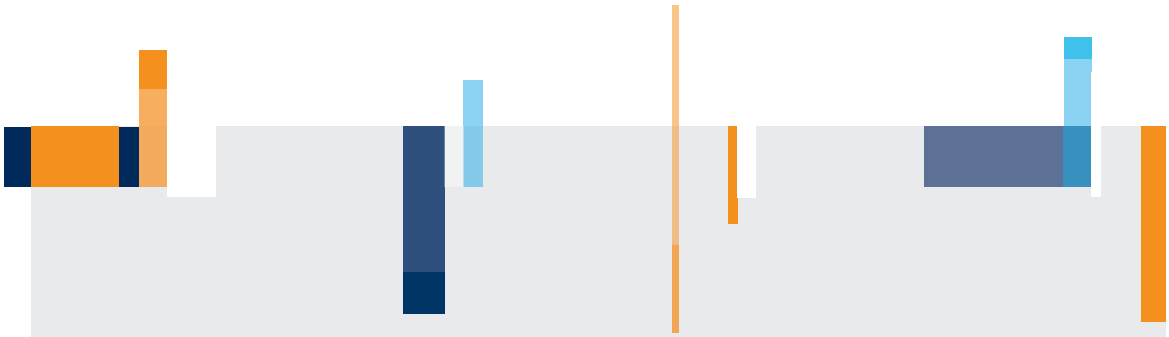
Contents



aea
EUROPE

The 7th Annual AEA-Europe Conference

assessmentandequity

9/11 November 2006
Sala Partenope, Via Partenope 36, Naples, Italy

Introduction

The Association for Educational Assessment – Europe arranges a conference each year to further its aims of fostering communication in the area of educational assessment. The six previous conferences have been held in Prague (2000), Krakow (2001), Frascati (2002), Lyon (2003), Budapest (2004) and Dublin (2005). It is with great pleasure that the AEA-Europe Executive Council welcomes you all to the magnificent surroundings of the city of Naples and the historic accommodation kindly provided by the University of Naples for its seventh conference. We are sure that you will find this a relaxing and stimulating environment for the event.

The conference theme this year is "Assessment and Equity". We have chosen this theme because we believe that as professionals we bear a responsibility for the use of the instruments we produce and study, and should understand how their use may produce effects that are adverse to our intentions. Equity concerns have led to increased attention to possible sources of bias, greater use of adaptations and accommodations, and using value-added approaches rather than league-tables to judge institutions. The selection of keynote speakers and their topics has been undertaken to illuminate and promote this discussion. The Discussion Groups also pick up this theme, as well as others, to continue debate and encourage collaboration and integration.

A new program component at this conference is the 'Open Paper Presentation' on Thursday and Friday. We have introduced this with the aim of giving members more opportunities to present and share ideas. We received many reactions in response to the call for proposals which suggests that this opportunity has been much appreciated. We also believe that the final result adds much to the richness and diversity of the program.

AEA-Europe is grateful to BENECON (Centro Regionale di Competenza per i Beni Culturali Ecologia Economia) for its generous support of this conference, to the University of Naples 'Federico II' for providing the conference rooms and to the Centro Museale delle Scienze Naturali for hosting the Saturday lunch. We also would like to thank Stephen Austin Ltd for providing conference folders, DRS Data and Research Services PLC for providing conference bags, Cambridge Assessment for providing floral decoration, Medias for providing the graphic design of this programme and banners and the Center for Educational Technology CET for providing additional textile bags.

Programme

WednesdayNovember8

9.30-16.30 Pre-Conference Professional Development Workshops

ThursdayNovember9

8.00-9.15 Registration and coffee

9.15-9.30 Opening
Emma Nardi and Carmine Gambardella, Director Benecon

10.00-11.00 Keynote session: Equity and the Role of Assessment in Italian Educational Policy
Luigi Berlinguer (Italy)

11.00-12.00 Keynote session: Assessment in the Service of Education Policy: Paradox and Potential
Henry Braun (USA)

12.00-13.30 Lunch

13.30-15.30 Open papers sessions A and B

Session A
Chair: Gordon Stobard

Equity Issues in Different Contexts

1 Educational Equity Balance in Finland; Jakko Hautamäki and Sirkku Kaupiainen et al (Finland)
2 Learning assessment: the role of 'assessment for learning' in young people's learning careers in the UK; Kathryn Ecclestone and Roger Murphy (UK)
3 Evaluation and Equity in Local Education systems; Tali Freund (Israel)
4 Automated Predictive Systems in the Prediction of Educational Outcomes; Eduardo Cascallar (Belgium) and Tracy Costigan (US)

Session B
Chair: Jan Wiegers

Part 1: University Entrance

1 A New Model of Admission Examinations for Universities in Georgia; Maia Miminoshvili and Iamze Kutaladze (Georgia)
2 Equity in Unversity Entrance Examinations; Grace Grima and Frank Ventura (Malta)

Part 2: Uses of the Common European Framework of Reference

3 The Common European Frame-work of Reference: A tool for value-added assessment? Therese Hopfenbeck and Elisabeth Ibsen (Norway)
4 Valid comparison of learner groups in a multilingual proficiency framework; Neil Jones (UK)

Coffee break during discussion groups

15.30-17.30 Discussion groups *1-4*

1 Partnership with parents in assessment. Examples from Scottish nurseries and schools
Carolyn Hutchinson, Louise Hayward and Norman Emerson (Scotland)

2 The Common European Framework of Reference for Languages: transparency and equity
Jenny Bradshaw (UK), Jenny Dalalakis (Belgium), Eli Moe (Norway) & Saskia Sluiter (Netherlands)

3 Evidence or consequence? Validity issues in educational measurement.
Simon Wolming & Christina Wikström (Sweden)

4 Professional Development (continued)
Report on progress made by the Professional Development Subcommittee and discussion of proposals for an AEA-E accreditation scheme.
Chris Whetton (UK), Gabriella Agrusti (Italy), Frans Kleintjes (Netherlands), Kiril Bankov (Bulgaria), Andrew Watts (UK), Marian Sainsbury (UK) Alastair Pollitt (UK) & Eduardo Cascallar (Belgium)

FRIDAY NOVEMBER 10

9.00–10.00 Keynote session: Ethical and Equity Issues in Assessment
Kari Smith (Norway)

Coffee break during discussion groups

10.00–12.00 Discussion groups *5-8*

5 Towards more valid and equitable systems of summative assessment
Wynne Harlen, Gordon Stobart, Richard Daugherty, Judy Sebba, Paul Black, John Gardner and Carolyn Hutchinson (UK)

6 Curriculum Design and Evaluation
Charles Briffa, Josette Farrugia and Martin Musumeci (Malta)

7 Selecting reliable markers – some studies in UK public examinations
Michelle Meadows and Jo-Anne Baird (UK)

8 Establishing European Quality Standards for Examinations and Assessments (continued)
Frans Kleintjes, Anton Béguin, Piet Sanders (Netherlands), Eduardo Cascallar (Belgium) and Gerben van Lent (Netherlands)

12.00–14.00 Lunch (buffet)

12.00–14.00 Poster presentations *a-i*

a Incorporating test users in the test development process – a case study
Lise-lotte Appelgren (Sweden)

b A Computer-assisted Test Design and Diagnosis System for Classroom Teachers
Qingping He (UK)

c Classroom materials to support Assessment for Learning through Literacy
Juliet Sizmur and Claire Hodgson (UK)

d An Evaluation System for Computer-Based Tests
Piet Sanders (Netherlands)

e The Predictive Validity of the SweSAT
Per-Erik Lyrén (Sweden)

f Presenting a picture – teacher assessment in Wales.
Hilary Cox, Juliet Sizmur, Claire Hodgson and Bethan Burge (UK)

g Measuring concepts and factual principles compared to measuring understanding and application
Lina Wahlgren, Ingemar Wedman and Sara Franke-Wikberg (Sweden)

h Developing electronic marking and management solutions
Graham Hudson and Brian Carbarns (UK)

14.00–15.00 Keynote session: Gender-assessment interaction: repositioning the focus of attention in equity studies
Patricia Murphy (UK)

15.00–15.30 Coffee break

15.30–17.00 Open Papers sessions *C* and *D*

Session C
Chair: Kathleen Tattersall

Approaches to Improving Examinations

1 Developing and piloting a methodology to enable students to participate in test development
Alison Wood and Alastair Pollitt (UK)

2 Assessment of writing proficiency
Ragnar Thygesen and Rolf Fasting (Norway)

3 A taxonomy of Sources of Difficulty in the assessment of ICT
John Threlfall and Nick Nelson (UK)

Session D
Chair: Grace Grima

Aspects of Science Assessment

1 PISA's Computer-based Assessment of Science (CBAS)-A gender equity perspective
Are Turmo and Svein Lie (Norway)

2 An evaluation of bookmarking as a judgemental equating method in KS2 science
Catharine Parkes (UK)

3 Towards gender inclusive assessment in science
Deborah Chetcuti (Malta)

17.00–18.00 AEA-Europe business meeting

Evening CONFERENCE DINNER
La Bersagliera (across Castel dell'Ovo, Borgo Marinari n. 10/11)

