose-designed game-like software (e.g., Allan, 1995) have been utilised to test cognitive functions such as memory (Ryan, 1994; Washburn & Gullede, 1995), skill acquisition (Donchin, 1995) and strategy development (Gonzalez & Cathcart, 1995). Social behaviour has also become a target of game-embedded assessment, albeit recently. For example, a conflict/co-operation simulator game, Mimics (Shmelyov & Aidman, 1997) implements embedded measurement to assess social behaviour in contexts of potential interpersonal conflict. Mimics requires the player to manipulate schematic facial expressions of their Avatar in order to negotiate a maze inhabited by several hosts. The hosts' reactions to the Avatar depend on both their and Avatar's expressions and range from friendly/supportive to obstructing and even expressly aggressive. The player can choose between negotiating with a host, attacking it, or taking an escape route. Comprehensive recording of player moves and interactions enables

computation of several indices of interactive behaviour such as aggressiveness, efficiency and motivation.

Initial validation (Aidman, 2000; Aidman & Shmelyov, 2002) has demonstrated considerable utility of the software's capacity to discriminate between unprovoked attacking (aggression as an intrinsic choice), retaliatory attacking (defensive aggression), and frustration-driven attacking (lashing-out at non-attacking obstacles). The method's applications in training and assessment are discussed, including its considerable potential to capture behavioural signatures of extraversion, assertiveness and other dispositions, as well as its capacity to minimise self-presentation distortions inherent in self-report measurement.

Friday, June 14

Winchester Conference Chamber, 16.20 – 16.45 PM

M3

PASS THIS WAY! A BRIDGE BETWEEN PUPIL ATTAINMENT & ATTITUDES

Presenters: Glen Williams, Wolverhampton LEA, UK
Robert Whittome, W3 Insights Ltd, UK
Phil Watts, UK, Wolverhampton LEA, UK

Increasingly, attention is being paid to pupil attitude and its role in raising attainment and improving behaviour, particularly given the proposed new OfSTED framework, where this will be an important element in school inspection.

The Pupil Attitude to School Survey (PASS) is a powerful, multifactorial measure of pupil attitudes. Its derivation is underpinned by item and factor analyses, undertaken in collaboration with both Exeter and Birmingham Universities, resulting in nine distinct emergent factors - feelings about school, positive self-regard as a learner, negative self-regard as a learner, approach to learning situations, attitudes to teachers, attitude to work, confidence in learning, attitude to attendance and attitude to work demands.

Three versions of the survey have been developed, each independently piloted and standardised on large samples ($n > 6000$), to cover primary, secondary and FE sectors. Sampling controlled for population variations in ethnicity, socio-economic status, teacher: pupil ratios and level of learning difficulties.

Implemented as a computerised assessment tool, it is possible to quickly establish a percentile score, for an individual pupil, on all nine factors. The questionnaire can also be used with groups or as a whole school survey. It is sufficiently robust to be utilised as a method of establishing baseline positions and then evaluating the effectiveness of school or individual development activities.

Evaluation of the PASS indicates high validity and reliability, ease of administration, and transparency of results, promoting rapid, easily derivable, intervention strategies. Applications, for both teachers and practitioners, are consistent with the revised Code of Practice. Suggested foci include individual casework, systemic whole school improvement and multidisciplinary resource allocation at a service level.

Friday, June 14

Winchester Conference Chamber, 16.45 – 17.10 PM

M4

THE RELATIONSHIP BETWEEN PATTERNS OF RESPONSE LATENCIES AND SCORES IN COMPUTERIZED PERSONALITY TEST

Presenters: Yehuda Esformes & Danit Bunin, Til International Ltd, Israel

The use of computers in the administration of personality tests enables recording of the time between item presentation and the examinee's response, thus opening new opportunities for exploring the information embedded in response latencies. Previous studies tried to connect between response latencies and accuracy of responses. Hsu et al. (1989) suggested a self-schema model to account for differences in response latencies. They suggested that responding honestly would involve accessing an elaborate information about the self, while non-honest responses would involve less complex information, and therefore shorter response latencies.

