- INTERNATIONAL SYMPOSIUM -

METHODOLOGICAL TOOLS FOR ACCOUNTABILITY SYSTEMS IN EDUCATION

6-8 February, 2006
Anfiteatro - Building 36

Joint Research Centre
Via E. Fermi 1, Ispra (VA) - Italy
# Table of Contents

**General Information** 3

**CRELL** 4

**Symposium Objectives** 5

**Programme** 6

**List of Speakers** 8

**Booklet of Abstracts** 9

- **Trevor G. Bond**, Hong Kong Institute of Education 10
- **Magdalena M. C. Mok**, Hong Kong Institute of Education 11
- **Carla Rampichini**, University of Florence 12
- **Giorgio Vittadini & Piergiorgio Lovaglio**, University of Bicocca-Milan 14
- **Leonardo Grilli**, University of Florence 16
- **Jaap Scheerens**, University of Twente 18
- **Min Yang**, University of London 20
- **David Andrich, Sandra Heldsinger, Stephen Humphry**, Murdoch University, Western Australia 21
- **Jeroen K. Vermunt**, Tilburg University, the Netherlands 23
- **Gage Kingsbury**, Northwest Evaluation Association, US 24
- **Enrico Gori**, University of Udine 27
- **Francesco Bartolucci**, University of Perugia 28

**List of Participants** 30
CRELL

- INTERNATIONAL SYMPOSIUM -
METHODOLOGICAL TOOLS FOR ACCOUNTABILITY SYSTEMS IN EDUCATION

Scientific Committee
Andrea Saltelli
Fulvia Pennoni
Daniele Vidoni

Organising Committee
Sarah Moore
Ute Cecchini
Ulrike Winter

For information related to logistics, and especially for accommodation and transportation, please contact:

Ulrike Winter
Tel: +39 0332 78 63 17
Fax: +39 0332 78 54 09
e-mail: ulrike.winter@cec.eu.int

For information related to Crell please contact:
Sarah Moore
e-mail: sarah.moore@jrc.it

or visit the Crell website:
http://farmweb.jrc.cec.eu.int/CRELL/
CRELL is the result of cooperation between the European Commission’s Directorate General for Education and Culture and Joint Research Centre. It is located at the JRC site in Ispra, northern Italy, in the Institute for the Security and Protection of the Citizen (Unit of Econometrics and Applied Statistics). CRELL brings together a diverse team of educationalists, social scientists, economists and statisticians in a multi-disciplinary approach to education research. It began operation in August 2005 and will reach full capacity by early 2006.

**WORK PROGRAMME**

In line with the needs of the Commission in the field of the monitoring and evaluation of education and training, the CRELL work programme follows four main tracks:

Establishing networks of researchers and institutions active in the field of indicator-based evaluation of education and training systems.

Conducting research on indicators, including scientific analysis to support indicator development in key areas of monitoring of education, e.g. investment efficiency, ICT, mobility, social inclusion, active citizenship, making learning more attractive, learning-to-learn and vocational education and training. This will also include preparatory research to support new data collection and the development of composite indicators in key areas of the knowledge society.

Forecasting and modelling, to quantitatively assess and analyse the impact of education and training objectives on economic performance, and to analyse the impact of important trends on education and training systems. European education and training systems and policy objectives will also be analysed in a worldwide perspective.

Providing support to the Commission in its monitoring and reporting activities, specifically in relation to the current monitoring instrument in the domain of education and training.
THE NEED FOR QUALITY EVALUATION METHODOLOGIES IN EDUCATION AND TRAINING

The first objective of the Lisbon strategy for Education and Training is to increase the quality and effectiveness of education and training systems in the European Union.

Mutual cooperation among EU countries is an essential condition for achieving this objective. The Recommendation of 12/2/2001 on European cooperation in quality evaluation in school education and the Recommendation of 24/9/1998 on European co-operation in quality assurance in higher education explicitly recognise the importance of joint actions in the field of evaluation of education systems. Moreover, the Council Resolution on the Promotion of Enhanced European co-operation in Vocational Education and Training (VET) identifies as crucial the task of “promoting co-operation in quality assurance with particular focus on exchange of models and methods, as well as common criteria and principles for quality in vocational education and training.”

Thus, quality assurance systems have been recognised as an essential part of an effective education and training systems. Techniques that allow quality to be measured are available, although not all countries have the same experience of their use in education and training.

Availability of longitudinal data, reliable measures of student achievement, and the analysis of the hierarchical structure of educational systems are three essential ingredients in the creation of reliable accountability systems in education. Multilevel modelling and Rasch analysis are the two statistical methodologies that provide the greatest insight with respect to this objective. The conference aims to stimulate debate among researchers using these methodologies to increase the quality of existing educational systems.
## PROGRAMME

### MONDAY, 6 FEBRUARY 2006

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.00</td>
<td>Registration and opening remarks:</td>
</tr>
<tr>
<td></td>
<td>Andrea SALTELLI</td>
</tr>
<tr>
<td></td>
<td><strong>TREVOR G. BOND:</strong> Constructing achievement scales: an introduction to the Rasch Model.</td>
</tr>
<tr>
<td>16.30</td>
<td>Coffee break</td>
</tr>
<tr>
<td></td>
<td><strong>MAGDALENA MOK:</strong> Multilevel modelling for accountability systems in education.</td>
</tr>
<tr>
<td></td>
<td>Discussant: Ruggero BELLIO</td>
</tr>
<tr>
<td></td>
<td>Free evening</td>
</tr>
</tbody>
</table>

### TUESDAY, 7 FEBRUARY 2006 - MORNING: STUDENT RESULTS AND THEIR COMPONENTS

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>09.00</td>
<td><strong>CARLA RAMPICHINI:</strong> Multilevel models for the evaluation of university courses.</td>
</tr>
<tr>
<td></td>
<td><strong>GIORGIO VITTADINI &amp; PIERGIORGIO LOVAGLIO:</strong> Human capital growth as a measure of university educational effectiveness.</td>
</tr>
<tr>
<td>10.30</td>
<td>Coffee break</td>
</tr>
<tr>
<td></td>
<td><strong>LEONARDO GRILLI:</strong> Multilevel models for the analysis of the transition from school to work.</td>
</tr>
<tr>
<td></td>
<td>Discussants: Alex MICHALOS &amp; Magdalena MOK</td>
</tr>
<tr>
<td></td>
<td>Coordinator: Fulvia PENNONI</td>
</tr>
<tr>
<td>12.30</td>
<td>Lunch</td>
</tr>
</tbody>
</table>

### TUESDAY, 7 FEBRUARY 2006 - AFTERNOON: INTERNATIONAL COMPARISONS OF STUDENT ACHIEVEMENT

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.30</td>
<td><strong>ANDERS HINGEL</strong> To be defined.</td>
</tr>
<tr>
<td></td>
<td><strong>JAAP SCHEERENS:</strong> The usefulness of international comparative assessment studies for answering questions of educational productivity and effectiveness.</td>
</tr>
<tr>
<td></td>
<td><strong>MIN YANG:</strong> Statistical tools in evaluation of the Gansu Basic Education Project</td>
</tr>
<tr>
<td>16.00</td>
<td>Coffee break</td>
</tr>
<tr>
<td></td>
<td><strong>HEIKO SIBBERNS:</strong> The IEA and the TIMSS project</td>
</tr>
<tr>
<td></td>
<td><strong>ANDREAS SCHLEICHER:</strong> PISA and related international surveys.</td>
</tr>
<tr>
<td></td>
<td>Discussants: Ulf FREDRIKSSON &amp; Daniele VIDONI</td>
</tr>
<tr>
<td></td>
<td>Coordinator: Fulvia PENNONI</td>
</tr>
<tr>
<td>18.00</td>
<td>Bus to hotel</td>
</tr>
<tr>
<td>20.00</td>
<td>Social dinner—bus from hotel</td>
</tr>
</tbody>
</table>
**WEDNESDAY, 8 FEBRUARY 2006 - MORNING: OBJECTIVE MEASUREMENT OF STUDENT ACHIEVEMENT**

