Journal of Educational and Behavioral Statistics
September 2007, Vol. 32, No. 3, pp. 252–286
DOI: 10.3102/1076998606298042

Multilevel Structural Equation Models for the Analysis of Comparative Data on Educational Performance

Harvey Goldstein
University of Bristol

Gérard Bonnet
Thierry Rocher
Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche, Direction de l'Évaluation et de la Prospective, Paris

The Programme for International Student Assessment comparative study of reading performance among 15-year-olds is reanalyzed using statistical procedures that allow the full complexity of the data structures to be explored. The article extends existing multilevel factor analysis and structural equation models and shows how this can extract richer information from the data and provide better fits to the data. It shows how these models can be used fully to explore the dimensionality of the data and to provide efficient, single-stage models that avoid the need for multiple imputation procedures. Markov Chain Monte Carlo methodology for parameter estimation is described.

Keywords: *international comparisons; factor analysis model; educational assessment; structural equation model; item response model; multilevel model; Markov Chain Monte Carlo*

The principal purpose of this article is to explore methodological issues in the analysis of large-scale data of comparative educational performance using multilevel structural equation models. To illustrate our approach, we analyze data from the Programme for International Student Assessment (PISA) survey of reading performance, which represents a very ambitious and wide-ranging attempt to measure and compare 15-year-olds in 32 countries and which employs procedures and models used widely in analyzing educational performance. We begin by describing the data and raising some preliminary methodological issues.

This research was partly supported by the Ministère de l'Education Nationale, de l'Enseignement supérieur et de la Recherche, direction de l'évaluation et de la prospective, Paris, and by a research grant from the Economic and Social Research Council (RES-000-23-0140). We are very grateful to Fiona Steele, William Browne, and David Thissen for helpful comments and to anonymous referees.

Under the auspices of the Organization for Economic Cooperation and Development (OECD), the testing for PISA was conducted in the first half of 2000, and the study was intended to be the first of a series. Although PISA concentrates on reading, it also has mathematics and science components. The second study, conducted in 2003, concentrated on mathematics, and the third, conducted in 2006, concentrated on science. The sampling design selected schools as first-stage units and sampled 15-year-old pupils within schools with a maximum of 35 students in each school. Extensive piloting of test items and general procedures, including translations, was carried out. The first comprehensive report (OECD, 2001) appeared in 2001, and an extensive (300-page) technical report (Adams & Wu, 2002) provides detail about the procedures used. In addition, data are available for secondary analysis from the OECD Web site (www.pisa.oecd.org/).

The PISA 2000 (OECD, 2001) analyses have concentrated on computing student proficiencies and country means for the three reading proficiency subscales, Retrieving Information, Interpreting Texts, and Reflection and Evaluation, as well as a combined scale. Each subscale is defined by a different set of items. In this article, we analyze data from the first subscale, Retrieving Information, containing 35 items. Details of this subscale can be found in the PISA 2000 technical report (Adams & Wu, 2002). The full scale contains 36 items, but one of these (R076Q03) was eliminated from the England file as "dodgy" because it did not fit well using the one-dimensional scaling procedure applied in the study.

Two countries, France and England, were chosen for this purpose. In PISA itself, Wales did not participate, and according to the technical report (Adams & Wu, 2002, p. 191), Scotland did not properly follow the sampling procedures. Unfortunately, the main OECD reporting only refers to the United Kingdom, that is, the average over England, Scotland, and Northern Ireland; because these have very different educational systems, interpretation is complicated. There is a separate country report (Office for National Statistics [ONS], 2002), however, that does allow direct comparisons with our analysis. The data used in the present article consist of 326 schools (141 in England and 185 in France) and 8,299 students (4,070 in England and 4,229 in France); further details can be found in Adams and Wu (2002).

One problem that arises in comparing France and England (as well as in other country comparisons) is that students move through the systems in different ways. PISA samples by age, namely, all children born in 1984. In England, most children start school in September of the school year in which they reach the age of 5. There is almost no repeating of years, so that a 15-year-old at the time of the PISA survey in April/May 2000, born in August 1984, will start school in September 1988 and be in Grade 11 at the time of the PISA survey and in a class where there are a number of older children (not in PISA) born in September 1983 to December 1983. However, the first year of schooling is designated as reception, so that, in fact, that child will have been in formal schooling for 12 years. A child born in September 1984 will start school 1 year later and be in Grade 10,

and this latter child is about the same age as the former but has had 1 year less schooling.