20.00–20.30 Reception

20.30–23.00 Conference Dinner

saturday NOVEMBER 11

9.00– 10.00 Keynote session: Using data from international comparative studies to investigate the equity issue: evidence from PISA results
Bruno Losito (Italy)

10.00–11.45 Discussion groups *9-12*

9 Value added measures and the enhancement of equity in assessment
Harvey Goldstein (UK)

10 Guarantees for equity of national exams?
Jo-Anne Baird (UK), Anton Béguin (Netherlands), Theo Eggen (Netherlands) and Claire Whitehouse (UK)

11 Development of a new model of national standards-based educational and assessment system: the path to quality and equity
Eduardo C. Cascallar (Belgium) and Michael Fast (UK)

12 Gender in Interaction with Assessment: Implications for how we think about gender equity in assessment
Jannette Elwood and Patricia Murphy (UK)

10.30–11.00 Coffee break during discussion groups

11.45–12.45 Keynote session: The Role of International Survey Comparisons in Accountability
Michal Beller (Israel)

12.45–13.30 Concluding remarks and farewell

14.00–15.30 Lunch
Real Museo Mineralogico (via Mezzocannone n. 8 - AEA-Europe bus)

Afternoon *Opportunities for sightseeing in Naples*

Keynotes
Abstracts and Biographies

Thursday

10.00–11.00 Equity and the Role of Assessment in Italian Educational Policy

Luigi Berlinguer, University of Siena, Italy

Political decision making in the field of education is always difficult, because the results of a reform bill necessarily can't be evident for a long time. Therefore equity can only be tackled through to a deep reflection on assessing data collected through strict national and international procedures. This presentation will introduce the problem of equity at national level from the Italian standpoint, in connection with the civil and economic needs of a developed country in the course of a real expansion of education and training for all citizens.

Luigi Berlinguer

Luigi Berlinguer is a full professor at the Faculty of Law of University of Siena, where he also acted as Rector from 1985 to 1994. He has always been engaged in politics. From 1963 to 1969 and from 1994 to 2000, he was a member of the House of Commons. From 1991 till the present he has been a member of the Board of the Left Democratic Party. From 1996 to 2000 he was Minister for Education and Research. During his mandate, he strongly supported the Italian involvement in international comparative surveys (OCDE-Pisa, OCDE-Sials; IEA-Civic Education) and launched the Italian National Institute for Evaluation of School System (INVALSI).

11.00–12.00 Assessment in the Service of Education Policy: Paradox and Potential

Henry Braun, ETS, USA

Education policies have four fundamental goals: Entrance, Excellence, Equity and Efficiency. In principle, assessment can play a constructive role in promoting each of these goals but, paradoxically, too many implementations result in negative, unintended consequences. This presentation will discuss the reasons for this paradox and describe some of the steps that can be taken at both the policy and technical levels to enhance the positive contributions assessment makes to educational improvement.

Henry Braun

Dr. Henry Braun, a Distinguished Presidential Appointee at ETS, is a noted statistician and psychometrician. He served as vice-president for research management at ETS from 1990-1999. He is a co-winner of the Palmer O. Johnson award from AERA (1986) and of the NCME award for Outstanding Technical Contributions to the Field of Educational Measurement (1991). Dr. Braun's current interests include the interplay of technology and assessment, the analysis of large-scale assessment data and education policy.

Friday

9.00–10.00 Ethical and Equity Issues in Assessment

Kari Smith, Department of Education and Health Promotion, University of Bergen, Norway

This presentation will discuss two studies focusing on ethics and equity in assessment. The first study addresses current assessment practices in light of selected paragraphs in the UN Declaration of the Rights of the Child (1989). The study examines if teachers' assessment practices are in accordance with the rights expressed in §3,

§12 and §13, serving the child's best interest, the child's right to express her/his view freely in all matters affecting the child, the right to freedom of modes of expression, and to seek, receive and impart information and ideas through any media of the child's choice. Where these rights are not observed, some children will be disadvantaged and we face a problem of equity.

The second part of the presentation will discuss the problem of equity in relation to university students whose first language differs from the language of instruction. Most institutions of higher education have students from multiple language and cultural backgrounds. The number of students studying in a language other than their first language is on a steady increase in Europe and elsewhere.

This study addresses the problem of equity related to assessment of these students and suggests ways in which universities can meet the challenge of seeking a balanced assessment which reduces the many difficulties foreign students meet when they enroll in universities outside their own country. There are different reasons for engaging in higher education in an additional language, and the paper recommends universities and teachers to adjust their programmes and assessment activities according to the specific needs of various groups of students.

Kari Smith

Kari Smith is professor at the department of Education and Health Promotion, University of Bergen. Her main research interests focus on evaluation and assessment for and of learning, teacher education, and professional development.

She worked as a teacher in a kibbutz school in Israel for 18 years, has worked as a teacher educator at Oranim Academic College of Education in Israel where she acted as the Dean of the Teacher Education Department for secondary school from 1995-2003. She has been and still is a counsellor to the National Institute of In-service Education for Teacher Educators (Mofet) in Israel. She is appointed as a professor by the Council of Higher Education in Israel, and was given the position of Professor with the University of Bergen, Education Department where she currently works.

Prof. Smith is active in EARLI (The European Association for Research in Learning and Instruction) where she has acted as the Coordinator for the Special Interest Group for Assessment and Evaluation for the last four years. Prof. Smith has also served as the Coordinator of the Testing, Evaluation and Assessment Special Interest Group of IATEFL (International Association for Teachers of English as a Foreign Language) for eight years till 2003, and she served on the Management Committee for that association till spring 2005.

Prof. Smith has published widely in international refereed journals and has acted as reviewer to a number of academic journals. A recent publication of her is a book written jointly with Harm Tillema from Leiden University; Portfolios in Professional Development, A research Journey.

14.00–15.00 Gender–assessment interaction: repositioning the focus of attention in equity studies

Patricia Murphy, Open University, England

Studies of equity in assessment have typically relied on assessment outcomes to explore the sources and nature of differences between sub groups. Concern with difference has been predicated on assumptions that statistical similarities in scores, or profiles of scores, provide evidence of an equitable instrument and process. This assumption continues to dominate the discourse about differences between boys' and girls' achievements (Murphy and Ivinson 2004). Equity studies have therefore been constrained both by the data available and by representations of assessment and of gender. The paper examines how developing understandings of learning challenged the way in which assessment was represented and the consequences for how equity was researched and findings interpreted (Gipps and Murphy 1994; Murphy 1995; 2000). These findings suggested ways in which gender mediated learning but did not illuminate the process of mediation. There continued to be a divide between research into assessment as a social process and research into equity and assessment and progress in the former has not been reflected in the latter. The paper argues that this divide can be understood in two ways. First in terms of the way that 'social' is understood in the model of learning typically applied in assessment research and second the focus in

socio-cultural analyses on assessment acts rather than the processes by which representations and practices of assessment mediate individual knowledge construction. This maintains the focus on outcomes and the consideration of gender equity issues to sex groups. The paper concludes by looking at case study research that offers an alternative socio-cultural perspective on gender and assessment that makes visible the social process of knowledge construction. Educational achievement is seen as a product of the interrelation between gender and assessment as social forces operating within classrooms that mediate how teachers represent valued knowledge outcomes and their appropriation by students. To extend our understanding of equity in educational assessment the paper argues that we have not only to make problematic the assessment process but also the teaching and learning process wherein what is made available to learn and what it is possible to achieve, and by whom, is realised.

Patricia Murphy

Patricia Murphy is Professor in Education (Pedagogy and Assessment) in the Faculty of Education and Language Studies at the Open University, England and director of the research group Pedagogy, Learning and Curriculum in the Centre for Research in Education and Educational Technology. She writes postgraduate courses and supervises in the areas of curriculum, learning and assessment generally and science and technology specifically. Patricia has a background in research in assessment and effective teaching and learning particularly in relation to equity and gender and has published widely in these areas. She was deputy director of the national Assessment of Performance science unit; national expert for the Programme of International Student Achievement (PISA); reviews the national science tests for 13-14 year olds; is a member of the national gender policy group for the Qualifications and Curriculum Authority and advises on assessment internationally.

saturday

9.00–10.00 **Using data from international comparative studies to investigate the equity issue: evidence from PISA results**

Bruno Losito, Roma Tre University, Italy

Data collected in international comparative studies can be used to analyse some of the features of the education systems in the light of the equity issue. In particular, data collected through student tests and questionnaires and school questionnaires can be used in order to investigate the effect of some variables that can be considered directly or indirectly related to the equity issue.

In particular, the PISA studies enable us to deal with these issues not just in terms of international comparisons, but also from a diachronic standpoint for the individual countries concerned, in this way offering indications on the effects and effectiveness of policies adopted at national level.

In this perspective, the PISA 2000 and 2003 results will be used to illustrate to what extent equity actually is (or is not) a characteristic of the Italian education system.

A comparison will also be made between the PISA results and those of the OECD-ALL (Adult Literacy and Life Skills) study, which provides information and data on the cultural levels of the Italian population. This will be the basis for outlining analysis and research lines aiming to integrate – in terms of equity – the data emerging from studies on school systems and those on adult population literacy levels.