<table>
<thead>
<tr>
<th>Time</th>
<th>Speaker</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>09.00</td>
<td>David Andrich</td>
<td>Maintaining an invariant unit in Rasch measurement.</td>
</tr>
<tr>
<td></td>
<td>Jeron K. Vermunt</td>
<td>Multilevel variants of discrete and continuous latent variable models, with an application in education assessment.</td>
</tr>
<tr>
<td>10.30</td>
<td></td>
<td>Coffee break</td>
</tr>
<tr>
<td></td>
<td>Gage Kingsbury</td>
<td>The basics: building accountability systems from the student up.</td>
</tr>
<tr>
<td></td>
<td>Enrico Gori</td>
<td>School effectiveness: empirical evidence and theoretical considerations in the light of Rasch-based longitudinal measures.</td>
</tr>
<tr>
<td></td>
<td>Francesco Bartolucci</td>
<td>Likelihood inference for a latent Markov Rasch model with an application in educational assessment.</td>
</tr>
<tr>
<td></td>
<td><strong>Discussants:</strong> Stefania Mignani &amp; Roberto Ricci</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Coordinator:</strong> Daniele Vidoni</td>
<td></td>
</tr>
<tr>
<td>12.30</td>
<td></td>
<td>Lunch</td>
</tr>
</tbody>
</table>

**WEDNESDAY, 8 FEBRUARY 2006 - AFTERNOON: ROUND-TABLE: TOWARDS A JOINT APPROACH**

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.30</td>
<td>The discussion will be opened by: Daniela Cocchi</td>
</tr>
<tr>
<td></td>
<td><strong>Coordinator:</strong> Daniele Vidoni</td>
</tr>
<tr>
<td>16.00</td>
<td>Coffee break</td>
</tr>
<tr>
<td></td>
<td><strong>Concluding remarks:</strong> Andrea SalteLLi</td>
</tr>
<tr>
<td>18.00</td>
<td>Close</td>
</tr>
<tr>
<td>18.30</td>
<td>Bus to airport / train station</td>
</tr>
</tbody>
</table>
THE SPEAKERS

David ANDRICH Murdoch University, Australia
Francesco BARTOLUCCI University of Perugia, Italy
Ruggero BELLIO University of Udine, Italy
Trevor G. BOND Hong Kong Institute of Education, Hong Kong
Daniela COCCHI University of Bologna, Italy; President, Italian Statistical Society
Ulf FREDRIKSSON JRC - CRELL
Enrico GORI University of Udine, Italy
Leonardo GRILLI University of Florence, Italy
Anders HINGEL European Commission, DG Education and Culture
Gage KINGSBURY Northwest Evaluation Association, United States
Piergiorgio LOVAGLIO University of Milan-Bicocca, Italy
Alex MICHALOS University of Northern British Columbia, Canada
Stefania MIGNANI University of Bologna, Italy
Magdalena MOK Hong Kong Institute of Education, Hong Kong
Carla RAMPICHINI University of Florence, Italy
Roberto RICCI University of Bologna, Italy
Andrea SALTELLI JRC - Head of Unit for Applied Statistics & Econometrics
Jaap SCHEERENS University of Twente, the Netherlands
Andreas SCHLEICHER Head of Indicators and Analysis (Education), OECD
Heiko SIBBERNS IEA Data Processing Centre, Germany
Jeroen K. VERMUNT Tilburg University, the Netherlands
Daniele VIDONI JRC - CRELL
Giorgio VITTADINI University of Milan-Bicocca, Italy
Min YANG University of London, United Kingdom
Booklet of Abstracts

- INTERNATIONAL SYMPOSIUM -

METHODOLOGICAL TOOLS FOR ACCOUNTABILITY SYSTEMS IN EDUCATION

6-8 February, 2006
Joint Research Centre, Ispra
Anfiteatro
Building 36
THE CENTRAL ROLE OF RASCH MODEL INVARIANCE FOR EDUCATIONAL ACCOUNTABILITY SYSTEMS

Trevor G. Bond
Hong Kong Institute of Education

This presentation argues that the important central construct, essential in any educational testing and reporting system is the requirement that each individual child’s test results can be plotted on a calibrated interval scale of achievement. For most in this audience, whose focus on accountability systems for education, a basic premise is that we must be able to compare the performance of one class, school, or region against some national average. PISA sets out to compare nation against nation. But much more fundamental to any accountability system in education is the basic requirement to determine the effect of the engagement of any individual child in the education process. Fundamental to the PISA comparisons, and to any valid and useful conception of an educational assessment system, is a well-maintained, calibrated, interval scale educational. It is also the key attribute of any assessment and reporting system that is to be appropriate for accountability purposes: a scale that has measurement qualities similar to those we take for granted in the metric system of weights, distances, volume and temperatures.

While ease of use, portability, and robustness are expected characteristics of physical measurement devices such as the metre rule and the thermometer, the core underlying property is measurement invariance: measurement scales, and their component measurement units, maintain their value between places and over time. This is central to educational accountability systems which require that the measures must function from place to place, and more importantly for accountability purposes, must remain stable from time to time. Indeed, it is quite counter-intuitive to set about to measure change, if the measure itself changes over time. That is the fundamental challenge to any educational assessment system, and particularly one that focuses on measuring change.

The Rasch model for measurement in the human sciences provides the measurement standards by which we can calibrate the assessment system’s items, tasks and tests as well as the performances of children, classes, schools and regions on those tests. While a Rasch-based measurement system shares many qualities with other commonly used quantitative methods of data recording and analysis, it also has some properties which are unique to it. It is the benefits that derive uniquely from the Rasch model that commend it as the basis for the analytical techniques described by other expert colleagues in this form. For every individual item, test prompt, writing situation or science experiment that a child attempts, there is an estimate of the difficulty of each of those items. And for every child the Rasch model provides an estimate of the underlying ability of that child when confronted with that task, test, or performance opportunity. For every test occasion – for each item, and for each child, there is a precision estimate for each Rasch ability and difficulty estimate that estimates the measurement precision (or tolerance) in the testing system. And the crucial property of measurement invariance relies on the Rasch measurement principle of parameter separation: item and person estimations remain independent of the distribution of those properties amongst persons or items.
Data collected for accountability purposes often involve nested or hierarchical data structure. For instance, in the development of a value-added indicator system, achievement data are collected at the student-level but inferences are to be made at the school and system levels. Here, students are nested within classes, classes are nested within schools, and schools within regions. Since the late 1980s multi-level modelling has been used as a standard approach to handle such nested data structure. Research has shown that failure to apply multilevel models to nested data structure leads to wrong statistical inferences. Furthermore, much valuable information may be lost if the nested data structure is not modelled explicitly.

In the last twenty years, the application of multilevel modelling has been extended to areas other than the handling of nested data structure. Multilevel modelling can now be applied for effective analysis of complicated data structures, including longitudinal data (e.g. growth trajectory), multivariate data (e.g. achievement in the four skills in the learning of English), discrete outcome data (e.g. students’ choice of entry to universities versus work), multiple membership data (e.g. the modelling of students’ mobility across schools in the estimation of the value added by schools to academic achievement), and cross-classified data structure (e.g. secondary schools taking students from a number of ‘feeder’ primary schools).

The presentation will provide a non-technical introduction to multilevel modelling. Examples will be drawn from the value-added indicator system of the Hong Kong Special Administrative Region, and research on self-directed learning using the latest version of the MLwiN software package (Goldstein, 2003). Strengths and weaknesses of multilevel modelling will be discussed in the analyses of these real data.
MULTILEVEL MODELS FOR THE EVALUATION OF UNIVERSITY COURSES

Carla Rampichini
Department of statistics, University of Florence

This contribution presents a methodology for the analysis of student ratings of university courses. Student ratings are an old and widely recognized instrument to evaluate university courses (Emerson, 2000). However, the statistical analysis of student ratings needs special techniques which take into account the ordinal nature of the ratings, the multivariate nature of the data (when the questionnaire includes several items) and the hierarchical structure of the phenomenon (ratings are nested in courses which are nested in schools). Moreover, if one wishes to use the student's satisfaction as a measure of course quality, it should be recognized that the satisfaction of a student, as expressed by the ratings, depends not only on the course characteristics of interest (lecture hall, clarity of the teacher, readings and so on), but also on the student's traits and expectations. Therefore, a fair comparison among courses requires the calculation of net measures that adjust for individual characteristics. One way of obtaining such measures is by means of multilevel models (Goldstein, 2003).

