

Wilfried Bos, Miriam M. Gebauer & Tobias C. Stubbe

Modelling Longitudinal and Trend Data

Symposium in Network 9 (Assessment, Evaluation, Testing and Measurement) at the European Conference on Educational Research (ECER) 2010 in Helsinki, Finland.

Chair (Session I): Tobias C. Stubbe (Institute for School Development Research)

Chair (Session II): Miriam M. Gebauer (Institute for School Development Research)

Discussant: Eugenio J. Gonzalez (IEA-ETS Research Institute)

The majority of current empirical educational research is cross-sectional in nature. Studies of this kind do not allow evidence regarding the development of students' achievement. In order to gain empirical information on educational processes longitudinal studies assessing the same students at different points of time are needed. From the methodological point of view however the implementation and the analysis longitudinal studies implicate numerous challenges.

Modelling repeated measurements with IRT scaling allows avoiding massive problems the classical test theory cannot prevent. Bereitner (1963) and Lord (1963) postulated that (1) the reliability of the difference between the tests show negative relation to the correlation between tests. That means the lower the correlation between the tests – the higher the reliability of difference. But high correlation between tests of different time points is desirable. (2) Different qualifications of the persons tested cannot be taken into account, since person and items difficulty are measured on one scale. (3) The shared error variance of the scores of the first measure point and the differences cause negative correlation between first measure point scores and differences, even when no change appears.

Especially for modelling test scores and data of competence testing IRT offers a sum of elegant methods, that eliminate or at least reduce the above mentioned problems. Currently multiplicities of models exist to measure change, which are realized in different applications:

(1) The Linear Logistic Test Model (LLTM or linear logistic RM) introduced by Fischer (1983, 1995a, b) and Spada & McGaw (1985) has the basic assumptions that the person parameter is considered constant over time and change in the overall item difficulty is

equivalent to global change in latent person ability. Therefore change is homogeneous across persons which lead to the condition of one-dimensionality. The position of person v on the one latent trait underlies his or her response to all items i at all occasions to another. This model can be realized with the applications which are called virtual items or virtual persons.

(2) The Linear Logistic Testmodel with Relaxed Assumptions (LLRA) by Fischer (1983, 1995c) and Andersen (1985) follows the assumptions that the change is person-specific and the item difficulty remains constant over time. Item form a one-dimensional scale at each time point which leads to a multidimensional latent space for each latent-trait continuum. Also item parameters are constant over time points. Virtual items or one test-treaty are applications to implement this model.

(3) The Loglinear conditional Rasch Model by Cressie & Holland (1983) and Kelderman (1983) models homogeneous change, with no person change considered. It is expressed without latent-person parameter, because the total score represents the person parameter.

The expected probabilities of the response vectors are reparameterized as linear combinations of item parameters.

(4) Multidimensional Rasch Model for Learning and Change (MRMLC) by Embretson (1991) models change as a latent dimension takes person qualification into account.

This symposium gathers an international panel of speakers. The altogether six papers – organized in two sessions – address specific methodological challenges in modelling longitudinal data in educational research.

Session I – Paper 1

Wilfried Bos (Institute for School Development Research)

Miriam M. Gebauer (Institute for School Development Research)

Tobias C. Stubbe (Institute for School Development Research)

Modelling Longitudinal and Trend Data – An Introduction

Numerous desiderata for future research in educational science require data from longitudinal studies. These data allow for the modelling of change in students characteristics (e.g. academic achievement). As the analysis of longitudinal data is much more sophisticated than the analysis of cross-sectional data scientists working with this kind of data have to be careful

when choosing the appropriate model(s). With this paper we present an overview on this field of research and thereby give an introduction to the topic of this symposium “Modelling Longitudinal Data”. We will outline the central problems arising when modelling longitudinal data and name the most popular state-of-the-art methods for the analysis of change in academic achievement of students. Furthermore we will introduce the topics of the other five symposium papers.

Session I – Paper 2

Gabriel Nagy (Max Planck Institute for Human Development, Berlin, Germany)

Jürgen Baumert (Max Planck Institute for Human Development, Berlin, Germany)

Rainer Lehmann (Humboldt University, Berlin, Germany)

Modelling Social Disparities in Student Achievement Growth: Multilevel Applications of Growth Curve and Simplex Models

Social inequality in student achievement is a core topic in the educational sciences. Various developmental mechanisms have been proposed to underlie its emergence. First, the Matthew effect assumes a path-dependent process of cumulative advantage, in which achievement gains are a function of prior achievement. This mechanism predicts increasing inequality as initial differences amplify over time. Second, mechanisms of cumulative advantage due to exposure effects have been proposed, in which the additive effects of social class lead to higher levels of inequality. Third, it has been proposed that schooling has inequality-reducing effects that compensate for differences in developmental relevant home resources. This paper discusses the possibilities of modelling these mechanisms when only a limited number of repeated assessments are available. It illustrates the use of multilevel formulations of growth curve and simplex models, drawing on annual assessments of reading and mathematics achievement (grades 4 to 6; N=3169). The results indicated decreasing individual differences, thus refuting a key aspect of the Matthew hypothesis. Results from growth curve models were in line with the compensation hypothesis as long as initial achievement was not controlled for. Simplex models reflecting path-dependent developmental processes suggested that social background is generally unrelated to achievement growth.