Bruno Losito

Bruno Losito is Associate Professor in Pedagogy at the Roma Tre University. Prior to his present position he worked as researcher at the Italian National Institute for the Evaluation of the Education System (INVALSI).

He participated in several international research projects and in evaluation studies.

He worked in the IEA Civic Education Study as Italian National Co-ordinator (phase 1 and 2). As a member of the international Steering committee (phase 2), he was responsible for designing the teacher questionnaire and for the analysis of the teacher data.

He worked with the Council of Europe within the framework of the Education for Democratic Citizenship Project

and is one of the writers of the "All-European Study on Education for Democratic Citizenship Policies". He also worked with UNESCO for the evaluation of the projects "Intercultural and Human Rights in Albania" (2003) and "Education for Democratic Citizenship: From Policy to Effective Practice through Quality Assurance" (2004-2005). He is currently the Italian National Project Manager for PISA (Project for International Student Assessment).

11.45–12.45 The Role of International Survey Comparisons in Accountability

Michal Beller, Director-General of the Israeli National Authority for Measurement and Evaluation in Education (RAMA), Israel

Evaluation plays a major role in initiating, implementing and monitoring educational reforms. Local media and public attention is constantly drawn to the state of education when achievement on international assessments (such as TIMSS and PISA) falls below expectations. This often results in calls for educational reforms.
In Israel, a special national task force – the Dovrat Committee – was set in 2004, after the release of the PISA 2003 results. The committee's charter was to propose guidelines for a comprehensive educational reform (structural, organizational and pedagogical) aimed at: (a) narrowing gaps between socio-economic groups; (b) upgrading teachers' professionalism and status; (c) enhancing schools autonomy and authority; (d) augmenting accountability and transparency. One of the recommendations, establishing a National Authority for Measurement and Evaluation in Education (RAMA), is currently underway, while other recommendations are still publicly debated and negotiated with the teacher unions.
The presentation will focus on the mission, vision and actions of RAMA in setting an accountability system, and assisting the ministry of education in meeting some of the reform goals. Opportunities of constructing an accountability system where assessment plays a dual role - as an "assessment of learning" and "assessment for learning" - will be discussed. Special attention will be devoted to issues of equity and to conditions under which accountability policies can foster a more just educational system. National challenges, such as the absorption of immigrants from all over the world and the advancement of the Arab sector, will provide some background for the discussion.

Michal Beller

Michal Beller is the director-general of the Israeli National Authority for Measurement and Evaluation in Education (RAMA). She assumed this position in 2005 upon her return to Israel from ETS in Princeton, where she was a Senior Research Director, with responsibility for the Assessment & Evaluation Research Cluster. Her research interests include assessment and measurement in K-16, admissions to higher education, validity research, test adaptation to different languages, test bias issues, program evaluation, and e-Learning.
Beller received her doctorate in psychology from the Hebrew University. From 1995 until joining ETS in 2001, she was associate professor, Department of Education and Psychology, at the Open University of Israel. From 1987 to 1995, she was the director-general of the Israeli National Institute for Testing and Evaluation (NITE), which administers university admissions tests, K-12 assessments, and conducts psychometric research.
Michal has been constantly active in educational testing both in Israel and internationally, having served on various committees amongst them are the International Advisory Board for the SweSAT, the Council for the International Test Commission, and the Executive Committee of the International Association for Educational Assessment.

Discussion Groups Abstracts

Thursday

15.30–17.30

1. Partnership with Parents in Assessment. Examples from Scottish nurseries and schools

Carolyn Hutchinson (QAC, Scotland), Louise Hayward (University of Glasgow, Scotland) and Norman Emerson (Learning and Teaching Scotland)

The term "partnership between schools and parents" is commonly used in both policy and research. In Scotland, the new Parental Involvement Bill is part of a commitment to strengthen the link between parents and schools. In practice, very different models of partnership between home and school exist in Scottish schools. Some schools work collaboratively with parents to build a more complete picture of a child; more commonly partnership is perceived as schools informing parents about what goes on in the school and seeking their support in achieving the school's aims. In the early stages of Assessment is for Learning (AifL), the programme involved parents in a variety of ways to explore effective partnership, building on evidence of the importance of such links for the success of children's learning. Teachers and headteachers identified parents as potentially one of the biggest barriers to the successful development of AifL in schools, perceiving that parents would question the new approaches to assessment, but the fears expressed were never realised and schools now speak very positively about the support they received from parents. This workshop will reflect on some examples of ways in which AifL schools set about engaging with parents and on the impact of their different strategies, from both parental and professional perspectives. Participants will be invited to reflect on the different models of partnership emerging, contributing perspectives and experience from their own countries.

Contact: Carolyn Hutchinson, carolyn.hutchinson@scotland.gsi.gov.uk

2. The Common European Framework of Reference for Languages: Transparency and Equity

Jenny Bradshaw (NFER, UK), Jenny Dalalakis (ETS-Europe, Belgium), Eli Moe (University of Bergen, Norway) and Saskia Sluiter (CITO, The Netherlands)

The Common European Framework of Reference for Languages (CEF) was published by the Council of Europe in 2001, followed by a draft manual for relating tests to the CEF. At the AEA-Europe conference in Dublin in 2005, we explored some of the work which has been done so far on relating foreign language tests to the framework. This was followed by a lively discussion on the quality of current use of the CEF. In this discussion, we will give a brief update on the projects in which we have each been involved. Group members will then be invited to contribute their own experiences of using the CEF. This will be followed by an evaluation by group members of the extent to which the CEF is helping to improve the transparency and portability of qualifications, and the type of quality control necessary to ensure equity and fairness for test-takers. We will consider questions such as the following:

- Is the CEF being used responsibly by test developers?
- What are the implications for test-takers if claims of links with the CEF are inaccurate?

- What sort of evidence should be available to support a link with the CEF?
- Should there be regulation and accountability – or should it remain open for all to use the CEF as they wish?

Contact: Jenny Bradshaw, j.bradshaw@nfer.ac.uk

3. Evidence or Consequence? Validity issues in educational measurement

Simon Wolming & Christina Wikström (Umeå University, Sweden)

Very simplified, the meaning of validity is if an instrument "measures what it purports to measure". The traditional way of defining the concept of validity is that it includes three main aspects; content validity, criterion-related validity, and construct validity. These aspects are all tightly connected to the instrument, its construction and how the outcome correlates with some external criteria. However, in recent years, many argue that validity incorporates a much wider view, by also emphasising consequential validity aspects of how the assessment outcome is interpreted and used, hence accepting Messick's arguments (see the 3rd ed. of Educational Measurement by Linn, 1989). When validating an instrument, it is often argued that all validity aspects should be included in the analysis. The decision on what approach to take and what validity aspects to include will therefore greatly affect process, methods, as well as the interpretation of the results. However, quite often, the outcome from the analyses within the validation process is not consistent. Some validity aspects may seem acceptable or good, while others are problematic. This discussion group will focus on how to interpret the outcome when a wide validity approach is accepted and the outcome of the validation varies, for instance when traditional validity aspects seem acceptable, but the consequential validity aspects are problematic, or when predictive validity is high but other aspect of construct validity is low. Some empirical examples will be given and participants are welcome to contribute with own examples.

Contact: Christina Wikström, christina.wikstrom@edmeas.umu.se

4. Professional Development (continued)

Chris Whetton, (NFER, UK), Gabriella Agrusti (Università degli Studi Roma Tre, Italy), Frans Kleintjes (CITO, The Netherlands), Kiril Bankov (University of Sofia, Bulgaria), Andrew Watts (Cambridge Assessment, UK), Marian Sainsbury (NFER, UK), Alastair Pollitt (UK), and Eduardo Cascallar (AGI, Belgium).

This session will outline the progress made in AEA-Europe in the areas of Professional Develoment in Assessment. Current plans will be outlined and members will be asked to provide their perspectives and outline their own needs.

The proposals for an AEA-Europe Accreditation Scheme will be described in detail, including the criteria and links to the European Qualifications Framework. We are also seeking members willing to participate in the Professional Affairs Board. The nature of applications for accreditation and the profiles which have to be submitted will be outlined and discussed.

A possible further development for AEA-Europe is to encourage and arrange mentoring of individuals engaged in professional development. Members will be asked to contribute to these proposals.

Finally, the session will explore other avenues through which the association can contribute to the professional development of its members.