The ordinal two-level model

In course programme evaluations based on student ratings, it is usual to suppose that the observed ordinal measurement \( Y \), with \( K \) levels, comes, through a set of \( K-1 \) thresholds, from a latent continuous \( \tilde{Y} \) following a variance component model:

\[
\tilde{Y}_{ij} = \alpha + \beta x_{ij} + u_j + e_{ij}
\]

(1)

With \( i = 1,2,\ldots,n_j \) elementary units for the \( j \)-th cluster (\( j=1,2,\ldots,J \)). In model (1) \( \alpha \) is the intercept; \( x_{ij} \) is a covariate and \( \beta \) the corresponding slope; the random variables \( e_{ij} \) and \( u_j \) are the disturbances, respectively at the first (subject) and second (cluster) level; and \( \tau^2 \) is the second level variance. In order to make that model identifiable, it is necessary to make some assumption on the latent variable distribution, e.g. normal with first level unit variance.

In the following I consider some papers in which the ordinal two-level model and its stratified and multivariate extension are used to analyse data gathered in the survey of course evaluation carried out by the University of Florence. The survey form has a main section with 26 items concerning the evaluation of various aspects of the course and some students’ characteristics. All the items in the main section require the same type of ordinal response on a four-point scale. The data have a hierarchical structure: respondents are nested in courses that are nested in schools. First of all I consider a single item for classes of a single school, so it is not necessary to adjust for characteristics of the school or the university. The method and the results are fully described in Rampichini et al. (2004).

The second-level residuals of the model are used to point out the ‘anomalous’ courses, both positive and negative. These residuals can be interpreted as a measure of the effect of the course on the evaluations expressed by students, conditional on
the variables inserted into the model; therefore they can be considered as ‘net’ indicators of the course quality with respect to the given item (Goldstein, 2003). The 95% pairwise confidence intervals around the residuals are constructed so that two residuals are considered significantly different from one another if and only if the respective intervals are disjoint (Goldstein, 2003). The ranking obtained from the model should be interpreted with caution, due to the uncertainty of the residuals. The main use of this ranking is to characterize clusters and to find extreme cases.

**Ordinal variance component models for stratified hierarchical data**

When we compare different schools, the student ratings are grouped into strata. If the model parameters are allowed to vary with such strata, a new identification problem comes up, requiring some additional assumption. Grilli and Rampichini (2002) discuss alternative assumptions that overcome the identification problem and illustrate a general strategy for the model selection. The proposed methodology is applied to the analysis of course programme evaluations based on student ratings, referring to three different schools of the University of Florence. Results show that both the latent average evaluation of the courses and the measurement scale (i.e. the model thresholds) vary with the school, suggesting that one must be careful when interpreting the raw ratings based on an ordinal scale.

**Multidimensional indicator of course quality**

The quality of a course (or the set of courses for a given degree programme, school or university) is a multidimensional concept: multiple factors contribute to form the final teaching result of a class. This aspect can be taken into account by devising a multidimensional indicator of course quality, based on a battery of items that characterize the quality of a course, such that the information provided is not redundant and is sufficiently explanatory of the general level of satisfaction in the class. Rampichini et al. (2004) proposed a two-step process, while Grilli and Rampichini (2003) present a joint analysis of the items by means of a multivariate multilevel model. A promising route for the construction of a unidimensional scale is the integration of Rasch and multilevel modelling (Fox and Glas, 2001).

**References**


HUMAN CAPITAL GROWTH AS A MEASURE OF UNIVERSITY EDUCATIONAL EFFECTIVENESS

Giorgio Vittadini, Pietro Giorgio Lovaglio
Department of Statistics, University of Bicocca-Milan

One of the most widely accepted currents of thought considers internal effectiveness and efficiency to be a fundamental tool for the self-evaluation of universities, whereas it suggests the use of external effectiveness and efficiency for a true university evaluation, comparative and longitudinal (Elias, 2002). As of today, such a system does not yet exist, and although there are numerous studies regarding the external effectiveness of education, there are very few that provide an analysis of external efficiency, i.e., having as a primary objective the evaluation of investment in higher education in terms of the monetary return (as earned income) to graduates over their life-cycle.

The present paper aims to furnish an initial contribution in this direction, utilizing the criteria of evaluation studies about organizations that distribute services of public utility. The evaluation of external efficiency should be based more on investment in higher education than on the earned income of graduates (Garen, 1984), or, in other words, on Human Capital (HC), defined as an individual’s expected earned income, relative to the skills and abilities acquired through education (Becker, 1964). Nevertheless, the evaluation of university education in this context does not coincide with the problem of HC estimation, but with the two following aims: first, the evaluation “ceteris paribus” of what universities furnish to their graduates as investment in HC with returns in terms of earned income (relative efficiency); and second, the evaluation of whether university education contributes to a real advantage in monetary terms (impact of efficiency).

This paper proposes a method for the evaluation of relative efficiency and impact of efficiency in universities, defined as the effects of higher education on the long-term income of university graduates. The nature of the problem, the debates in the literature, the need to evaluate different universities “ceteris paribus,” suggest the utilization of Multilevel Longitudinal Model with random effects.

From the methodological point of view, in order to overcome the problem of “selection bias” (Garen, 1984) due to the non-randomization of the individuals between treatments, the evaluation of effectiveness and efficiency must be addressed "ceteris paribus,” adjusting the “outcome” by the effects of individual characteristics, resources of universities, local markets where graduates find jobs (Fitz-Gibbon, 1997; Scheerens and Bosker, 1997). To this aim, even if statistical literature addresses linear models, a debate has recently been raised regarding the nature of the dependent variable in these evaluation models.

First of all, after having shown the major drawbacks of the classical approaches based on the gain score or the added value (residual gain score), we justify the methodology that utilizes outcomes measured in a longitudinal framework (time series), an approach that allows insertion of factors, as covariates, linked to the insti-
tutions (schools, universities) that can affect the performances of the micro-units (student, class), as recent studies have demonstrated (Thum, 2002; Bryk et al., 1998).

The advantage of multilevel models is that they have helped accountability agencies avoid the misleading picture conveyed by computing score aggregates such as the school or district means and to focus on student instead (level and/or change) (Meyer, 1996). Many current accountability systems employ some form of the multi-level model (Bryk, et al., 1998).

The traditional model is extended (Lovaglio, 2004) to the stage where the interest of the stakeholders is addressed not only to the evaluation of efficiency among a set of institutions (relative evaluation) but also the efficiency of the treatment (distributed service) regarding the non-intervention (impact evaluation); in this case the model also allows each institution that distributes a service (higher education) to see whether their clients obtain better performances as compared with a control group composed of subjects with similar characteristics to whom the service was not been distributed.

The application proposed concerns an external efficiency study of Italian universities using the Bank of Italy Survey (1998, 2000, 2002); nevertheless, because of the low number of "panel graduates", it is not possible to deepen the analysis of the external efficiency of the higher education relative to the university’s effect or the effect of the type of degree; the analysis is therefore limited to the effect of the higher education tout court on the dynamics of incomes (impact analysis that compares the level and the growth rates of earnings between graduates and non-graduates in Italy) and to a limited set of type of degree.

In particular the results show that:

- a great part of the within-subjects’ income variability is determined by individual characteristics that change over time and to a lesser extent by the temporal dynamic of incomes;
- the differences between means of income in every temporal moment and for the growth rates has been estimated for the two groups: they show that for all three temporal moments the differences between income means are not only meaningful, but increase in time, and also that the income growth rate of graduates is higher than that of non-graduates, even if near to the limits of the customary levels of significance;
- to quantify the extent to which the labor market rewards university education we confront the growth rates of incomes between a graduate and a non-graduate in the age class under 30 years: the impact of the Bachelor degree appears highly significant in terms of absolute value (increasing over time), while in relative terms (in comparison with the non-graduate rates) this difference remains significant, but constant in the older age groups;
- medicine and law show growth rates of incomes significantly higher than mean growth rate for graduates.
MULTILEVEL MODELS FOR THE ANALYSIS OF THE TRANSITION FROM SCHOOL TO WORK

Leonardo Grilli
Department of Statistics, University of Florence

The analysis of the transition from school to work plays a crucial role in the evaluation of higher education institutions, since one of their aims is to endow graduates with the appropriate skills for the labour market. The analysis can focus on various aspects, e.g. the employment status on a given date, the time needed to find the first job, the degree of consistency between the current job and the curriculum, satisfaction in the current job.