Session I – Paper 3

Sarah Frahm (University of Hamburg)

Stephan Jarsinski (Institute for School Development Research)

Andreas Voss (Hamburg University of Applied Sciences)

Modelling Longitudinal Orthography Data with IRT – Results of an Intervention Control Study in Germany (HeLp 2007/08)

In the German HeLp-study (2007/08), a longitudinal intervention and control study, orthography data of more than 1000 fifth graders has been collected by using a systematic orthography test based on the graphematic linguistic theory and on a differentiated competence model. Within the data, an intervention group and a control group have been differentiated. The data has been assessed on three different time points in grade five which is for the intervention group before, during and after the intervention. It will be shown how the complex data can be analyzed and compared within the two groups and the three time points with the help of the Item Response Theory (IRT) in order to gain a maximum insight into the students' development of differentiated orthography skills. The IRT analysis was done on a whole-word level and then differentiated into four subskills. Thereby, the distinct development of subskills can be further analyzed. On top of that, anchor items were compared. We want to present the methodology and also demonstrate, interpret and discuss selected results of the study. It will be shown in how far the differentiation of subskills is important as well as how the two groups differ in their development.

Session I – Paper 4

Anabela Serrão (GAVE – Gabinete de Avaliação Educacional)

Olívia Sousa (GAVE – Gabinete de Avaliação Educacional)

Carlos Pinto Ferreira (GAVE – Gabinete de Avaliação Educacional)

How do Portuguese students perform on Mathematics since grade 4 to grade 6?

Primary education students are exposed to universal assessment tests in Mathematics. The outcomes of these tests provide information which can be used to improve the efficacy of the educational system. The differentiated processed information is sent as a feedback to policy makers, school principals, teachers, students and parents. The present study aims at analyzing the performance evolution of about 100,000 Portuguese students from grade 4 to 6 in what respects competences (concepts and procedures, problem solving, reasoning and

communication) and contents (numbers and operations, geometry, statistics and algebra), tested by those assessment instruments. The identification of possible problems relating either to specific competences or to particular thematic domains can help to improve successful classroom strategies. A comparative analysis of the students' performances in 2007 (Grade 4) and 2009 (Grade 6) tests will be performed, linking these data sets with IRT models. The intent of this paper is to explain the evolution of elementary school students' performances in Mathematics aiming at the identification of possible problems and the proposal of adequate strategies.

Session II – Paper 1

Monica Rosén (University of Gothenburg)

Rolf Strietholt (Institute for School Development Research)

Linking Reading Literacy Tests for a 35 Year Trend Study. Analyses of the Bridge Items

IEA-studies on reading literacy of 9-10-year-old students provide an extensive source for trend analyses over a period of 35 years. A precondition for such analyses is to measure the students' abilities on the same metric. Since the tests in the studies from 1970, 1991, 2001 and 2006 are not identical, this condition is not fulfilled from the outset. However, the IRT-technique provides an approach that allows to link tests and to establish a comparable measure on the same scale and over time. Such an analysis is based on the prerequisite that at least some of the items are equal in the different test forms and that they have similar psychometric properties over time. Bridges of items are available between all these studies by the links in the international studies, RL1991 and RL2001 (Martin, Mullis, Gonzalez, & Kennedy, 2003), PIRLS2001 and PIRLS2006, (Mullis, Martin, Kennedy, & Foy, 2007), and by Swedish extensions of the international designs which provide further bridges between the tests from 1970s reading comprehension and subsequent studies (Rosén, 2006; Taube, 1993). The research question addressed in this study concerns the reliability of these bridge items, i.e. to what degree they are sufficient to form a common IRT-scale.

Session II – Paper 2

Pierre Foy (TIMSS & PIRLS International Study Center)

Michael O. Martin (TIMSS & PIRLS International Study Center)

Ina V. S. Mullis (TIMSS & PIRLS International Study Center)

Measuring Trends in Mathematics and Science Achievement in an Evolving World

The Trends in International Mathematics and Science Study (TIMSS) measures trends in mathematics and science achievement of fourth- and eighth-grade students. TIMSS was first administered in 1995 and has been repeated every four years since then. Most recently, TIMSS 2007 was conducted in 60 countries, including 22 European countries. Measuring trends over time in student achievement while keeping the assessment relevant in a continuously evolving educational environment is a challenging endeavour requiring sophisticated statistical methods. To report mathematics and science achievement for student populations, TIMSS uses item response theory scaling together with conditioning and multiple imputation methodology. This paper describes the concurrent calibration scaling methodology (Foy, Galia, & Li, 2008) used by TIMSS to link successive assessments to a common scale, enabling the measurement of growth or decline in student achievement over time. An important feature of the TIMSS 2007 assessment was a change in the data collection design (Mullis, et al., 2005) necessitating the implementation of an additional “bridging component” to link the 2003 and 2007 assessments. This paper describes the design and implementation of the bridging study, and the steps taken to use the bridging data to establish the link between the 2003 and 2007 assessments.

Session II – Paper 3

Monica Rosén (University of Gothenburg)

Rolf Strietholt (Institute for School Development Research)

Choosing between the 1-, 2- and 3-PL Model in a Trend Study

In our study we apply different IRT-models on trend data from 1970 to 2006 in order to analyze the impact of the different models on the results. We use data from the four countries that took part in all five IEA-studies on reading literacy of 9-10 year-old students in the mentioned period. Basically, IRT-models vary in the number of parameters that describe the relation between the latent trait and the observed response behavior on an item. There is an ongoing discussion which model is most appropriate. In recent large-scale assessment studies,

for instance, IEA uses the 3-PL model in studies like PIRLS and TIMSS whereas the OECD employs the 1-PL model when analyzing PISA achievement tests. Defenders of the 1-PL model highlight its straightforwardness and simplicity. Those who prefer more complex models argue that the assumptions of the 1-PL model are overly restrictive and do not meet actual data. In our view, in social science models hardly ever fully capture reality (Forster, 2004) and it is hard to weigh the arguments against each other. For pragmatic reasons, we will therefore focus on the question if the chosen IRT-model affects the country level achievement estimates with practical relevance.