The distinction between response latencies of "accepted" and "rejected" items revealed a more complex picture than the simple self-schema model. By using this categorization, Popham and Holden (1990) found negative correlations between scale scores and mean latencies for accepted items and positive correlations for rejected items. They proposed an elaborated model, suggesting that schema-relevant items are quickly endorsed and slowly rejected.

We found the same results with a computer administered self-report personality test encompassing 450 items and 31 scales grouped into 5 empirical factors. The subjects were 839 adults in The Netherlands. Item latencies were double standardized to control for both item characteristics and subject characteristics. All 31 scales yielded significant and similar results. Subjects scoring high in a given scale required relatively less time to endorse items on that scale and more time to reject items on that scale.

An alternative model of "testing of limits" is suggested to account for the results and new applications of response latencies are discussed in light of the model.

N: HIGH STAKES EXAMS
17.30 PM

15.30 –

Chair: Professor Anita Hubley, University of British Columbia, Canada

Friday, June 14

Walton Room, 15.30 – 15.50 PM

N1

A RETROSPECTIVE ON PERFORMANCE-BASED TESTING

Presenters: Robert Hunt, & Cristina Goodwin, Certiport.com, USA

As the exclusive, worldwide administrator of the Microsoft Office User Specialist (MOUS) certification exams, Certiport Inc. has spent the last four years refining a series of performance-based exams leading to certification on Microsoft Office desktop business applications. The experience of developing and deploying emulation or “live” performance-based exams to a worldwide network of testing centers has been a frustrating, time-consuming, and expensive relative to the production of knowledge-based exams. In view of this, the question begging an answer is: “is it worth the effort?”

The proposed presentation will introduce the audience to the widely perceived, but seldom studied benefits of performance-based tests; the challenges and alternatives in producing such tests; and the returns to the test candidate and the employers who hire them.

The presenters will also share some of their experiences with the issues relating to performance-based item types, including the challenges associated with the specification of testing objectives, creating the items, and using the item format itself to the greatest effect.

To date, more than 400,000 individuals have earned MOUS certification in over 100 countries and more than 30,000 exams are administered each month in Certiport’s 10,000 strong worldwide testing center network.

Friday, June 14

Walton Room, 15.50 – 16.10 PM

N2

TECHNOLOGY ADVANCES IN INNOVATIVE CERTIFICATION ITEMS

Presenter: Charles B Johnston, NCS Pearson, USA

Many certification programs have seen the benefits of computer based testing and have transitioned their existing programs to that new testing format. Slower to occur, though, is the increased utilization of the computer for presenting novel items, improved test formats, and generally using the power of the computer for more than presenting the same multiple choice items. Dr. Johnston will demonstrate several interesting new item formats and describe how they might be used to better measure your certification content. He will also share his real-life experiences and provide guidelines to assist you in determining whether and how to apply this new technology to your program. Attendees will leave the session with an enhanced understanding of the real technological possibilities for CBT certification items and how the new item formats can enhance their program's measurement. Attendees will also have an opportunity to have their item-related questions answered by an experienced practitioner.

Friday, June 14

Walton Room, 16.10 – 16.30 PM

N3

**TECHNOLOGY AND TEST ITEMS IN COMPUTER-BASED ASSESSMENT:
ENHANCING MEASUREMENT THROUGH FORMAT AND DESIGN**

Presenter: April L Zenisky, University of Massachusetts, at Amherst, USA

In computer-based testing (CBT), many technologies are integrated at once to define the format of the test items. First, there are many ways in which the item stem can be structured with regard to the nature of the task and how it is presented to examinees. Secondly, CBT assessments can be designed so that examinees have access to online reference materials such as calculators and spell-checkers. Lastly, technology also impacts the ways in which examinees can respond to different items, such that examinees can complete different cognitive tasks such as selecting the correct response, reordering data, manipulating onscreen information, and creating an original answer in a number of interesting ways that can be differently implemented to ensure that the test is appropriately measuring the construct of interest. Furthermore, within this framework of how assessment tasks are put together, there may be other innovative prospects for computer-based measurement to further improve the amount and quality of measurement information that is gained through the process of administering test items to examinees as computers become more powerful and programmers more adept. The purpose of this poster presentation is to provide an overview of recent developments in assessment tasks with particular emphasis on the current state of research on emerging item types and to identify several areas for additional research.