In France, on the other hand, children start school in September of the calendar year in which they reach 6 years. Thus, a child born in August 1984 who does not repeat a year will be in Grade 10, as will be a child born in September, and both will have received the same amount of schooling. In France, the 1st year in school is counted as Grade 1. Any child who repeats a year (approximately one third do so) will be in Grade 9. Because the normal transition from *collège* to *lycée* occurs after Grade 9, these children will be in collège along with children who have not repeated, that is, those born in 1985. Thus, for the children born between September 1984 and December 1984, the French and English students will have been in formal schooling for the same length of time in terms of grades, although if reception is counted, the English will have been in school 1 year longer. For those born between January 1984 and August 1984, the French students will have been in schooling 1 further year less than the English, whether or not they repeat. However, 100% of French children are in preschool provision (*école maternelle*) for 2 years prior to formal schooling and 94% of French children 3 years prior. In England, about 80% of 3-year-olds are in part-time nursery education. This makes comparisons very difficult, and we discuss later how we attempt, at least partially, to take account of this.

The first section of this article performs some simple analyses, effectively replicating those of the OECD, and goes on to perform some relational analyses. The second section introduces the multilevel (school and student) structure of the data and shows how a valid analysis can be performed. The third section explores the dimensionality of the data at both the student level and the school level. The fourth section shows how a constrained multilevel model can be fitted to make comparisons that have a consistent interpretation. The final section discusses some implications of the findings for international comparisons.

One-Dimensional Latent Variable Models for Student Performance

The standard psychometric procedure for the modeling of test item responses is commonly known as item response theory (Lord, 1980). A simple, basic, latent-trait model of this type relates the responses on a set of test items to one or more underlying latent abilities. A basic version can be expressed as follows.

For a student (i) who responds to item (r), the probability (π_{ri}) of a correct response is given by

$$\begin{aligned} g(\pi_{ri}) &= \beta_{0r} + \lambda_r \theta_i \\ \theta_i &\sim N(0, \sigma_\theta^2) \\ y_{ri} &\sim \textit{binomial}(1, \pi_{ri}), \end{aligned} \tag{1}$$

where g is a link function, typically the logit, and the response, y_{ri}, is 1 if the item is correctly answered and 0 if not and the y_{ri} are mutually independent.

This is just a binary factor model with a single factor (θ) and a set of loadings (λ_r). We refer to the term β_{0r}, often referred to as the "facility" for Item r, as belonging to the fixed part of the model, which will later be augmented with further predictors. In the PISA data, we have some graded or partial-credit items where the correct responses are either partially correct (coded 1) or fully correct (coded 2). In this case, for such an item, the first line of the model can be written, for a response coded s ($s = 0, 1$), as

$$\begin{aligned} g(\gamma_{ri}^{s}) &= \beta_{0r} + \alpha_s + \lambda_r \theta_i \\ \gamma_{ri}^{s} &= \sum_{f=0}^{s} \pi_{ri}^{f},\ \alpha_0 = 0, \end{aligned} \tag{2}$$

where π_{ri}^{f} is the probability of a response in Category f. We here model the cumulative probabilities where cumulation is from Category 0 (incorrect) rather than in the binary case where, by convention, the first category is the correct response. Thus, in the fixed part of the model, the probability of an incorrect response is simply $g^{-1}(\beta_{0r})$, and the probability of an incorrect or partially correct response is $g^{-1}(\beta_{0r} + \alpha_1)$.

An important assumption in Models 1 and 2 is that the responses y_{ri} are conditionally independent. Because some of the items involve responses to the same text or figure, it is possible that this assumption will be violated, as the (conditional) probability of a correct response to one item may depend on the outcome with respect to an earlier item. Thus, for example, Items R104Q01, R104Q02, R104Q05, and R104Q06 all refer to a passage about telephone use and feature the same person (Pedro) in each question.[1] To avoid this problem, an alternative is to consider the complete set of item responses, treating the number of correct responses to this set of four binary items as a single-item-graded response. Thissen, Steinberg, and Mooney (1989) discussed this, and Steinberg and Thissen's (1996) study patterned item combinations as elementary response units (testlets). In the above example, we could, for instance, compute a total score ranging from 0 to 4. Scott and Ip (2002) considered an alternative approach to what they term the "item clustering" effect. For each specified item cluster, they added an individual-level random effect designed to identify an individual's additional response to each item as a member of that cluster (see below). We have not pursued these possibilities here, but they are an interesting area for further work.