Contact: Chris Whetton, c.whetton@nfer.ac.uk

Friday

10.00–12.00

5. Towards more Valid and Equitable Systems of Summative Assessment

Wynne Harlen (University of Bristol, UK). Others taking part are: Gordon Stobart, Richard Daugherty, Judy Sebba, Paul Black, John Gardner and Carolyn Hutchinson

This discussion session will focus on the characteristics that a fair, informative and affordable system of summative assessment should have. An assessment system is taken to include how assessment is used: to help learning and foster deeper engagement with it (formative assessment or assessment for learning); summative assessment for keeping records or making decisions about individual students (assessment of learning); for evaluation of teachers, schools, local/district authorities; for year on year comparison of students' achievements for monitoring national or regional standards. While most attention is given in this discussion to the 'external' summative assessment (that is, for certification, selection, meeting statutory requirements) it is clear that how this is conducted often influences the way in which teachers conduct 'internal' summative assessment (for recording students' progress and reporting to parents) and the 'assessment culture' of the school. To open the discussion, reviews of research findings about the validity, reliability, impact and cost of summative assessment conducted in various ways will be briefly summarised. Some advantages and disadvantages of basing a system on tests and examinations compared with those of a system which make use of assessment by teachers will be discussed. Discussion will focus on the impact of assessment systems on students' motivation and learning goal orientation and on teachers' implementation of formative assessment. Questions for discussion include:

- impact on the curriculum
- differential effect on students arising from how summative assessment is carried out
- issues around reliability, both of tests and teachers' assessments
- problems related to 'high stakes' use of results
- the role of student self-assessment
- workload for teachers and costs to institutions and the system

Contact: Wynne Harlen, wynne@torphin.freeserve.co.uk

6. Curriculum Design and Evaluation

Charles Briffa, Josette Farrugia & Martin Musumeci (University of Malta)

One of the roles of an examination board is curriculum design and evaluation. The board must present syllabi that are relevant to the students' lives and that cater for the needs of the students who will stop studying the subject at that level as well as the needs of those students who will study the subject at a higher level. This discussion group will consider the issues that must be dealt with when constructing and updating syllabi. The Maltese context will be taken as the starting point of the discussion. Participants are encouraged to share their experience, expertise and concerns with those present. The main points that will be tackled are:

- What measures should be taken to ensure equity? [assessment: Paper A, Paper B; coursework; syllabus production; etc; access to exam: gender difference in registration;]
- What importance should be given to subject matter content versus skills in the syllabus?
- What should be the relationship between the different levels of the same subject (eg secondary level - age 16+ and advanced level - age 18+)?
- Should the syllabus be more concerned with providing education for students who will stop studying at that level or to prepare students for higher education?
- What are the characteristics of the current language syllabi in the MATSEC examinations? Should they be changed? What questions do we need to consider about these syllabi?

- What are the characteristics of the current science syllabi in the MATSEC examinations? Should they be changed? What questions do we need to consider about these syllabi?
- What are the characteristics of the current syllabi for commercial subjects in the MATSEC examinations? Should they be changed? What questions do we need to consider about these syllabi?

Contact: Charles Briffa, charles.briffa@um.edu.mt

7. Selecting Reliable Markers – some studies in UK public examinations

Michelle Meadows & Jo-Anne Baird (the Assessment and Qualifications Alliance, UK)

In an equitable assessment system the mark a candidates' work receives is accurate and independent of which examiner does the marking. For this a rigorous examiner selection procedure is necessary. The Assessment and Qualifications Alliance, Research and Policy Analysis Department have conducted a number of studies attempting to identify factors which might allow the prediction of those examiners who are likely to mark most reliably and those who are likely to require additional training or monitoring. Most of the work, however, is not yet in the public domain. This session will review this research and internationally published studies relevant to predicting marker reliability. The relative importance of the following influences will be discussed: examining experience, teaching experience, subject knowledge, previous performance, personality traits and trainability. In particular the findings of two studies will be reviewed. The first study explored whether it is necessary for a marker of Key Stage 3 English to be a qualified teacher with three years' teaching experience (Royal-Dawson and Baird, 2006). The marking reliability of four types of markers with an academic background in English but different amounts of teaching experience were compared: English graduates, PGCE graduates, teachers with three of more years' teaching experience and experienced examiners. Overall there was little difference in the marking reliability of the different marker types suggesting that the criterion of teaching experience could be relaxed to allow markers with graduate-level subject knowledge to mark Key Stage 3 English tests. The second study examines whether psychometric measures of personality can be used to predict marking reliability in GCSE English of individuals with distinctly different levels of examining experience, teaching experience and subject knowledge. That is experienced GCSE English markers, PGCE English undergraduates, English undergraduates and Non-English undergraduates. Preliminary findings from this study will be available. The studies have been funded by AQA and the National Assessment Agency and are highly relevant to the work of the newly established Institute of Educational Assessors. Implications of the findings for the Institute will be discussed.

Discussion will focus upon the following kinds of issues:

- What marker selection criteria are used in different countries/assessment systems? To stimulate discussion and aide research in this area, participants are asked to bring details of the marker selection arrangements in their country to the workshop.
- Are these criteria valid? What criteria should be used?
- Can awarding bodies recruit non-teachers as markers without losing public confidence and current examiners' goodwill? How?
- Which personality measures might be important in predicting marking reliability?
- Can we generalise the predictors of marking reliability in Key Stage 3 English and GCSE English to other subjects/qualifications?
- In which subjects/qualifications do markers require subject knowledge and/or teaching experience?
- Could generic guidelines as to the level of subject knowledge and teaching experience needed to mark different types of items be developed? What might these guidelines consist of?
- Is it possible to employ markers from overseas? What quality control procedures would be needed?

Contact: Michelle Meadows, mmeadows@aqa.org.uk

8. Establishing European Quality Standards for Examinations and Assessments (continued)

Frans Kleintjes, Anton Beguin, Piet Sanders (CITO, the Netherlands), Eduardo Cascallar (AGI, Belgium) & Gerben van Lent (ETS Europe)

Contact: Eduardo Cascallar, agi_group@msn.com

saturday

10.00–11.45

9. Value Added Measures and the Enhancement of Equity in Assessment

Harvey Goldstein, University of Bristol, UK

The use of 'value added' procedures for comparing schools is well known and embodies the principle that 'fair' comparisons based upon pupil performance should adjust for the initial achievements of those students because these may differ greatly between schools and because initial achievement is a powerful determinant of final performance. We note that this makes the assumption that the reasons for initial achievement differences are unrelated to the 'effectiveness' of the school, and this is an issue for discussion. Another issue is whether further factors, such as gender or ethnic origin, should also be adjusted for. A common feature of all assessment systems is the 'uncertainty' or 'measurement error' associated with any judgement. It would seem a fundamental principle for equity that such uncertainty is reflected in any judgements or public exposure and this is another issue for discussion. A related issue is that of stating the strengths and weaknesses of the assumptions made by any particular assessment system, including reasons for choosing particular instruments, alternative views etc. Likewise the opportunity to contest an assessment, whether by an institution or an individual, should be part of any system claiming to be fair. Nevertheless, this set of issues has not been widely discussed in the context of value added performance and the forum will bring together a range of views and experiences to try and move the debate forward. The idea of 'value added' also plays a role in individual assessment debates, although not usually discussed in such terms. For example, concerns about gender fairness have often centered on the use of separate gender 'norms' for selecting or comparing students. In effect this involves an 'adjustment' for gender, or for example ethnic, differences and can be viewed as a form of 'value added' measurement with an analogous set of issues to those outlined above. Another aim of the forum will be to pursue analogies between such debates about assessment equity for individuals and debates about equity of value added measures for institutions.

Contact: Harvey Goldstein, h.goldstein@bristol.ac.uk

10. Guarantees for Equity of National Exams?

Jo-Anne Baird (AQA, UK), Anton Béguin (CITO, The Netherlands), Theo Eggen (CITO, The Netherlands) and Claire Whitehouse (AQA, UK)

In this discussion group we introduce a number of ways to set standards on a form of an exam:

- Standard setting using the judgement of experts
- Standard setting using assumptions on the population (equating on the curve)
- Standard setting supported by application of IRT: conditions and methods.

We show where these procedures are (or can be) used in setting the standard on the GCSE exams of AQA in England and in the central examinations in secondary education in the Netherlands. The discussion part will be on the pro's and con of the different procedures in the given situations.

Contact: Theo Eggen, theo.eggen@cito.nl

11. Development of a New Model of National Standards-based Educational and Assessment System: the path to quality and equity

Eduardo C. Cascallar (AGI, Belgium) and Michael Fast (UK)

The main theme of this discussion group will focus on a new model for the development of national educational standards compatible with an international vision of educational standards, and a compatible set of assessments to examine the performance of students at various grade levels on the standards implemented in the programme. The model depends upon the establishment of a "common language framework" to facilitate comparison of curriculum indicators with standards of educational systems recognized for best practice. One of the goals was to achieve national, regional, and cross-national integration of the educational effort, with solid process and outcome measures. Emphasis was placed on the use of computer technology to enable efficient and rapid comparison of standards across systems, both in terms of horizontal coherence (within a grade level) and vertical coherence (across grades and grade clusters). The model then highlights the role that standards play in determining alignment between classroom pedagogy and formative and summative testing. These two lines of interrelated standards-based assessments were developed with the collaboration of local and international experts. Extensive training in advanced psychometric methods was provided to local staff and professionals. The formative set of tests was developed for use in the classroom, based on IRT conceptualizations, but geared to provide the teacher with an evaluative tool for individual diagnosis and to inform classroom pedagogy. In addition, annual summative tests were designed and developed to establish baselines and rates of progress centred on the standards developed. It is argued that this process is fundamental in the attempt to establish high quality and equitable educational and educational assessment systems.