Friday, June 14

Walton Room, 16.30 – 16.50 PM

N4

**USING INNOVATIVE ITEM TYPES TO IMPROVE THE VALIDITY OF CERTIFICATION
TESTS**

Presenter: David Foster, Galton Technologies, USA

This session talks about the innovative item types made available today by computerized testing technologies. It provides examples of how these item types can be used to measure performance objectives better than the standard multiple-choice format. Data on the effectiveness of these item types will be presented.

Friday, June 14

Walton Room, 16.50 – 17.10 PM

N5

MAINTAINING SECURITY IN COMPUTERIZED ASSESSMENTS: PRACTICAL CONCERNS ADDRESSING MULTIPLE NEEDS

Presenter: Jon S Twing, NCS Pearson, USA

With the proliferation of computerized testing, particularly when such testing occurs via the Internet, there are a number of psychometric challenges that must be addressed in order to ensure the security of those assessments. Although technological issues (related to software and connectivity) present unique security challenges in on-line assessment, the psychometric challenges of maintaining security for computerized tests is truly multifaceted. This presentation will focus on the types of security concerns that are expressed by users of large-scale assessments and how those concerns are being addressed in many different ways in operational on-line assessments. Psychometric challenges such as item exposure control methods, issues associated with item pool management, and issues associated with properly collecting data on newly developed test items are among some of the issues that must be addressed. Security issues will be discussed in the context of on-line versions of high-stakes educational testing programs in the United States.

O: **PRACTICAL ASPECTS OF CBT**
17.10PM

15.30 –

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

Friday, June 14

Wintonian Room, 15.30 – 15.55 PM

O1

THE HOGREFE TESTSYSTEM: TOWARDS AN INTEGRATION OF TEST DEVELOPMENT, ADMINISTRATION AND DECISION MAKING

Presenters: Klaus-Dieter Haensgen, Jérôme Frossard, Sebastien Simonet,
 Katharina Stress & Ralf Zumbunn, Zentrum für Testentwicklung,
 University of Fribourg/Switzerland

The benefit of computer based diagnostics will at least depend on its utility - how diagnostic information can really help to support decisions (of several kinds). Developers of diagnostic systems have to take into account this more pragmatic aspect as well as the scientific criteria. But the expectancies of "consumers" in German speaking countries seem to be more influenced by this pragmatism - scientific developers have to argue against misuses of tests with computer (often by insufficiently qualified persons) as well as to make it better! Some examples were given to show the gap to existing "promises" in non-scientific approaches, which never will be fulfilled.

During the last 5 years we have developed an integrative system for test development, test administration and - beginning - for providing help in decision making (Hogrefe TestSystem). The "TestFactory" is an author ware for test declaration. Due to standard tools, there is a similar technology and ergonomy for about 100 tests. Tests were compiled (using HTML) and the "TestSystem" is able to administrate the tests with high quality. Procedures for reporting, ranking and profile comparison will improve the utility for decision making. We will show some examples.

Due to standards the test author is able to "declare" (not necessary to "program") the test and result presentation in "clear text". Starting as a developing tool at first, we have now the experience, that quality management of test development itself can be improved. Authors have to declare samples, sample sizes, parameters, norms etc. in a standardized way and so they are "forced" to give all information (and to calculate, if forgotten).

According to our approach, it is necessary to have more impressive criteria for "customers" of diagnostics - not to see only the surfaces but also the "well-known criteria" for good diagnostics. We will make some proposals for technical quality standards.