A special case of Model 1 is the so-called Rasch model, where the loadings λ_r are constrained to be equal. This is the model, with a logit link, used in the PISA analyses.

In this article, we use a probit link rather than a logit (see, e.g., Lord & Novick, 1968, chap. 16). The two link functions are, in fact, very similar so that we can expect resulting estimated probabilities to be very close. The probit, however, has certain advantages computationally and also has a useful

interpretation in terms of an underlying normal "propensity" distribution for the responses. Thus, for a binary response, we can suppose that there is an underlying continuous response for an item with a threshold value (X) such that responses above that value are correct and those below are incorrect. Formally, we write the probability of a correct response as

$$\int_{X}^{\infty} \phi(x)dx,$$

where ϕ is the standard normal density. This model is discussed, among others, by Fox and Glas (2001) and by Goldstein and Browne (2004) and is here extended to the ordered category case given by Model 2 (see the appendix for a complete specification).

The above model can be extended in several ways. First, we can add further fixed-part explanatory variables such as gender, country, age, and so on. Second, we can make the model multilevel by recognizing between-school variation and explicitly incorporating school-level latent variables or factors. Third, we can add further factor dimensions along which student responses can vary. Fourth, we can allow the loadings to be functions of explanatory variables. Finally, we can allow the factor values or scores also to be functions of explanatory variables. It is possible to extend the model to consider more general structural equation models, but we shall not pursue this here. Steele and Goldstein (2006) gave a further discussion and an application to women's status data.

Generalizing the notation of Goldstein and McDonald (1988) and McDonald and Goldstein (1989), a basic multidimensional two-level model for a continuous normal response is given by Model 3. In the case of binary or ordered responses, this models the underlying propensity as defined above, and our exposition thus will be expressed in terms of the following basic model.

$$\begin{aligned} &y_{rij} = \sum_{h} \beta_{hr} x_{hrij} + \sum_{f=1}^{F} \lambda_{fr}^{(2)} v_{fj}^{(2)} + \sum_{g=1}^{G} \lambda_{gr}^{(1)} v_{gij}^{(1)} + u_{rj} + e_{rij} \\ &u_{rj} \sim N(0, \sigma_{ur}^2), e_{rij} \sim N(0, \sigma_{er}^2), v_{j}^{(2)} \sim MVN_F(0, \Omega_2), v_{ij}^{(1)} \sim MVN_G(0, \Omega_1) \qquad (3) \\ &r = 1, \ldots, R, i = 1, \ldots, n_j, j = 1, \ldots, J, \sum_{j=1}^{J} n_j = N, \end{aligned}$$

where h indexes the fixed-part explanatory variables, R is the number of responses, F the number of Level 2 factors, and G the number of Level 1 factors. The $v_{fj}^{(2)}$, $v_{gij}^{(1)}$ are respectively sets of common factors at Levels 2 and 1 with corresponding uniquenesses u_{rj}, e_{rij}. Note that correlations between the factors are allowed, although we do not fit correlated factors in the present article. Where the response is binary or ordered, then, for the underlying propensity distribution, we have $e_{rij} \sim N(0, 1)$. The subscript r refers to the item, i to the student, and j to

the school. In this article, we shall assume that the diagonal terms of Ω_1, Ω_2 are 1. The alternative is to fix one or more loadings and allow Ω_1, Ω_2 to be general covariance matrices to be estimated. Goldstein and Browne (2004) gave a further discussion, and the steps needed for both formulations are described in the appendix.

Several other authors have studied Model 3 and extensions to it. Zhu and Lee (1999) and Fox and Glas (2001) used Markov Chain Monte Carlo (MCMC) estimation, the former based on Gibbs sampling for a single-level factor model, whereas the latter authors consider the binary response two-level model with a single factor at Level 1 and use Gibbs sampling with a probit link function.