The exact definition of the aim of the analysis depends on the purpose of the evaluation and ultimately on the policy objectives: for example, in the last few years there has been a tendency to narrow the definition of ‘employed graduate’ by considering only jobs which are permanent or consistent with the curriculum. Moreover, a key point in the analysis of employment rates is the definition of the denominator, i.e. the criteria to identify the subset of graduates interested in finding a job (this is crucial to compare degree programmes whose graduates have markedly different propensities to keep on studying versus searching for job).

Whichever aspect is studied (in statistical terms the “outcome variable”), the phenomenon has a relevant hierarchical structure (graduates nested in degree programmes nested in universities and so on). Such a structure calls for multilevel modeling, which allows analysis at all the hierarchical levels at the same time. Multilevel models are especially suited for comparing effectiveness (Goldstein and Spiegelhalter, 1996), as the relative effectiveness of an institution is explicitly represented in the model as a random effect (a latent variable at the institution level). Even if multilevel models represent a theoretically satisfactory tool for the assessment of educational institutions, their implementation faces serious problems such as misspecification due to omitted variables and low power in ranking the institutions.

In this contribution I review my work on the time taken to obtain the first job. Other studies on graduate employment are Grilli and Rampichini (2004) (a multilevel polychotomous response model for studying the question of where the skills needed for the current job have been acquired) and Grilli and Rampichini (2005) (a multilevel factor model for studying satisfaction in various aspects of the current job).

Survival models for the time taken to obtain the first job are potentially more informative than logit or probit models for the employment status on a given date, since they also give insight into the dynamics of the process, allowing explanatory variables with time-varying effects and even time-varying explanatory variables (though, unfortunately, they are rarely available in datasets).

The time taken to obtain the first job is usually reported in months so it requires discrete-time survival models (continuous-time models are inadequate due to the large number of ties). Discrete-time survival models and multilevel models each have a long history, but their joint use is quite recent (Barber, Murphy, Axinn and Maples, 2000).
Biggeri, Bini and Grilli (2001) analyse a large dataset (13,511 individuals) from a survey on job opportunities of Italian graduates of 1992 conducted by the Italian National Statistical Institute (Istat) in 1995. The data makes it possible to determine, for each graduate, the time needed to find the first job (in months) or the censored time for those still unemployed at the date of the interview. The graduates are nested in course programs which are grouped into universities, so that the dataset has a hierarchical 3-level structure. The adopted model is a 3-level logit survival model.

The available explanatory variables include demographic features, family background, educational history and work experience. They are all measured at the graduate level and at a fixed time point, so the potentialities of the model are only partially exploited. Nevertheless, the application reveals some interesting dynamic patterns, notably: the hazard of getting job decreases as time elapses (except for the males involved with the military service); and the female disadvantage in getting a job vanishes as time elapses. Plots of residuals at the university and course programme levels, along with their confidence intervals, are used to compare effectiveness.

In a later work Grilli (2005) considers data from a survey on the high school graduates of year 1995, carried out by the Italian National Statistical Institute (Istat) three years later. The graduates are nested in schools, so a 2-level model is adopted. Two grouped-time versions of Cox proportional hazards model are compared, showing the higher flexibility of the so-called ‘continuation ratio’ version.

The set of available explanatory variables is wide, including many variables at the graduate level and a few variables at the school level (technical vs. gymnasium, public vs. private). Moreover, an attempt to control for local labour market conditions is made by using an external source to derive further explanatory variables for the regional-level youth unemployment rate and its trend. Compared with university graduates, for high school graduates the hazard of getting job has a more complex pattern, since many such graduates do not immediately search for a job, but are involved in training or educational activities (often of short length).

References
THE USEFULNESS OF INTERNATIONAL COMPARATIVE ASSESSMENT STUDIES FOR ANSWERING QUESTIONS OF EDUCATIONAL PRODUCTIVITY AND EFFECTIVENESS

Jaap Scheerens
University of Twente

International comparative assessment studies are particularly useful for assessing the productivity of education systems, in terms of average achievement in a specific subject matter area or literacy domain. Countries can pick and choose the benchmarks they would like to use, in order to compare themselves: the international average, the score of a neighbour country or the average of the highest scoring country. As the illustrative data from PISA and TIMSS show, large differences exist between the highest and the lowest scoring countries. For resource-poor countries this might be problematic, because students might feel discouraged in not being able to do a substantial part of the items. At the same time it could be seen as important that such international comparisons can be made. A possible solution might be to expand the difficulty range in the sense of including sets of easier items for countries that are expected to score relatively low. If tests confirm to the assumptions of item response models, these easier item sets could then be equated to the general international tests. International assessments for specific regions, like the PASEQ and SACMEC studies in Africa, and the Primer Estudio Internacional Comparativo in thirteen Latin American countries, have the advantage of being able to adapt the difficulty level of the achievement tests, and include perhaps more environmentally valid items in the context of schooling in resource-poor countries.

Not only achievement levels such as the country averages are useful, but also the patterns of variability that the score distributions of international comparative assessments show. As will be illustrated, interesting conclusions can be drawn on the basis of the total between-school variance, the proportion of the variance that is between schools, (usually indicated as the between-school variance), and sometimes also the variance between parallel classes in one school. If the data can be broken down according to regions within countries, such analyses of the patterns of variability gain in relevance. Variability measures provide insights into the inequality among students in their achievement results, the degree of segregation of the system of schools and practices like ability grouping and streaming within schools. In theory it would also be feasible to set benchmarks for keeping the different types of variability of and within school systems within limits. Such benchmarks would address the interpretation of educational quality in terms of equity.

Comparing the impact of malleable school variables on the one hand, and student background conditions and composition factors on the other, reveals something about the margins of control and change in education. On further reflection these different effects can be related to two different strategies to influence outcomes: productivity improvement on the one hand, and selection and admission policies on the other. The size of the composite effects, as will be illustrated on the basis of the PISA data set, may call more attention to “selection management” and establishing fixed “quota” of students with specific background characteristics. It cannot be discounted that the impact of student background and compositional factors is over-
rated in international comparative assessment studies, because of weaknesses in the operationalization of the school variables. Besides, as will also be illustrated, the two types of factors overlap in their impact on achievement, which further complicates the interpretation. In any case, do international assessments provide the occasion to examine globally the margins of “malleability” in schooling, as well as the degree of dependency of results on student background characteristics and their aggregates? The latter provide an additional interpretation relevant for the equity perspective, implying that systems in which achievement results depend to a larger degree on “given” student background conditions like their socio-economic status are considered to be less equitable than systems for which this association is lower.

The global international assessment studies from IEA and OECD have yielded relatively disappointing results with respect to confirming the effectiveness-enhancing factors that are part of the school effectiveness knowledge-base. This applies both to the size of the association of these variables with performance, after controlling for student background conditions, and to the weak consistency of the significance of the effects of these variables across countries. Regional studies, like the Latin American PEIC, however, do show results that are more in line with results of school effectiveness research studies. One way of improving the relevance of international comparative assessment studies to questions of educational effectiveness would be to invest more in measuring school factors and processes, using more extensive scales and perhaps also to conduct classroom observations. Another alternative would be to consider stand-alone school surveys and classroom observation studies to yield information on effectiveness-enhancing process indicators. An example is the school and teacher survey in the countries united in the World Education Indicator Project of UNESCO, the OECD and the World Bank.