More information: <http://www.unifr.ch/ztd/HTS>

Friday, June 14

Wintonian Room, 15.55 – 16.20 PM

O2

HUMAN COMPUTER INTERFACING FACTORS IN CBT: IMPLICATIONS FOR FAIR TESTING PRACTICES

Presenter: Professor Cheryl Foxcroft, University of Port Elizabeth, South Africa

CBT can potentially introduce construct irrelevant difficulties that can impact negatively on test performance. This paper will summarise research related to the impact of human computer interfacing (HCI) factors on computer-based test performance, especially for test-takers with differing levels of technological sophistication. Data will also be presented from a case study in which the impact of differential levels of computer familiarity, culture and gender on computer-based test performance will be presented. The findings of both the literature survey and the case study raise critical issues related to fair assessment practices in CBT, which will be discussed. Among these issues are how to adequately prepare test-takers for CBT, especially those with low levels of computer familiarity, the need for user-sensitive interfaces, and how to use test scores in a fair way by taking HCI factors into account.

Friday, June 14

Wintonian Room, 16.20 – 16.45 PM

O3

CHILDREN'S PERCEPTIONS OF COMPUTER-BASED ASSESSMENTS

Presenter: Mary Richardson, AQA, UK

Research on children's' views of a new computer based tests for gifted and talented pupils was carried out to investigate the children's' perceptions of the tasks. Computer-based assessment for gifted and talented children is a new venture in the UK. The current government has recently launched a project designed to support the education of gifted students including 'World Class Tests' in mathematics and problem solving. It was found that students enjoyed the computer-based tests and that they particularly liked the use of colour, graphics and interactivity. The vast majority of participants said that they would prefer to take tests on computer rather than on paper. Students gave a variety of justifications for this preference and were able to comment on the possible disadvantages of computerised testing. Some potential implications of the computerised presentation format of the tests were noted.

Friday, June 14

Wintonian Room, 16.45 – 17.10 PM

O4

COMPUTER-BASED TESTING - THE CANDIDATE PERSPECTIVE

Presenters: Sarah Heywood & Niall Leavy
Office of the Civil Service and Local Appointments Commissioners, Ireland

This paper outlines the main findings of a number of recent studies carried out in advance of the introduction of electronic assessment to support large-scale recruitment in the Irish Civil Service. The focus of this paper is on candidates' perceptions of a multi-media approach to test administration, responses to different types of test items across different test batteries and formats. Participants came from a wide range of age groups, educational backgrounds and experience with computers.

Data is available on participants who undertook the tests on both computer-based and paper and pencil formats. As well as gathering information on views and experiences, it also enabled us to explore the equivalence of computer and paper and pencil-based tests. Participants' perceptions were collected through questionnaire and focus group discussions.

Issues that were addressed included, inter alia, format of pre-test familiarisation material (electronic versus paper-based), attitudes towards the on-screen instructions (e.g. participants preferred the instructions read out rather than working at their own pace using the screen instructions), preference of large group vs small group testing environment and preference of testing method. The studies also show consistent concern over practical issues in the testing environment. Data is also available revealing that electronic testing options are not meeting candidate expectations of what technology could provide.

The implications of the findings for test designers are presented along with issues relating to maximising candidate performance and managing testing in a more cost-effective way through computer-based testing.

Friday, June 14

Keats Room, 13.00 – 17.30 PM

P11

ON CUT-SCORES FOR THE MODIFIED CAUTION INDEX

Presenters: E Doval & M C Viladrich, Universidad Autònoma de Barcelona, Spain
J Renom, Universidad de Barcelona, Spain
E Torres, Universidad del País Vasco, Spain

Modified Caution Index (Harnisch & Linn, 1981) is one of the most popular group-based person-fit statistics used when evaluating the validity of response patterns obtained in ability tests. Some characteristics of MCI are appealing, i.e. MCI values lie between 0 (perfect fit) and 1 (maximum misfit), its performance has been widely studied, and it is useful in a wide range of applied evaluation situations. MCI values depend on the ability level of examinees less than other group-based indexes and its detection rates seem to be excellent for extreme cases.