Some suggested topics for discussion:

- How to define national educational standards for testing purposes.
- How to establish national priorities for educational standards and their assessment.

Contact: Eduardo Cascallar, agi_group@msn.com

12. Gender in Interaction with Assessment: Implications for how we think about gender equity in assessment

Jannette Elwood (Queen's University, Northern Ireland) and Patricia Murphy (The Open University, England)

The theme of this discussion group will be 'Gender in interaction with assessment'. The discussion group is linked to the keynote address given by Patricia Murphy with the idea of extending the ideas raised in this keynote for further exploration. Gender is a always positioned as a key variable in the analysis of much assessment and testing data and is thus commonly treated as just another dimension on which to report data. This discussion group will be organised around three core ideas:

(i) The need to move away from this limited understanding of how gender is considered in the field of assessment and testing.
(ii) That assessment is a cultural process and one that mediates understandings of gender through teachers' and students' views of success.
(iii) That a socio-cultural position on the definition of gender and assessment helps us better understand the complex patterns of performance that emerge through the interaction of teachers, students and assessment tasks.

Participants will be asked to consider these three core ideas during the discussion group with the aim of acquiring some degree of 'European' perspective on these issues which is lacking across the present literature.

Contact: Jannette Elwood, j.elwood@qub.ac.uk

Open Papers Abstracts

ThursdayNovember9

Session A: Equity Issues in Different Contexts

Chair: Gordon Stobard

1. Educational Equity Balance in Finland

Jakko Hautamäki and Sirkku Kaupiainen et al (University of Helsinki, Finland)

Our research group has developed a framework for learning-to-learn and the Finnish Learning-to-Learn Scales for assessing this. We have data from representative samples, sizes from 2000 to 4500, of 6th graders (the last year of the primary school), collected 1997-98 and 2002-03, 9th graders (the last year of the lower secondary school), collected 1997-98 and 2001-02, and of 6th form and vocational education, 2000-01. Taken together, the studies constitute one of the largest assessment and data collection exercises of competencies and beliefs, relevant for later engagement in learning-demanding situations, and with norm- and criterion referenced scales. Major results from our earlier studies have been to show that Finnish schools are surprisingly equal, but that on the epidemiological level factors (which an individual cannot escape) like gender, parents' education, and the language-of-schooling have explanatory power, i.e., independent effects. In order to make clear the issues of this proposal we shall present some of the results. One solution to condense our results is to use the concept equity balance. With the concept we refer to the distribution of results and outcomes analysed from the point of view of several educationally relevant context factors: the equity balance is a table where the explained variances are given for these context factors in relation to several outcome criteria. We will present a comprehensive analysis of the equity balance, and will compare, with comment, our national results with PISA estimations of between-school variations in PISA 2000 and PISA 2003, where, on avarage, the between-school variation in Finland is around 6-7 %.

2. Learning assessment: the role of 'assessment for learning' in young people's learning careers in the UK

Kathryn Ecclestone and Roger Murphy (University of Nottingham, UK)

The concept and practice of 'assessment for learning' have moved from the field of academic educational research to policy and the creation of assessment regimes in the UK's education and training system. The idea that well-executed, well-designed formative assessment both raises achievement and enhances motivation and engagement with learning is widely supported at all levels of the system. There is strong theoretical and empirical evidence that supports the promotion of assessment for learning. Despite its obvious appeal, theoretical and empirical research by the presenters in the compulsory, post-compulsory and higher education sectors shows a tendency to distil principles of assessment for learning into techniques. This instrumental approach avoids dealing with the socio-cultural and political factors in learning and assessment contexts that affect students' achievement, motivation and engagement.

This paper draws on several studies in post-compulsory and higher education to explore how assessment for learning in different socio-cultural contexts has to be related to young people's 'learning careers', where identity, motivation and broader influences interact with teaching and assessment activities. The paper argues that over-regimented, prescriptive assessment regimes in the UK distort the educational principles behind assessment for learning and contributes to the shaping of students' learning careers in particular assessment driven ways. It

offers ideas for how findings from research about themes explored in the paper might be used to enhance the design and practice of assessment.

3. Evaluation and Equity in Local Education systems

Tali Freund (Centre for Educational Technology, Israel)

Up until few years ago, the responsibility for the education system in Israel was almost exclusively in the hands of the central authority. In recent years, local authorities are more and more involved in education, showing growing interest in systematic evaluation of educational achievements, and in allocating of resources according to identified cores of difficulty and strength. Evaluation experts can assist leaders of local education systems with the implementation of an "evaluation culture", and in particular with setting the goals they wish to promote, defining quality indicators for the local education system, developing evaluation tools, gathering data, analyzing it, drawing conclusions, and presenting it to the clients.

This lecture will demonstrate some evaluation processes performed in local education systems in Israel, guided by the ambition to maintain equity in each stage – from design to implementation, using the following:

- Diverse evaluation objects reflecting different indicators chosen to assess the quality of the local education system. Equity is obtained through evaluating not only students' scholastic achievements but also by evaluating dropout rate, graduates' integration in the academic system and the workplace market.
- Two-steps evaluation model which includes scholar achievements' measurement in the beginning and end of the year. This model ensures equity as it is based on evaluating schools' relative progress, as opposed to the national model in Israel which is based on evaluating schools' absolute achievements at a certain point of time regardless of any inhibiting background variables prevalent in weaker schools.
- Diverse evaluation tools including external tests to assess achievements developed by professionals, and tools for internal evaluation based on teachers' judgments. Equity towards teachers is obtained through the trust expressed by the system in their evaluation skills.
- Evaluation products in various resolutions ensure equity through evaluation reports: A general one reflecting the whole system, an individual one for each school, for each class and for each student. And also through data analyzed by groups (new immigrants, students with special needs), enabling customizing solutions for differential needs.

4. Automated Predictive Systems in the Prediction of Educational Outcomes

Eduardo Cascallar (Assessment Group International, Belgium) and Tracy Costigan (American Institute For Research, US)

Neural network analysis is a machine-learning approach to modeling complex patterns from data, with the aim to predict outcomes. There are several advantages to such machine learning approaches in comparison to more traditional general linear model statistical approaches, and more traditional statistical methods which are limited by parametric assumptions, as well as by the number of inputs and interactions that can be included in a single model. A general model is presented for the analysis of complex predictions and classifications in educational settings. In particular results from a pilot study to detect and classify students with potential reading problems will be presented and explained. A back propagation (BPN) algorithm model was used, due to its excellence in generalization, and its ability to classify extreme cases despite lack of data. Since model validation is critical in establishing the optimal solution for a neural network, a number of critical steps were taken to examine the results and the model proposed: (a) an innovative methodology was used to get to "see" inside the neural networks "black-box" and establish the relative importance of weights by modeling the predictive values using a rule-induction technique to develop classification systems in which the "decision rules" identifies the variables of importance; (b) to validate the neural network model, it will be important to partition the data into two samples: a training/testing sample and a validation sample. This technique, common in more traditional statistical approaches, protects against over-fitting of the model; (c) causal relationships among components/constructs in this model

as well as the underlying variables within these components/constructs were examined; (d) finally, in order to evaluate the performance of the neural network system, a number of measures used provide a means of determining the quality of the solutions. These measures will be explained in the context of the study presented.

Session B, Part 1: University Entrance

Chair: Jan Wiegers

1. A New Model of Admission Examinations for Universities in Georgia

Maia Miminoshvili and Iamze Kutaladze (National Assessment and Examination Centre, Georgia)

A new model of admission examinations was introduced in Georgia when secondary and higher education were experiencing a significant crisis. Corruption in university entrance examinations was at its peak. These tests had lost their basic function – to select entrants based on their knowledge and skills – resulting in complete disintegration of fundamental principles of fairness. Schools were failing to teach important parts of the curriculum, and universities were requiring students to cram knowledge that was not part of it. As a consequence entrants had to seek the services of private tutors who quite often happened to be the same as the ones setting the entrance exams. However, many students could not afford this. The new model replaces the former exams set by each separate university by one unified exam. Both in terms of content and way of administration, it introduces several significant changes, which contribute to elimination of disparities that were characteristic for the old system. It is based on the principle of complex assessment of entrants' achievements (subject knowledge) and skills. Such an approach balances inequality between students caused by differences in access to educational resources or differences in family background. Significant changes were made to the administration model of the examinations guaranteeing secure sessions, confidentiality during marking, unbiased markers, availability of copies of the original student scripts for appeal procedures, etc. The exam revolution seems to pay off. In comparison to previous years, the number of students from rural regions admitted to higher education institutions has increased and so has the number of students from the poorer families. The unified entrance exams have become a powerful tool for decision makers in the field of education. It is a fair and effective means of allocation of students to educational resources. And for the first time in Georgian history there are comparable results that shed a light on the outcomes of secondary education in the various regions and cities of the country.