A number of authors have extended Model 3, using maximum likelihood procedures, to include categorical responses and more general structural equation formulations. Thus, Muthen (1997) considered applications to latent growth curve modeling, and a more general discussion of these models is also given by Muthen (1989, 2002). More recently, a very general framework for multilevel structural equation modeling is provided by Rabe-Hesketh, Skrondal, and Pickles (2004) that includes most of the models to be discussed below. These authors obtain maximum likelihood estimates, typically based on quadrature, although other authors (e.g., Raudenbush, 1995) use an expectation maximization (EM) algorithm. Song and Lee (2004) fit this model for mixtures of normal, binary, and ordered responses using a mixture of Gibbs and Metropolis-Hastings sampling.

An advantage of carrying out the estimation using MCMC methods, as in the present article, is the ability to incorporate prior information in a Bayesian sense and to provide exact interval estimates for parameters or functions of parameters. Also, because of the modularity of the algorithm, it is possible to add additional complexity relatively straightforwardly, including the possibility of incorporating distributional assumptions other than normality. The algorithm described in the appendix assumes diffuse priors but is readily extended to incorporate informative prior distributions. It extends previous work in particular by allowing for missing data and parameter constraints among fixed coefficients. It also proposes the use of the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & Van der Linde, 2002) that provides a measure of model complexity and can be used for comparing nonnested models. In the single-level case, the DIC is analogous to the Akaike Information Criterion (AIC) and can be considered a generalization of this. Unlike other model fit procedures such as Bayes factors, it does not require improper (diffuse) priors such as are used for some of the parameters.

The procedures have been implemented using MATLAB (Mathworks, 2004) and MLwiN software (Rasbash, Browne, & Steele, 2004). Although MLwiN has some basic facilities for multilevel factor modeling, the algorithm as written in MATLAB is more flexible, although computationally slow.

In the next section, we describe the analysis models used in PISA and conduct some simple comparisons using the above formulation, before describing our more detailed analyses.

The OECD PISA Models

The unidimensionality assumption was used by PISA to determine the structure of the tests as well as the subsequent modeling. The analysis essentially involves two stages (Adams & Wu, 2002). The first stage consists of fitting Models 1 and 2 to the complete data set using equal loadings. This provides estimates of the intercept parameters β_{0r}. Treating these estimates as known parameter values, a further unidimensional Rasch model is fitted to the responses, but this time including as fixed explanatory variables (conditioning variables) a set of scales formed from a principal components analysis of the data in the student questionnaire. These data include items related to both the school and social background of the students. This is done separately for each country. To take account of the uncertainty in the factor scores, an estimate of the posterior distribution of the factor score for each student is obtained, and five values are sampled randomly from this distribution for each student. For example, if MCMC were used for the analysis, these could be five (approximately) independent values from the chain of factor scores for each student, obtained by selecting values suitably far apart in the chain. Alternatively, values can be selected from parallel chains. These plausible values are then used in subsequent analyses to compare countries and so on. Essentially, five analyses are performed, each one using just one plausible value for each student, and the analysis estimates are then combined to obtain inferences. Mislevy (1991) described the procedure.

Certain problems arise with this approach. The first is that although the plausible values may be expected to perform reasonably well for models that use a subset of the conditioning data, this will not generally be true for variables not in the conditioning data set. This includes school-level variables and in particular applies to multilevel analyses where school is the higher level unit (see Mislevy, 1991). The second problem is that the uncertainty attached to the intercept parameter estimates is not taken into account, although it is not clear whether this is a serious problem. A related issue is that the intercept parameter estimates (item difficulties) are based on a model that does not include the conditioning variables. The PISA analyses use sampling weights that reflect the achieved sample characteristics. In the present article, we shall ignore these because they appear to make only small differences.