For practical uses, it is necessary to determine an optimal cut-score in order to identify aberrant response patterns. Harnisch & Linn suggested a cut-score of 0.3 even though based on only one empirical dataset. Meijer, Muijtens & Van der Vleuten (1996) compared the relative detection rate of 3 of cut-scores based on a series of simulation studies.

The aim of our work is to provide further evidence about appropriate cut-scores when evaluating person-fit in pass/fail tests, with emphasis on educational testing.

Samples of normal and aberrant response patterns were simulated, considering as influential factors the type and severity of aberrance, the test length, item difficulties spread, item discrimination, and the normal sample homogeneity. After obtaining MCI for every normal and aberrant pattern, ROC curves of detection rates were obtained comparing aberrant versus normal samples at selected cut-scores. Suggestions for practical use were derived.

Friday, June 14

Keats Room, 13.00 – 17.30 PM

P12

X-PAT USES IN APPLIED SETTINGS

Presenters: E Doval, Universidad Autònoma de Barcelona, Spain
J Renom, Universidad de Barcelona Spain
M I Núñez & A Solanas, Universidad de Barcelona Spain

The purpose of this paper is to present several uses of X-PAT in exploring and detecting unlikely response patterns in test or exam databases. X-PAT is a Windows computer program developed from Universidad de Barcelona and Universidad Autònoma de Barcelona for exploratory response pattern analysis in absence of IRT conditions. After importing a database, X-PAT offers users two types of analysis:

1. Aberrant Response Patterns (ARP): to identify individual responses in a test that doesn't fit with a predetermined model.
2. Similar Error Patterns (SEP): it detects individual error patterns almost identical between pairs of examinees.

Usually X-PAT results add new information that allows users to understand examinee's answering behaviour and possible presence of test misconduct (cheating, guessing, etc.). In a test construction process, X-PAT's results will also serve to detect subgroups of suspicious response patterns that must be removed from the database in order to improve psychometrical item analysis. A small number of ARP's can disrupt item parameter calibration (conventional or IRT). In ARP analysis, X-PAT output produces six group-based person fit statistics for every examinee distributed in three types according Harnisch and Linn's proposal (1981). ARP statistics are: Norm-Conformity Index (Tatsuoka and Tatsuoka, 1980), U index (Van der Flier, 1977), Loevinger Errors (EL), Caution Index (Sato, 1975), Modified Caution Index (Harnish and Linn, 1981) and Point Personal Biserical Correlation. For SEP analysis, X-PAT computes a squared matrix based in Bellezza and Bellezza's (1989) statistics that compares all possible pairs of examinee patterns. Several graphics complete this output.

Friday, June 14

Keats Room, 13.00 – 17.30 PM

P13

BLOC-INFO: SOFTWARE FOR BLOC-C AND BLOC-S SPANISH LANGUAGE TESTS.

Presenters: J Renom, Universidad de Barcelona, Spain
S Rodriguez, A Solanas, Universidad de Barcelona, Spain
M Puyuelo, M, Universidad de Zaragoza
E Wiig, Boston University / Knowledge Institute, USA

BLOC-INFO is a pioneer software project in Spain. This program allows one to administrate and correct by computer both versions of the most complete language competence test for children developed at the moment in Spain (BLOC-C and the new form BLOC-Screening). BLOC-INFO facilitates test correction tasks and produces several examinee profiles for both tests. These reports may combine and integrate, in the same graphic environment, complete or partial administrations of both versions. Another useful possibility is to integrate in the same graph-report several longitudinal administrations for every examinee. In this way, it should be considered that the amount of data used in BLOC-C can be large. Every complete administration yields 580 item responses and 62 scores, and every screening version yields 150 responses and 4 scores. All this information must be added to personal and clinical variables. BLOC-INFO allows the user to create an examinee's assessment file. The main component is the examinee profile with his/her current and past scores, response patterns and personal descriptive data. Automatically, all this information may be coded and sent by Internet to a central processing station in order to construct a database with all examinee response patterns. The target is to distribute, using personal and clinical variables, these data patterns into several smaller databases more representative of examinee's typologies. This new information will permit periodical re-estimating of BLOC-C and BLOC-Screening psychometrical properties and perform differential norms for every typology. The last step of this updating sequence is to "recharge" technical data of BLOC-C and BLOC-S in order to improve its assessment properties.