2. Equity in University Entrance Examinations

Grace Grima and Frank Ventura (MATSEC Examination Board, University of Malta)

In 'A Fair Test? Assessment, Achievement and Equity', Gipps and Murphy (1994) provide an understanding of the multi-faceted dimensions of equity in examinations. Their comment that "A level is hardly researched at all" (p. 276) may be dated in the UK context, but still holds true as a focus of study in the local context, with reference to the 18+ system of examinations. Taking the position that we cannot develop a totally equitable examination system, in this paper we discuss equity issues within the context of a system of university entrance examinations. In particular, the issues will be discussed within a bilingual context and focus on an examination structure that differs significantly from the three 'A' levels formerly required for University admission in Malta. The current system requires students to gain passes in a Language subject, a Mathematics or Science subject, a Humanities subject as well as Systems of Knowledge. In addition to this Matriculation Certificate (MC) certificate, students also need passes in English, Maltese and Mathematics at Secondary Education Certificate (SEC) in order to gain entry into university.

The issues to be discussed relate to three phases of the process leading to university entrance:

Phase I: Access issues (the teaching/learning experience, school types, the length of the school year, the language issue in a bilingual context, private tuition, fees, regional distribution of candidates, the need to widen access)

Phase II: The examination process (the structure of the examination per se, syllabus construction, the quality of

the examination papers, the marking and grade awarding processes, professional training of the examiners, special needs)

Phase III: Consequences of the examination (monitoring results, maintaining standards, course entry statistics, graduation statistics).

The paper will explore the extent to which the examination board has control of or influence over particular equity issues and what measures are possible to maximize equity within a small state context

Session B, Part 2: Uses of the Common European Framework of Reference

3. The Common European Framework of Reference: A tool for value-added assessment?

Therese Hopfenbeck and Elisabeth Ibsen (University of Oslo, Norway)

The Common European Framework of Reference (CEF) was used in the Norwegian National tests of written English for the first time in Norway in 2004. It can be argued that the implementation of CEF represents a move towards a more equitable form of assessment, yet there remain several obstacles, including; reliability, lack of common assessment culture among stakeholders and competency in applying the CEF. The majority of the tests were assessed by the students' own teachers. A sample of the tests was assessed by experts-teachers who did not know the students. The reliability in 2004 was estimated to be too low for the evaluation of schools and the required publication of results, and also too low for learners to measure their own progress. There was also a clear trend that the class teacher gives the highest level. The evaluation report of 2005 shows similar results as the report of 2004. Due to these findings, there has been a temporary stop of national tests in Norway. It might be argued that the use of CEF-levels can be accepted as a useful and equitable tool, if the teachers, pupils and parents are more involved in the process of using it. It can also be argued that CEF is better suited for portfolio assessment within the class, than as a tool for assessment of national tests. An experiment was set up at the Department of Teacher Education at the University of Oslo. Four newly graduated teacher students marked pupils' test responses with CEF-levels for the national tests. Results from the student teachers' assessment showed considerable discrepancy between the levels given for the same pupil's answer, but the ranking of pupils functioned reasonably well. The experiment, then, mirrors the national problems with discrepancy between levels given.

4. Valid comparison of learner groups in a multilingual proficiency framework

Neil Jones (Cambridge Assessment, UK)

A key element of the UK's national languages strategy is the Languages Ladder, a new voluntary recognition system. The Languages Ladder's proponents aimed to complement existing qualifications frameworks in two important ways. Firstly, they found these to be confusing or uninformative about the levels of competence they represented, and secondly they found them unsupportive of learning. Consequently the new framework should accredit clearly-defined functional proficiency levels and provide a "learning ladder" of bite-sized, accessible learning targets. They pointed to the Common European Framework (CEFR) as a model.

Asset Languages is the system being developed by Cambridge Assessment to implement the Languages Ladder. The scheme is comprehensive, including over 20 languages: both "modern foreign languages", and those spoken by particular communities in the UK. It targets three contexts (Primary, Secondary, Adult), with skills assessed separately. It offers two assessment strands: external assessment at six major stages, and more informal teacher assessment at 17 finer grades.

The Languages Ladder's proponents stressed its motivational role in learning at the lower levels. But every level of language proficiency has some practical use, and thus some value. Languages are problematic for current educational assessments, whose societal value relates importantly to notions of learning effort. Languages however may be acquired with more or less effort, depending on the language and the context of the learner. The societal value of Asset Languages qualifications should be simply the value attached to having particular language skills, however acquired.

Thus equity for Asset hinges on clearly identifying proficiency levels, enabling valid comparison across widely differing languages and learner groups.

I will discuss conceptual and technical issues in constructing this complex framework, relating these to current efforts in the European context to develop a methodology for linking language assessments to the CEFR.

FRIDAY NOVEMBER 10

Session C: Approaches to Improving Examinations

Chair: Kathleen Tattersall

1. Developing and piloting a methodology to enable students to participate in test development

Alison Wood (QCA, UK) and Alastair Pollitt (UK)

This paper is set in the context of general qualifications (such as General Certificate of Education Advanced levels) in England, Wales and Northern Ireland, but addresses the general issue of the validation of assessments and the avoidance of construct irrelevant variance. It is based on the premise that, for reasons of equity for candidates, robust methodologies for making judgements about the demand of questions are essential. It argues that, for the demands of an assessment to be valid ones, they must be intended by the question-setter who, therefore, needs to know which factors may (or may not) impact on the demand of a question and how they might do so. Current methodologies for generating this information, using expert judges, are not entirely satisfactory and this might well be because of the expert status of those judges. These shortcomings can be overcome by using student judges, but they need the support of a methodology which is tailored to their specific needs, as non/semi-experts: the Relative Judgements Construct Elicitation (RJCE) methodology. In particular, the methodology needs to empower the student and address the power imbalance between student and researcher. The paper sets out five proposed success criteria for a methodology, then reports the findings of a pilot study of this methodology, setting out the advantages and disadvantages and arguing that it meets the success criteria well. The methodology is direct, provides authentic data, gives access to the whole of the question-answering process, has explanatory power and is economical. The paper then argues that the methodology is generalisable beyond mathematics. The paper concludes that there is a clear rationale for developing a role for students in the test development/test evaluation/comparability process.

2. Assessment of writing proficiency

Ragnar Thygesen and Rolf Fasting (University of Stavanger, Norway)

The presentation deals with achievement testing of writing proficiency in students at the fourth, seventh and tenth level of primary education, including the theoretical rationale behind the tests, how the tests were implemented, and the procedures for the assessment of the texts produced by the students. The results from a representative sample will be presented and discussed. We will present a theoretical model of writing in the form of a dynamic, graphical figure. The model assumes that competence in writing expresses itself in the quality of acts of writing, like proficiency in presenting, exploring, describing, entertaining etc. through writing. The assessment model prescribes the way in which the examiners should assess the quality of the students' texts by using a categorical system, e.g. they look at coherence, at how well the text gets the message across, at spelling and other aspects. In accordance with fixed benchmarks, we define 'sufficient proficiency in writing' in accordance with a three-level model that differentiates between achievement in this way: The text is almost as can be expected from most students at the actual level/much better than expected/much poorer than expected. Well-trained, external examiners assessed the sampled texts. A striking result is the low agreement between the examiners, which emphasizes the need for improving the concepts that characterize writing competence. With improved concepts

we could develop national norms in this domain. Another surprising result is that a high number of texts on many categories at level 4 and 7 were generally assessed as 'much poorer than expected', which should be alarming for the schools. On the other hand, at level 10 the texts produced were generally better than expected. Perhaps not so surprising is the finding that girls in general show better writing competence than boys.

3. A taxonomy of Sources of Difficulty in the assessment of ICT

John Threlfall and Nick Nelson (University of Leeds, UK)

One of the considerations in any assessment is what determines performance, and one way of looking at this is in terms of Sources of Difficulty (SODs) (Pollitt and Ahmed, 1999) – that is, what prevents success on a task that has been set with assessment in mind. Some SODs are legitimate, such as not having the knowledge or skill that is being assessed; some are not legitimate, arising from factors irrelevant to the assessment, and which generate false negatives (failure on the task despite having the competence that is being assessed). This presentation is focused on computer administered and scored assessment of ICT, and what prevents success on ICT tasks other than absence of ICT capability. It addresses how legitimate and non-legitimate SODs might be distinguished – since the boundary between them is not as clear cut as one might hope – and goes on to consider how non-legitimate SODs might be classified, and what that suggests about how such SODs in computer-delivered ICT assessment might be removed or mitigated.

Session D: Aspects of Science Assessment

Chair: Grace Grima

1. PISA's Computer-based Assessment of Science (CBAS) – A gender equity perspective

Are Turmo and Svein Lie (University of Oslo, Norway)

In December 2003 the OECD put out a call for tender for an optional computer-based assessment of scientific literacy within the frames of the PISA study. In the spring of 2005 a field study was conducted in 13 of the PISA countries, among them Norway. The items for CBAS were based on the same framework used for the development of paper-based science items. In this framework, scientific literacy is defined as the capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity. The study was implemented in Norway in the period from March to June 2005 by the use of 6 carry-in laptops. 10 students in each participating school were sampled for the study. In total, 315 15-year old students participated. All students did a one hour paper-based test and a one-hour computer-based test. The students also responded to a paper-based background questionnaire. In addition to the cognitive response data, also behavioural data were recorded from the computer-based test. The Norwegian results show no statistically significant gender difference for the paper-based test, while a substantial gender difference in favour of boys is established for the computer-based test. The presentation will discuss these findings based on analyses of both the response data, the behavioural data and the student questionnaire data. The results show that boys express higher motivation and larger preference for doing the test on PC compared to on paper. Furthermore, the behavioural data show that boys more than girls make active use of the multimedia elements in the PC test. A significant positive correlation between the frequency of use of multimedia-elements and achievement is found for the girls. However, this relationship is not found for the boys.