Methodological advances in applying stronger psychometric models (IRT modelling) and analysis techniques (multipath) would be beneficial in the following areas:

- better control of unreliability in the independent variables (both adjustment variables like SES, and malleable variables, e.g. educational leadership at school);
- allowing for more secure interpretations of compositional effects;
- providing the appropriate statistical models for investigating more complex causal models where independent variables are defined at different aggregation levels;

For the international studies, however, improvements in the study designs would be even more important than the above improvements in psychometric and statistical modelling. For OECD’s PISA project, for example, grade- and class-based samples and a longitudinal design would be very important to meet countries’ interests in having clearer answers to questions of educational effectiveness.
STATISTICAL TOOLS IN EVALUATION OF THE GANSU BASIC EDUCATION PROJECT (GBEP)

Min Yang
University of London

The GBEP is the largest project sponsored by the UK Department of International Development (DFID) in the poorest region of China. The first phase was launched in early 2000, to run for six years. It covers four rural counties including over 60% of Muslim minority residences in the Gansu province. The overall goals of the project are to improve the net enrolment rate for all primary school children, particularly girls, to reduce dropout rates, and to get most children to complete nine years of basic education. There are six specific outcomes: 1) Improved school conditions; 2) Increased participation of poor and disadvantaged children, girls, and minority groups; 3) Improved teacher and pupil performance; 4) Strengthened capacity of teacher-training institutions; 5) Improved planning and management at all levels of the education system; and 6) Educational research influencing policy and practice in basic education. Scientific evaluation of the project at different stages is vital for the Project Managers to ensure that the project moves towards its final goals, and a major concern for the DFID is to ensure that the money be well spent at all times. It is also an important mechanism for capacity building and management training in the local education management system. This presentation covers the overall design of the project evaluation, the role of quantitative and qualitative statistics and, more specifically, the longitudinal design and statistical models of student achievement for examining the overall effects of the project on children’s learning.
MAINTAINING AN INVARIANT UNIT IN RASCH MEASUREMENT

David Andrich, Sandra Heldsinger, Stephen Humphry
Murdoch University, Western Australia

The distinctive feature of the Rasch models for measurement is that if the responses conform to the chosen Rasch model within a specified frame of reference of a class of persons and a class of items, then the comparison in locations among persons is independent of which subset of items is chosen for measurement, and the comparison in locations among items is independent of which persons are chosen for item calibration.

This principle of invariance is used in equating for the study of change, either of different cohorts of the same class of students, for example Year 5 students of two different years in the same education system, or Year 5 students in one year with Year 7 students in the same year of the same education system, or indeed for comparisons of students across education systems. In that case some common items might be used in the two testing regimes, and if (i) the common items show invariance across the two cohorts of students, and (ii) if the remaining items conform to the common items, then the measurement of the two cohorts can be deemed to be in the same unit of scale.

The unit of scale has two components generally not considered. First, there is a natural unit of measurement which is integral to a frame of reference, that is, for a test composed of items with respect to the group of persons. This unit reflects the precision of measurement. Second, there is an arbitrary transformation of this natural unit to an arbitrary unit which is implied in the Rasch model. If both conditions listed in the last paragraph hold, then these two features are taken into account implicitly. However, there are circumstances in which such conditions cannot be expected to occur by sound item construction and be simply confirmed by checking the invariance of common item locations. This paper shows that if these two features of the measurement with the Rasch model are ignored in such circumstances, they can create serious misinterpretations in the assessment of change.

The circumstances considered in this paper concern studying change in relation to specified benchmarks on a scale. For example, there might be a desire to set a benchmark, and to check if a subsequent cohort at the same Year level has a greater or smaller proportion of students below a specified benchmark, or to compare the proportions of two different Year levels above the specified benchmark.

In setting benchmarks it is necessary to conduct judgments independently of the actual measurements. Generally, in such judgments, this matter of unit of scale is generally ignored. This paper shows with a concrete example that the natural unit of scale, and the arbitrary transformation implied in the Rasch model, cannot be ignored.

This paper uses two data collection designs concerned with setting benchmarks and shows that the two designs have a very different unit of scale. The first is the a modified Angoff procedure which requires judges to estimate the probability that a
benchmark student will succeed on every item of a test, and the second is a pair comparison procedure in which items, including items deemed to be benchmark items, are compared in pairs regarding their relative difficulty. Not only does this show that the unit of scale from the two data collection designs are different, but also shows that the unit of scale is very different from the unit of scale that is obtained from the responses of the students to the same items. The paper also shows that if the unit of scale is not taken into account then the two approaches lead to very different benchmarks with significant misinterpretations. It is shown that by taking account of the different units, the benchmarks are much more closely aligned from the two approaches.

Three other points are made in the paper. First, the study of unit seems to explain two important observations in the literature – (i) that judges are not very accurate in predicting at an absolute level the probabilities that students will succeed on items; (ii), that judges are reasonably accurate in predicting the relative difficulties of the items for the students.

Second, although in this study where the data collection formats are explicitly and evidently different, it is readily understood that the unit of scale, which reflects the precision, might be different in the two formats, and different from that obtained from the student responses. However, the unit of scale can be an issue in circumstances that are more subtle where the formats are not so clearly different. Examples of such circumstances for studying growth are considered in the paper.

Finally, as indicated in the first part of the abstract, the unit of scale is not generally considered in Rasch measurement with the focus on the invariance of relative locations of the items. The paper stresses, however, that every analysis of the responses of a group of students to a set of items in principle has its own natural unit, and that these responses are transformed with an arbitrary transformation to locations on the scale. Taking account of both of these issues is potentially necessary in vertical equating, and generally necessary for the study of growth in learning. Taking account of the arbitrary unit of measurement in the Rasch model should lead to better and more stable measurement in monitoring and understanding learning.

Acknowledgements
The research reported in this paper arises from the Doctoral studies of Sandra Heldsinger and Stephen Humphry at Murdoch University and the data were made available by the Department of Education and Training in Western Australia. The research is supported in part by an Australian Research Council grant with the following industry partners, the national MCEETYA Performance Measurement and Reporting Task Force, IIEP (UNESCO) and the Australian Council for Educational Research (ACER).
MULTI-LEVEL VARIANTS OF DISCRETE AND CONTINUOUS LATENT VARIABLE MODELS, WITH AN APPLICATION IN EDUCATION ASSESSMENT

Jeroen K. Vermunt
Department of Methodology and Statistics, Tilburg University, the Netherlands

I will present a framework for multilevel latent variable modelling as (partially) implemented in the Latent GOLD software package. This framework includes models with discrete and continuous latent variables (LVs), as well as combinations of these.

One of the special cases, in which both the lower- and higher-level LVs are discrete, is the hierarchical variant of the latent class (LC) model proposed by Vermunt (2003). More specifically, lower-level units (cases) are clustered based on their observed responses as in a standard latent class model, whereas higher-level units (groups) are clustered based on the likelihood of their members being in one of the case-level clusters. Applications include repeated measures and three-way data collected via panel or experimental designs, as well as multilevel data, such as from pupils nested in schools, employees nested in firms, citizens nested in regions, and consumers nested in stores.

Another variant is the multilevel IRT model by Fox and Glas (2001), in which both the lower- and higher-level LVs are continuous and the response variables are discrete. Putting together all possible combinations, we get the following scheme:

<table>
<thead>
<tr>
<th>Higher-level LVs</th>
<th>Lower-Level LVs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>I. Multilevel random-effects IRT / FA</td>
<td>II. Multilevel random-effects LC</td>
</tr>
<tr>
<td>Discrete</td>
<td>III. Multilevel mixture IRT / FA</td>
<td>IV. Multilevel mixture LC</td>
</tr>
</tbody>
</table>

Whereas variants I, II and IV have already been explored, variant III – in which the lower-level LVs are assumed to be continuous and the higher-level LV is discrete – has not been studied. This specific combination gives rise to a mixture IRT or FA model in which the mixture components are formed by groups rather than by cases, as would be the case in a standard mixture IRT (say mixture Rasch) or mixture FA model; in other words, groups are clustered based on the trait/factor means and variances of their members, and possibly on (other) model characteristics that may differ across groups, such as item difficulties and discriminations. Overall, this seems to be a good way to identify group differences as well problematic groups when the number of groups is too large for a standard multiple-group analysis.

ML estimation of the multilevel LV models by means of the EM algorithm is straightforward, but requires a special implementation of the E step of the EM algorithm. The E step of the algorithm is similar to the well-know forward-backward algorithm for hidden Markov modelling.

Besides discussing the framework and estimation issues, I will present an empirical application from the educational field, in which I will focus on the unexplored, but rather interesting, variant III from the above table.
As long as schools have existed, we have asked that they be accountable for what they teach students. In ancient Athens, Socrates was accused of corrupting his students, and was later convicted of crimes against the state and sentenced to death. While consequences aren’t quite that serious for teachers today, we live in societies that are constantly trying to identify whether schools are doing a good job of preparing their students for life as adults. Beyond identifying good teaching, there seems to be more than a little competition to be the best school system, as evidenced by the popularity of recent studies of student performance in different countries.