Friday, June 14

Keats Room, 13.00 – 17.30 PM

P14

SELF-ASSESSMENT FOR BRAZILIAN PORTUGUESE, AND ITS RELATIONSHIP TO ADAPTIVE TESTING

Presenters: Elena C Papanastasiou & Antonio R M Simoes, The University of Kansas, USA

Self-assessment is an area of testing that has not been adequately researched because of the subjective nature of the tests. However, in non-competitive situations, self-assessments could be valuable tools in the area of language testing and in the area of adaptive testing.

This study is based on the results obtained from a subjective test of Brazilian Portuguese (BP) that is currently intended to be used as a supplement for making placement decisions for students who will be attending a foreign language program. The test is available on the internet at www.ukans.edu/~brasilis.

Based on the subjective test, the students will be placed in one of four levels of BP that range from basic to superior. This classification will be used as an initial ability estimate for an objective adaptive test that will be administered to the students after they have completed the subjective test. Therefore, the students who were placed in the basic category on the subjective test, will be assigned an initial ability estimate of -2.0 for the objective adaptive test. The students placed in the intermediate category from the subjective test will obtain an initial ability estimate of -1.0 . Students in the advanced and superior categories will obtain initial ability estimates of 1.0 and 2.0 for the adaptive test.

The assignment of initial ability estimates can be very beneficial for small scale adaptive tests. First, the items of average difficulty will not be overexposed to the students by continuously administering them at the beginning of the test. Second, the estimation of the ability estimates will be more accurate since there will be more items that will be targeted directly to an examinee's ability. Finally, the standard error of the ability estimates will be smaller, while the information obtained by the test would be larger. This paper will present the advantages and effects of using self-assessments to assist psychometricians in the adaptive testing process.

Friday, June 14

Keats Room, 13.00 – 17.30 PM

P15

**AACHEN TEST BATTERY OF VOCATIONAL APTITUDE OF THE DEAF
(AACHENER TEST SYSTEM ZUR BERUFSEIGNUNG VON GEHÖRLOSEN, ATBG)**

Presenters: W Iversen, F Kramer, S Lintz, K Grote, U Louis-Nouvertne, H Sieprath, I Werth,
U Zelle & L Jäger, Institute of German Studies Language, Media and Communication, Department of Neurology, Technical University Aachen, Germany
W Huber, Section Neurolinguistics, Technical University Aachen, Germany
K Willmes, Section Neuropsychology Technical University Aachen, Germany

The central aim of the ATBG-project is to improve the profession finding process for the Deaf, supporting the Deaf in making an independent career choice. The project focuses on an integration of sign language as the natural language of the Deaf.

For the Deaf, it is almost impossible to acquire full competence in spoken language. Due to their unusual circumstances of language acquisition and education, most of them also have lower levels of competence in written language.

Vocational aptitude tests require reading skills or spoken language competences. Even nonverbal tests are usually administered in written language.

A multi-media computer-based test system has been developed to measure deaf persons' most important job-oriented skills and abilities as well as vocationally relevant personality features independent of the level of reading skills and spoken language. In all 26 tests of the ATBG, all pieces of information can be called up in written language, German Sign Language or Signs Supporting German. The test person decides which linguistic form he/she prefers. The diagnosis of vocational aptitude can thus be improved substantially with respect to objectivity and validity.

Test results from a large validation study in young adult signers (n=652) indicate comparable levels of performance in ability tests between deaf and hearing subjects but the results of hearing subjects are much better in skill tests. If one assumes that the level of cognitive abilities attainable is also related to skill achievement, poorer performance of the deaf people must be largely attributed to the particular socialization of deaf people.