2. An evaluation of bookmarking as a judgemental equating method in KS2 science

Catharine Parkes (NFER, UK)

In England, bookmarking has been used in the development of the Key Stage 2 National Curriculum assessments for a number of years as means of involving teachers in the standard setting process. The bookmarking procedure is heavily reliant on experienced teachers' knowledge of the performance of pupils on the tests. The teachers are presented with a live test booklet ordered to difficulty and asked to predict whether pupils performing at certain

levels (according to the National Curriculum Programme of Study) could successfully answer each item. The bookmarked items identified by the teachers can be used to work out the equivalent cut score on the test. Over the years, a number of amendments were made to the process to deal with issues that seemed peculiar to the science papers in a key stage 2 context. Teachers perceived that particular items were more difficult than pupils actually found them. Also, there are only a small number of pupils at the lower levels and the teachers' lack of experience of this group meant that they underestimated what the pupils could achieve on the papers. The amendments will be discussed, as will their impact. Bookmarking and other judgemental methods are designed to provide an independent means of setting cut scores, but how do you verify the cut scores produced? In bookmarking, the cut scores varied widely when compared with the statistical predictions. When the process was carried out as a means of training panellists in the following year, the cut scores produced varied widely from the cut scores produced the previous year, implying that the process is not entirely reliable. We need to determine what the purpose of carrying out a judgemental exercise is: to provide independent evidence or corroborate statistical methods?

3. Towards gender inclusive assessment in science

Deborah Chetcuti (University of Malta)

The way in which we assess students, the assessment tasks we choose, and the context of the questions chosen all influence the performance and achievement of students. "Assessment tasks have social consequences…(which) manifest themselves in the form of differential performance between different sub-groups" (Elwood & Murphy, 2002, p. 396). One of the most important social consequences of assessment in science is the way in which achievement in science influences teachers' and students' views of their potential in science. Therefore in order to encourage more students to take up science subjects at post-16 level we need to ensure that our assessment practices in science are equitable and "fair and just for all groups" (Gipps & Murphy, 1994, p. 18). This paper explores whether the SEC science examinations in Malta are fair and just for girls and boys attending different Maltese schools. Using examination results for Biology, Chemistry and Physics at SEC level, the paper shows that Maltese boys are outperforming girls in Biology and achieving almost at par with the girls in Physics. In Chemistry it is the girls who are outperforming the boys. The examination results also show that achievement is not only related to gender but also to the type of school which the students attend. Using interview data from teachers teaching in different Maltese schools, the paper tries to provide an explanation for this difference in performance between the girls and boys attending the different schools. Based on the experiences of the participants of the study, the paper also tries to offer some suggestions for moving towards gender inclusive assessment in Maltese schools. The paper will offer examples of how key principles such as using a multi-sensory approach to teaching, using relevant contexts, relating assessment to out-of school experiences, and promoting independent learning can help both girls and boys achieve their best potential and redefine their perceptions of science.

Poster Session Abstracts

FRIDAY NOVEMBER 10

12.00–14.00

A. Incorporating test users in the test development process – a case study

Lise-lotte Appelgren (Enlight, Sweden)

Test development consultants are constantly facing new challenges in terms of new clients, new content areas, new test formats and new purposes for the test outcome and its interpretation. The content area can be very specific and previously unknown for the consultant and, for instance, focus on technical or medical areas, where the purpose of the future test is to be used for certifications, or for assessing knowledge levels for future educational investments. Clients, i.e. future test users, who seek the advice and help from these consultants, often have very limited understanding of the necessity and complexity of proper test development procedures, and often have the hope of simply ordering an instrument with a certain quality. They often know why they need this instrument and how they plan to use its outcome, but do not know how to define the objectives of the content area or why they should do it. Without this definition, it is difficult to develop a blueprint and hence almost impossible to end up with a valid test, irrespective of how much time and effort that is given to the item construction. In a case like this, where one part provides the content information and the other the test development skills, it is necessary that both parts are involved in the test development process. This presentation describes a model for how to analyse and define a new content area, and to develop test specifications ("the blueprint") in collaboration between the future test user and the test developer. The empirical evidence is provided by a case study describing a test development project conducted at Enlight. The presentation describes the test development process, the model for developing the blueprint and discusses the advantages and disadvantages of this model.

Contact: Lise-lotte Appelgren, lise-lotte.appelgren@enlight.net

B. A Computer-assisted Test Design and Diagnosis System for Classroom Teachers

Quingping He (Durham University, UK)

With the rapid development of information and communication technologies (ICT), computer-assisted assessment (CAA) is becoming increasingly important in education. One of the most valuable benefits of using CAA is that it can provide timely and specific information on the performance of each student, which can be used for diagnosing areas where students have individual difficulties. Such information can be used to provide guidance for future study. Recent years have seen increasing use of computer adaptive testing (CAT) in assessment. CAT is unique to CAA. The primary difference between an adaptive test and a classical test is that in an adaptive test each individual will answer a different set of questions drawn from an item bank based on his/her ability, whereas all students will take the same set of questions contained in a classical test. This poster presents a computer-assisted test design and diagnosis system which can be used by classroom teachers to design both classical and adaptive tests and analyze test results. The system provides the following main functions:

- Create items and construct item banks.
- Design tests effectively by selecting items from an item bank. Both classic and IRT based tests (including CATs) can be designed using the system.

- Conduct designed tests on computers or on paper.
- Analyse test results. In addition to providing basic test statistics, the system will also be able to undertake detailed diagnostic analysis on student's performance at both individual and class/school levels. The system has the ability to generate information on the performance of students and test items that can be easily used by teachers to identify curriculum areas where students are under performing.
- Undertake test item analysis using an IRT model - the Rasch Model. This will enable the establishment of large calibrated item banks for schools.

Contact: Quingping He, quingping.he@cem.dur.ac.uk

C. Classroom materials to support Assessment for Learning through Literacy

Juliet Sizmur & Claire Hodgson (NFER, UK)

This presentation will introduce Assessment for Learning: Literacy 7-11 a set of literacy resource materials designed to help children understand and monitor their own learning by giving and responding to feedback. The materials, published in January 2006, were developed by a team from NFER's Department for Research in Assessment and Measurement (English Teams), in consultation with class teachers and pupils as well as experts in the field of Assessment for Learning. They are specifically aimed at helping pupils develop skills in peer assessment and to encourage teachers in the development of focussed learning intentions and the sharing of success criteria. The classroom activities can be used differentially across the full range of abilities and promote skills that lay the basis for the development of personalised, self-directed learning. In addition to the materials:

- examples of the 8 colourful reading booklets – 2 booklets for each of Years 3, 4, 5 and 6 (ages 7 – 11)
- Posters of the following will be included:
- the structure of the materials – showing the simultaneous learning intentions based in literacy and in peer assessment
- an overview of a selection of activities across the packs
- an introduction to imaginary pupils (Tom and Sam) and their modelled comments (as presented in the resource packs)
- some examples pupil's work and the development of their peer assessment comments

Contact: Juliet Sizmur, j.sizmur@nfer.ac.uk

D. An Evaluation System for Computer-Based Tests

Piet Sanders (CITO, The Netherlands)

In the 'Evaluation System for Computer-Based Tests', five criteria are used to assess the quality of computer-based tests. The first criterion is concerned with the theoretical basis for test development, the second criterion with the quality of the test material and test manual, the third criterion with norm-referenced and criterion-referenced score interpretations, the fourth criterion with the reliability of the test, and the fifth criterion with the (construct, criterion, and consequential) validity of the test. The determination of the quality of the criteria, is based on the ratings given by qualified and independent raters on different indicators for the quality of the criteria. Dependent on the ratings, criteria will be assessed as good, satisfactory, or unsatisfactory. In the poster presentation, the five criteria, the corresponding indicators, and the rating procedure of the criteria will be presented.

Contact: Piet Sanders, piet.sanders@cito.nl

E. The Predictive Validity of the SweSAT

Per-Erik Lyrén (Umeå University, Sweden)

The Swedish Scholastic Assessment Test, SweSAT, has been used as an instrument for selection to higher education in Sweden since the late 1970's. The importance of the test, and hence its validity, cannot be exaggerated, and

one of the main indicators of the quality of the test is its predictive validity. Several studies have been performed over the years, but the methods used, including the choice of predictors and criteria, have differed. Also, there are issues regarding data considerations that need to be illuminated. To conclude, there are several problems connected with the predictive validity studies that need to be addressed and dealt with. The aim of this poster is to give a brief account of the methods, data considerations, results and possible problems of previous predictive validity studies. Also, some thoughts on how predictive validity studies may be improved, what consequences the introduction of ECTS grades may have on the predictive validity of the SweSAT, and alternatives to predictive validity studies in the validation of the SweSAT will be presented.