Approaches to school accountability have commonly suffered from being developed in a top-down fashion, to meet specific but very narrow goals. If we want to create a system for educational accountability that identifies and promotes good teaching practice we need to consider developing a system that starts with the student and moves out to the rest of the educational community. This paper suggests elements of such a system, and describes examples of each of these elements as they exist today.

Equi-Precise Measurement
If we measure each student accurately, the development of an accountability system is a straightforward process. Accurate measurement at the student level can roll up to the classroom level, the school level, and beyond. While it doesn’t help us determine what should be measured and how it should be measured, the very intent to measure each student’s capabilities with the same high level of accuracy improves our ability to create a high-quality accountability system.

Common test development practices do not result in equi-precise measurement. In a traditional test a common set of items is given to all students in a grade level. This design causes problems with respect to measurement and also with respect to motivation. With a traditional test, it is common for the information obtained for extreme students to be less that 25% of the information obtained for students in the middle of the distribution. This means that instructional decisions made for high and low performing students are substantially less accurate than those made for students near the average. It also means that high and low performing students are either frustrated of bored by most of the content on the test. The result of this type of test design is less-than-accurate measurement and less-than-optimal motivation for many of the students taking the test.

A simple solution to this difficulty is to change to an adaptive test design. An adaptive test with a reasonably deep item pool aligned to the content standards of interest can provide equi-precise measurement for virtually all students tested. In addition, it can provide a challenging test for all students, rather than just those in the centre of the achievement distribution.
**Stable Measurement Scales**

Another element that fosters the development of a strong accountability system is the presence of stable measurement scales. If we want to identify how effective a school is from one year to the next, we need to be able to say that our measures can be meaningfully compared from one year to the next. This is such a straightforward tenet of good measurement practice that it is surprising how few accountability systems make it a stated requirement. Over the past decade, a wide variety of agencies have had difficulties comparing scores from one year to the next because simple measurement practices were not followed.

One simple way to assure that measurement stability is maintained from one year to the next is through the use modern test theory to create a measurement scale that can be used with a wide variety of tests. The Rasch model and other IRT models allow this type of scale development, if used carefully. Much as in the building of a suspension bridge, the development of a measurement scale requires substantial support and cross-bracing, to assure that each portion of the scale is consistently strong and able to stand the strain it will encounter from one year to the next. In addition to providing stability, such a scale allows a test score to be directly related to the content that is challenging to the student at the time.

**Growth and Status**

Commonly, measures of status (proficiency) have been used within accountability systems. These approaches have the characteristic that differences in initial status influence the accountability outcome. To the extent that some schools have students that come to them with different preparation and interest in academics, the use of student status as a measure of school effectiveness is not recommended. It is fairly clear that a strong accountability system would set challenging, individualized growth targets for each student and would judge success by how well the targets are met. Examples of this type of approach are seen in a number of settings. The stumbling block for this approach is the accurate measurement of growth. The solution for this is partially statistical and partially psychometric. With tests that are accurate for each student, and with appropriate statistical models for growth, we can identify growth targets and measure how well each student can reach them.

**Consequences that Help Students**

An accountability system that is designed to measure the growth of students accurately on a stable measurement scale provides a firm foundation for accountability. The next question to be considered is how to use the information obtained from the accountability system. In general, accountability systems have been designed around systems of punishment and reward. If a school does poorly, it may be identified for sanctions, loss of funds, or closure. If a school does well, it may be identified for special recognition, additional funding, or merit bonuses. While reward/punishment systems may be effective in some circumstances, they aren’t the most useful systems to move education forward.

An alternative approach to use the information in an accountability system is to use it as a place to start to examine how well each student is being served by his/her school. If we have measured the growth of every student well, we should be able to identify where the deficits and strengths are in each school, and take action where it is needed to help students. Some of the questions that can be considered with this
approach include:

- Are students in all grades meeting their growth targets?
- If not, which students are missing their growth targets?
- Are students being given the instruction that is appropriate for them, based on their current achievement and their growth history?
- Do teachers have the materials they need for each of their students?
- Is the school in need of additional staff to enable learning?
- Should the makeup of classes in the school be changed to enhance learning?

Is the distribution of staff appropriate across schools?

In this approach, there is no overriding policy of reward or punishment. Instead, the goal of the accountability system is to enable and increase student learning. Over time, the growth and performance of students in schools can be monitored to identify those interventions that are most effective in improving and maintaining student growth. With this type of approach, we should be able to improve schools gradually, and improve their effectiveness with all of their students, rather than just a few.
SCHOOL EFFECTIVENESS: EMPIRICAL EVIDENCE AND THEORETICAL CONSIDERATIONS IN THE LIGHT OF RASCH-BASED LONGITUDINAL MEASURES

Enrico Gori
Department of Statistics, University of Udine

In modern societies, the role of governments in education is a matter of extensive debate and research (Glenn, De Groof, 2004); understanding which policies are fit-test for improving school quality leads public attention to swing from resources (Hanushek, 2003) and rigid control of the state over the organization of education to accountability (Hanushek & Kimko, 2000; Unesco, 2005; Bishop & Woessmann, 2001). With respect to the latter, recent studies have shown that school quality, as measured by test scores in basic competences (literacy and mathematics), is directly related to individual earnings, productivity and economic growth. The same studies underline that school quality is better in institutional settings characterized by strong accountability systems and by school autonomy in planning, staff hiring policies, and drafting of the curricula. The depicted landscape suggests the creation of quasi-markets in education where a plurality of providers manages the educational provision, while the state (which may still be one of the providers) finances the service provision and controls its quality. However, two inefficiencies hinder this approach. The first inefficiency is organizational: the state’s inherent organizational complexity and slowness results in an inability of the system to adapt to individual needs. The second inefficiency is structural: schooling affects the whole population, and if the state alone is responsible for the quality and the characteristics of the service, schooling will necessarily be shaped around the state’s vision of education. Even if it aims at the common good, the state is still one specific institution whose influence on the service may result in a bias in favour of some specific points of view. Thus, an alternative is necessary.

Preparatory to any alternative accountability structure, however, some issues need to be solved in order to understand the characteristics of the obtainable data and the analytical tools that may best fit its treatment:

School objectives. Knowledge and skills are often negatively considered only “one among many other” objectives of the educational systems.

Relative weakness of test scores in terms of measurement theory. Theoretically, testing practices favour some children or some specific groups and are not “valid” or “reliable” because they do not take into account crucial information on the dimensions that are being measured. In this regard, only Rasch analysis provides reliable measures of the degree of uncertainty of the estimated variables, and such information is crucial to correctly evaluate the correlations and, through these, assess the external validity of the constructed measures.

Negative consequences. Allowing test scores to have a bearing on students’ later careers may have negative consequences, by, for example, focusing the students’ efforts on passing the test rather than on learning (also known as learning/teaching to the test).

The present contribution examines these three issues, their counter-arguments and their possible solutions, and envisages an accountability structure responsive to the stakeholders rather than to the state but still producing data on student achievement comparable in space and time.
LIKELIHOOD INFERENCE FOR A LATENT MARKOV RASCH MODEL WITH APPLICATION TO EDUCATIONAL ASSESSMENT

Francesco Bartolucci
Dipartimento di Economia, Finanza e Statistica, University of Perugia

We illustrate a version of the Rasch (1961) model for longitudinal data based on the assumption that any occasion-specific response variable depends only on a discrete latent variable, which in turn depends on the latent variables corresponding to the previous occasions according to a first-order Markov chain. The model is in practice a constrained version of the latent Markov model which was introduced by Wiggins (1973) and has been successfully applied in several fields (for a review see Langeheine and van de Pol, 2002). The model may also be seen as an extension of the latent class version of the Rasch model (see, among others, Lindsay et al., 1991), in which a subject is allowed to move between latent classes from one time occasion to another according to a transition matrix.