Contact: Per-Erik Lyrén, per-erik.lyren@edmeas.umu.se

F. Presenting a picture – teacher assessment in Wales

Hilary Cox, Juliet Sizmur, Claire Hodgson and Bethan Burge (NFER, UK)

A research team at the National Foundation for Educational Research (NFER) have been working with Qualification Curriculum and Assessment Authority for Wales (ACCAC) to develop a set of materials to support teachers in gathering evidence to support both their on-going and summative teacher assessments. These tasks and activities are designed to help teachers assess the full range of attainment levels encountered at the end of primary school, including those children with special educational needs. These materials have been developed in response to changes to the assessment system in Wales where there has been a move away from statutory testing to teacher assessment in the subject of 'English' at the end of primary schooling. As part of the development, these materials were trialled with a number of children in schools across Wales. In order to demonstrate the versatility of the newly devised assessment materials, we will present collections of children's work as case studies showing how children of all abilities have tackled some of the tasks. These new support materials for teacher assessment have a central stimulus text and within the collection of tasks, specific activities are identified for higher and lower level pupils as well as those designed to meet the abilities of all pupils. Teacher feedback indicates that these new support materials are successful – no pupils are made to feel excluded or different. We will highlight the common tasks completed by children of a range of abilities as well as activities geared more specifically to specific ability levels.

Contact: Hilary Cox, h.cox@nfer.ac.uk

G. Measuring Concepts and Factual Principles Compared to Measuring Understanding and Application

Lina Wahlgren (Stockholm University College of Physical Education and Sports, Sweden),
Ingemar Wedman, (Gävle University College, Sweden) and Sara Franke-Wikberg (Umeå University, Sweden)

In this study we have focused on the examination process of tertiary education. We have followed four main programs and made certain in-depth follow-up investigations concerning two courses in the various programmes studied. The four general programmes that have been studied are medicine, psychology, engineering and teacher education. Our study has focussed on in-depth interviews with teachers and students, followed by psychometric analyses of the examinations carried out in the various programmes. Aside from the examination process per se, we also obtained information on economic matters connected with studies, and the difficulties of failing students. The examination process varies considerably between the different programmes. One is using a traditional examination through paper-and-pencil tests while another is using a so called "home examination" where the students take home the examination and take the examination in his or her own. In a way, this difference in using examinations goes back to a discussion on how to focus on knowledge and skills compared to understanding and application, a division which is hard to follow in the examinations used. From a psychometric perspective, many difficulties are noted. Questions concerning the empirical value of individual items are never asked and information on reliability and validity are never addressed. The thinking around these questions and the empirical results obtained in our study are illustrated in the report. In the final part of this report, we note that there are large

differences between the programs studied, and also that the students adjust to the manner in which the examination is carried out. We also discuss the way the economic resourcing of examinations in Sweden (and elsewhere) actually affect the way the examination is carried out and scored.

Contact: Ingemar Wedman, ingemar.wedman@gih.se

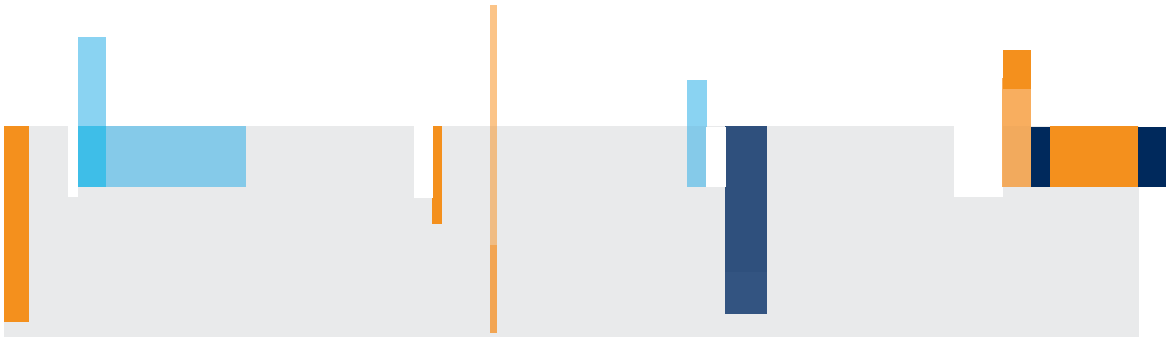
H. Developing electronic marking and management solutions

Graham Hudson and Brian Carbarns (DRS, UK)

The presentation will explain the progress made to date in the establishment of large-scale electronic marking services and the potential that this brings for improved examination management in the UK and elsewhere in the world. Discussion will cover how marking quality is monitored and maintained through multiple marking methods, ensuring anonymity, impartiality and avoidance of bias during marking. Some areas of potential future use of the electronic marking system for supporting other assessment approaches will be included. Questions and discussion will be welcomed.

The poster session will be given by Graham Hudson, National Business Development Manager for Education in DRS. Graham has over twenty years experience of implementing and managing large-scale assessments within the UK, including the national curriculum tests and establishing a government-funded programme for implementing the use of new technologies in marking.

Contact: Graham Hudson, graham.hudson@drs.co.uk



aea
EUROPE

About the Association for Educational Assessment-Europe

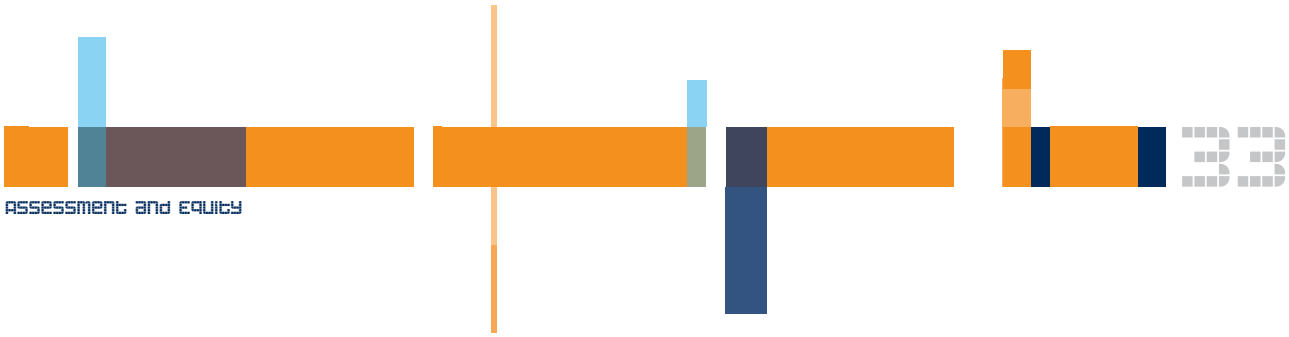
Goals of AEA-Europe

The general goal of AEA-Europe is to act as a platform for discussion of developments in educational assessment in Europe, fostering co-operation and facilitating liaison between organisations and persons active in educational assessment across the whole of Europe. AEA-Europe defines educational assessment in its broadest sense including academic, professional and vocational contexts and is equally concerned with both assessment processes and products.

Aims and Objectives

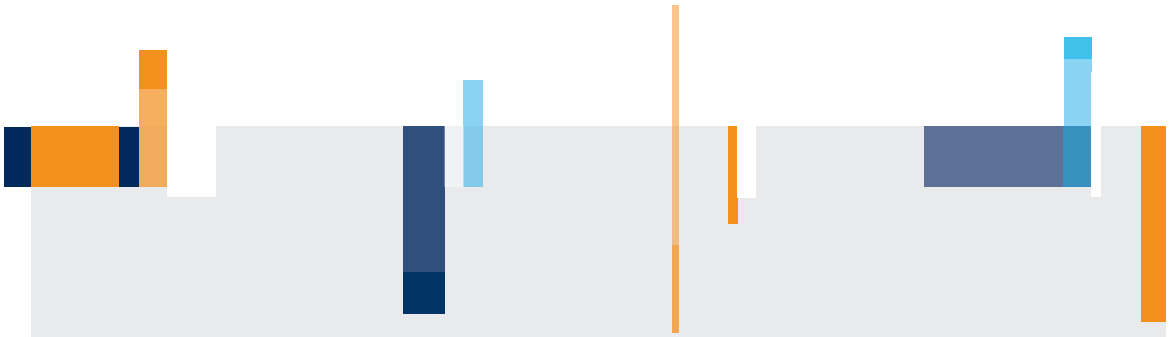
The aims of AEA-Europe are to:

- improve communication among European organisations and institutions interested in educational and occupational assessment through sharing of professional expertise, exchange of knowledge and collaboration between members through conferences and publications;
- provide a framework within which co-operative research, development, implementation and evaluation of projects, involving educational assessment can be undertaken;
- foster and enhance collaborative networks and projects between organisations and individuals across the whole of Europe;
- co-operate with other agencies having complementary interests;
- engage in a range of activities that will lead to the improvement of assessment processes and products and their appropriate use by organisations, institutions, agencies and other associations throughout Europe;
- and enhance awareness of assessment processes and products in relation to their impact on learning and understanding.



Council of AEA-Europe

www.aea-europe.net



aea
EUROPE