For the model above we illustrate maximum likelihood estimation, based on an EM algorithm (Dempster et al., 1977), and how to compare it with the latent class Rasch model through a likelihood ratio statistic. Note that this is equivalent to testing that the transition matrix of the latent Markov model is diagonal. The asymptotic distribution of this statistic under the null distribution is not of $\chi^2$-type as in standard cases. As shown by Bartolucci (2006), it is instead of $\bar{\chi}^2$-type, a well-known distribution in constrained statistical inference (Silvapulle and Sen, 2004). The case in which the transition matrix of the assumed model has all the off-diagonal elements equal to each other is discussed in detail. In this case, the asymptotic distribution is simply

$$0.5\chi^2_0 + 0.5\chi^2_1.$$

The test above may be used in at least two ways in education assessment. The first is for checking that the basic assumptions of the Rasch model are met, in particular for what concerns the assumption of local independence. According to this assumption, in fact, the variables representing the responses to the test items of the same subject must be conditional independent given his/her ability (or latent class). Implicit learning phenomena are therefore excluded as well as situations in which an item provides clues for responding to other items. So, if for a certain set of items administered to a sample of subjects we find that the latent class Rasch model has to be rejected in favor of its latent Markov version, we can conclude that local independence is not a plausible assumption for this set of items. This kind of test will be illustrate through an application to certain National Assessment of Educational Progress (NAEP) data (see also Bartolucci and Forcina, 2005).

The second possible use is for the analysis of data deriving from testing the same subjects at different occasions in order to measure their progresses in certain matters. In this situation we can have several response variables for any occasion and so the univariate latent Markov model above may not be directly applicable. The
extension of the univariate model to this case is therefore outlined. It turns out that likelihood inference of the extended model may still be made on the basis of the results of Bartolucci (2006). In particular, maximum likelihood estimation may be performed through the same EM algorithm outlined for the univariate model, with minor adjustment. Similarly, the likelihood ratio statistic for testing that the transition matrix is diagonal, and so transition between latent classes is not allowed, has still asymptotic distribution under the null hypothesis.

References
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
<th>Address</th>
<th>City</th>
<th>Phone</th>
<th>Fax</th>
<th>Email</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fabio AIELLO</td>
<td>Università degli Studi di Palermo</td>
<td>Viale delle Scienze 13</td>
<td>Palermo</td>
<td>+39 091 662 6303</td>
<td>+39 091 426 452</td>
<td><a href="mailto:fabiello@inwind.it">fabiello@inwind.it</a></td>
<td></td>
</tr>
<tr>
<td>Sara ALVAREZ</td>
<td>Dirección General de Ordenación Académica</td>
<td>Plaza de España 5-4 Planta</td>
<td>Asturias</td>
<td>+34 985 668 136</td>
<td>+34 985 668 138</td>
<td><a href="mailto:saraa@princast.es">saraa@princast.es</a></td>
<td></td>
</tr>
<tr>
<td>Ignacio ALVAREZ-PERALTA</td>
<td>Instituto Nacional de Evaluación y Calidad del</td>
<td>C/ San Fernando del Jarama, 14</td>
<td>Madrid</td>
<td>+34 91 745 9200</td>
<td>+34 91 745 9249</td>
<td><a href="mailto:ignacio.alvarez@mec.es">ignacio.alvarez@mec.es</a></td>
<td></td>
</tr>
<tr>
<td>David ANDRICH</td>
<td>Murdoch University</td>
<td>South Street</td>
<td>AUS - 6150</td>
<td>+61 8 360 2245</td>
<td>+61 8 360 6280</td>
<td><a href="mailto:D.Anrich@Murdoch.edu.au">D.Anrich@Murdoch.edu.au</a></td>
<td></td>
</tr>
<tr>
<td>Barbara ANTONIOLI</td>
<td>Università della Svizzera Italiana</td>
<td>Viale Buffi 13</td>
<td>Lugano</td>
<td>+34 955 423 7474</td>
<td></td>
<td><a href="mailto:barbara.antonioli@lu.unisi.ch">barbara.antonioli@lu.unisi.ch</a></td>
<td></td>
</tr>
<tr>
<td>Silvia BACCI</td>
<td>Università degli Studi di Firenze</td>
<td>Viale Morgagni 59</td>
<td>Firenze</td>
<td>+39 0332 78 9314</td>
<td>+39 0332 78 5733</td>
<td><a href="mailto:s.bacci@ds.unifi.it">s.bacci@ds.unifi.it</a></td>
<td></td>
</tr>
<tr>
<td>Francesco BARTOLUCCI</td>
<td>University of Perugia</td>
<td>Via A. Pascoli, 20</td>
<td>Perugia</td>
<td>+39 075 585 5202</td>
<td>+39 075 585 950</td>
<td><a href="mailto:bart@stat.unipg.it">bart@stat.unipg.it</a></td>
<td></td>
</tr>
<tr>
<td>Ruggero BELLIO</td>
<td>Università di Udine</td>
<td>Via Treppo 18</td>
<td>Udine</td>
<td>+39 0432 249 595</td>
<td>+39 0432 249 595</td>
<td><a href="mailto:ruggero.bellio@dss.uniud.it">ruggero.bellio@dss.uniud.it</a></td>
<td></td>
</tr>
<tr>
<td>Giovanni BOSCAINO</td>
<td>Università degli Studi di Palermo</td>
<td>Viale delle Scienze 13</td>
<td>Palermo</td>
<td>+39 091 662 6303</td>
<td>+39 091 426 452</td>
<td><a href="mailto:gioboscaino@unipa.it">gioboscaino@unipa.it</a></td>
<td></td>
</tr>
<tr>
<td>Michela BATTAUZ</td>
<td>Università degli Studi di Udine</td>
<td>Via Treppo 18</td>
<td>Udine</td>
<td>+39 0432 249 592</td>
<td>+39 0432 249 595</td>
<td><a href="mailto:battauz@dss.uniud.it">battauz@dss.uniud.it</a></td>
<td></td>
</tr>
<tr>
<td>Trevor BOND</td>
<td>Hong Kong Institute of Education</td>
<td>10 Lo Ping Rd.</td>
<td>Hong Kong</td>
<td>+852 2948 8473</td>
<td>+852 2948 7983</td>
<td><a href="mailto:tbond@ied.edu.hk">tbond@ied.edu.hk</a></td>
<td></td>
</tr>
<tr>
<td>Eugenio BRENTARI</td>
<td>Università di Brescia</td>
<td>Contada S. Chiara 50</td>
<td>Brescia</td>
<td>+39 030 298 8633</td>
<td>+39 030 240 0925</td>
<td><a href="mailto:brentari@eco.unibs.it">brentari@eco.unibs.it</a></td>
<td></td>
</tr>
</tbody>
</table>
Gasper CANKAR
National Examinations Centre
Ob Zeleznični 16
SLO - 1000 LJUBLJANA
tel.: +386 154 84 682
fax: +386 1 54 84 601
e mail: gasper.cankar@ric.si

Vincenza CAPURSI
Università degli Studi di Palermo
Viale delle Scienze 13
I - 90128 PALERMO
tel.: +39 091 662 6284
fax: +39 091 426 452
e mail: capursi@unipa.it

Maurizio CARPITA
Università degli Studi di Brescia
Contadina S. Chiara 50
I - 35122 BRESCIA
tel.: +39 030 298 8642
fax: +39 030 240 0925
e mail: carpita@eco.unibs.it

Daniela COCCHI
Dipartimento di Scienze Statistiche Paolo Fortunati
Via delle Belle Arti 41
I - 40126 BOLOGNA
tel.: +39 051 232 153
fax: +39 051 232 153
e mail: cocchi@stat.unibo.it

Javier CORTES DE LAS HERAS
Instituto Valenciano de Evaluación y Calidad Educativa
Av. Campanar 32
E - 46015 VALENCIA
tel.: +34 963 184 602
fax: +34 963 184 624
e mail: cortes_jav@gva.es

Rita DUKYNAITE
Ministry of Education & Science
A. Volano 27
LIT - 01516 VILNIUS
tel.: +370 5 274 3123
fax: +370 5 261 2077
e mail: rita.kukynaite@smm.lt

Vittorio FABRICATORE
Irre Lombardia
Via Rossini 16
I - 21100 VARESE
tel.: 
fax:
e mail: fabricatore@irre.lombardia.it

Mirco FRANCESCHI
Irre Lombardia
Via Leone XIII, 10
I - 20145 MILANO
tel.: +39 02 408 00 221
fax: 
e mail: franceschi@irre.lombardia.it

Ulf FREDRIKSSON
European Commission, DG JRC
Via Enrico Fermi, 1
I - 21020 ISPRA (VA)
tel.: +39 0332 78 6574
fax: +39 0332 78 5733
e mail: ulf.fredriksson@jrc.it

Debora GATELLI
European Commission, DG JRC
Via Enrico Fermi 1
I - 21020 ISPRA (VA)
tel.: +39 0332 78 6573
fax: +39 0332 78 5733
e mail: debora.gatelli@jrc.it

Guido GAY
Istituto Regionale di Ricerca della Lombardia
Via Copernico 38
I - 20124 MILANO
tel.: +39 02 673 830 229
fax: +39 02 673 830 299
e mail: gay@irer.it

Simone GERZELI
Università di Pavia
Corso Strada Nuova 65
I - 27100 PAVIA
tel.: +39 0382 984 350
fax: +39 0382 984 674
e mail: simone.gerzeli@unipv.it

Ornella GIAMBALVO
Università degli Studi di Palermo
Viale delle Scienze 13
I - 90128 PALERMO
tel.: +39 091 662 6342
fax: +39 091 426 452
e mail: giambi@unipa.it

Michela GNALDI
Comitato Nazionale di Valutazione del Sistema Universitario
Piazza Italia 25
I - ROME
tel.: +39 05 584 96498
fax: 
e mail: michela.gnaldi@miur.it
Lene MEJER  
European Commission, Eurostat  
Bech Building  
L - 2920 LUXEMBOURG  
tel. :+352 4301 35423  
fax: +352 4301 36049  
e mail: lene.mejer@cec.eu.int

Andrea MERCATANTI  
Università di Pisa  
Via Milano 60/6  
tel. :+39 06 479 253 79  
fax: +39 06 479 250 05  
e mail: mercatan@libero.it

Athena MICHAELIDOU  
Cyprus Ministry of Education and Culture  
P.O.Box 12720  
CY - 2252 NICOSIA  
tel. :+357 22 402 317  
fax: +357 22 480 505  
e mail: athens@cyearn.pi.ac.cy

Alexandros MICHALOS  
University of Northern British Columbia  
3333 University Way  
CA - V2N 4Z9 PRINCE GEORGE  
tel. : +1 250 960 6697  
fax: +1 250 960 5744  
e mail: michalos@unbc.ca

Stefania MIGNANI  
Università di Bologna  
Via Belle Arti 41  
I - BOLOGNA  
tel. :+39 051 209 8255  
fax: +39 051 232 153  
e mail: stefania.mignani@unibo.it

Magdalena MOK  
The Hong Kong Institute of Education  
10 Lo Ping Rd.  
HK - Tai Po HONG KONG  
tel. :+852 2948 7704  
fax: +852 2948 8744  
e mail: mmcmok@ied.edu.hk

Sarah MOORE  
European Commission, DG JRC  
Via Enrico Fermi 1  
I - 21020 ISPRA (VA)  
tel. :+39 0332 78 6572  
fax: +39 0332 78 5733  
e mail: sarah.moore@jrc.it

Anna MULAS  
Università di Siena  
P.zza S. Francesco 8  
I - 53000 SIENA  
tel. :+39 0577 235 083  
fax: +39 0577 232 626  
e mail: mulas@unisi.it

Cesare MURGIA  
Invalsi  
Via Borromini 5  
I - 00044 FRASCATI  
tel. :+39 06 941 85313  
fax: +39 06 941 85202  
e mail: acaputo@invalsi.it

Cristina MUSCHITIELLO  
Università degli Studi di Bari  
Via Camillo Rosalba 53  
I - 70124 BARI  
tel. :+39 080 504 9251  
fax: +39 080 504 9147  
e mail:

Michela NARDO  
European Commission, DG JRC  
Via Enrico Fermi, 1  
I - 21020 ISPRA (VA)  
tel. :+39 0332 78 5968  
fax: +39 0332 78 5733  
e mail: MICHEAL.NARDO@JRC.IT

Michela NATILLI  
Scuola Superiore Sant'Anna  
P.zza Martiri della Libertà 33  
I - 56127 CAMPOBASSO  
tel. :+39 050 883 388  
e mail: m.natilli@sssup.it

Erich NEUWIRTH  
University of Vienna  
Universitätsstr. 5-9  
A - 1010 VIENNA  
tel. :+43 1 4277 39464  
fax: +43 1 4277 39459  
e mail: erich.neuwirth@univie.ac.at

Daniela NOTARBARTOLO  
Irre Lombardia  
Via V. Foppa 10  
I - 20144 MILANO  
tel. :+39 02 4989 489  
fax: +39 02 438 00 205  
e mail: notarbartolo@irre.lombardia.it
Jose Maria NYSSSEN GONZALEZ  
ANECA  
c/Orense 11 - 7a planta  
E - 28020 MADRID  
tel.:+34 914 178 230  
fax: +34 915 568 642  
e mail: jmnyssen@aneca.es

Allan OLSON  
Northwest Evaluation Association  
5885 SW Mewadows Road, Suite 200  
USA - 97035 LAKE OSWEGO OR  
tel.: +1 503 624 1951  
fax: +1 503 639 7873  
e mail: allanolson@nwea.org

Omar PACCAGNELLA  
University of Padova  
Vía del Santo 22  
I - 35123 PADOVA  
tel.:+39 049 827 4050  
fax: +39 049 827 4211  
e mail: omar.paccagnella@unipd.it

Laura PAGANI  
Università di Udine  
Via Treppo 18  
I - 33100 UDINE  
tel.:+39 0432 249 576  
fax: +39 0432 249 565  
e mail: pagani@dss.uniud.it

Bo PALASZEWSKI  
Swedish National Agency for Education  
Alstroemergatan 12  
S - 106 20 STOCKHOLM  
tel.:+46 8 527 331 09  
fax: +46 8 24 4420  
e mail: Bo.Palaszewski@skolverket.se

Tiziana PEDRIZZI  
Irre Lombardia  
Via Leone XIII, 10  
I - 20145 MILANO  
tel.:+39 02 438 0235  
fax:  
e mail: pedrizzi@irre.lombardia.it

Fulvia PENNONI  
European Commission, DG JRC  
Via Enrico Fermi 1  
I - 21027 ISPRA (VA)  
tel.:+39 0332 78 6472  
fax: +39 0332 78 5733  
e mail: fulvia.pennon@jrc.it

Giorgio PLAZZI  
Università di Udine  
Via Treppo 8  
I - UDINE  
tel.: +39 0432 24 9583  
fax: +39 0432 24 9595  
e mail: gIORGIO.plazzi5@tin.it

Cherubino PROFETA  
Università L. Bocconi  
Vía Sarfatti. N° 25  
I - 20136 MILANO  
tel.:+39 02 583 623 04  
fax: +39 02 583 623 10  
e mail: cherubino.profeta@unibocconi.it

Carla RAMPICHINI  
University of Firenze, Department of Statistics "Giuseppe Parenti"  
Viale Morgagni 59  
I - 50134 FIRENZE  
tel.:+39 0554 237 246  
fax: +39 0554 223 560  
e mail: carla@ds.unifi.it

Roberto RICCI  
Istituto Regionale Ricerca Educativa Emilia Romagna  
Vía U Bassi 7  
I - BOLOGNA  
tel.:+39 051 227 669  
fax: +39 051 269 221  
e mail: rricci@stat.unibo.it

Andrea SALTELLI  
European Commission, DG JRC  
Via Enrico Fermi, 1  
I - 21020 ISPRA (VA)  
tel.:+39 0332 78 9686  
fax: +39 0332 78 5733  
e mail: andrea.saltelli@jrc.it

Fernando SANCHEZ-PASCUALA NEIRA  
Consejeria de Educacion de la Junta de Castilla y Leon  
Monasterio Nuestra Sra. De Prado  
E - 47071 VALLADOLID  
tel.:+34 983 411 518  
fax: +34 983 412 770  
e mail: sanneife@jcyl.es

Bruno SCARPA  
Università di Pavia  
Strada Nuova 65  
I - 27100 PAVIA  
tel.:+39 0382 984 332  
fax: +39 0382 984 767  
e mail: bruno.scarpa@unipv.it