In subsequent sections, we show how a fully efficient analysis, avoiding the use of plausible values, can be performed. Our first simple analysis, however, looks at the assumption of equal loadings. We fitted Model 1 across the whole data set with and without the equal loadings constraint, and the results are given in Tables 1 and 2, with the loading and intercept estimates together with their standard errors. For these and subsequent analyses, we used a "burn in" of 1,000 and a subsequent 5,000 iterations for the chain. For the main parameters of interest, the fixed-part coefficients and the loadings, the mixing of the chains is reasonable, as it is for the residual variance estimates. The mixing for the threshold

TABLE 1
Comparisons of Intercept Parameter Estimates

	Equal Loadings (Rasch Model)		Unequal Loadings	
Question	Estimate	*SE*	Estimate	*SE*
R040Q02	0.66	0.04	0.64	0.04
R040Q03A	0.34	0.03	0.36	0.04
R070Q02	0.33	0.03	0.34	0.03
R070Q03	0.93	0.03	0.90	0.03
R076Q05	0.05	0.03	0.07	0.03
R077Q02	0.83	0.04	0.84	0.04
R083Q02	1.33	0.03	1.24	0.04
R083Q03	1.13	0.03	1.15	0.04
R088Q03	1.00 (1.39)	0.066 (0.10)	1.00 (1.35)	0.08 (0.13)
R091Q05	2.18	0.06	2.00	0.07
R100Q04	0.31	0.03	0.31	0.03
R104Q01	1.31	0.03	1.56	0.07
R104Q02	−0.13	0.03	−0.10	0.03
R104Q05	−0.07 (1.73)	0.04 (0.15)	−0.11 (1.91)	0.03 (0.17)
R104Q06	0.93	0.03	0.89	0.03
R110Q04	1.30	0.04	1.45	0.06
R110Q05	1.35	0.04	1.44	0.05
R111Q04	1.01	0.03	0.98	0.04
R119Q06	1.36	0.04	1.31	0.04
R122Q03T	0.56 (0.84)	0.05 (0.09)	0.60 (0.89)	0.05 (0.08)
R216Q04	−0.05	0.04	−0.04	0.04
R219Q01E	1.38	0.03	1.29	0.04
R220Q01	0.34	0.03	0.33	0.03
R225Q03	1.78	0.05	1.82	0.06
R225Q04	1.12	0.03	1.16	0.04
R227Q02T	0.48 (0.91)	0.04 (0.07)	0.46 (0.87)	0.04 (0.07)
R227Q06	1.20	0.03	1.34	0.06
R234Q01	1.37	0.04	1.42	0.05
R234Q02	−0.76	0.03	−0.72	0.03
R237Q01	0.75	0.03	0.83	0.04
R238Q01	0.40	0.03	0.40	0.03
R239Q02	0.35	0.03	0.34	0.03
R245Q01	0.84	0.03	0.79	0.03
R246Q01	0.92	0.03	1.05	0.05
R246Q02	−0.21	0.03	−0.20	0.03

Note: Threshold parameter estimates for four ordered category items are in parentheses. Burn in = 1,000; sample = 5,000. Single-level factor model with 8,299 pupils.

TABLE 2
Comparisons of Parameter Estimates—Loadings

Question	Equal Loadings (Rasch Model = 1)	Unequal Loadings	
		Estimate	*SE*
R040Q02	1	0.55	0.05
R040Q03A	1	0.69	0.05
R070Q02	1	0.73	0.05
R070Q03	1	0.60	0.05
R076Q05	1	0.82	0.06
R077Q02	1	0.62	0.05
R083Q02	1	0.43	0.04
R083Q03	1	0.63	0.05
R088Q03	1	0.63	0.06
R091Q05	1	0.32	0.08
R100Q04	1	0.77	0.05
R104Q01	1	1.10	0.08
R104Q02	1	0.35	0.03
R104Q05	1	0.92	0.09
R104Q06	1	0.54	0.04
R110Q04	1	0.90	0.07
R110Q05	1	0.82	0.06
R111Q04	1	0.54	0.05
R119Q06	1	0.56	0.05
R122Q03T	1	0.77	0.06
R216Q04	1	0.70	0.06
R219Q01E	1	0.47	0.05
R220Q01	1	0.53	0.05
R225Q03	1	0.70	0.06
R225Q04	1	0.76	0.05
R227Q02T	1	0.50	0.04
R227Q06	1	0.88	0.08
R234Q01	1	0.77	0.06
R234Q02	1	0.52	0.05
R237Q01	1	0.85	0.05
R238Q01	1	0.68	0.05
R239Q02	1	0.59	0.04
R245Q01	1	0.44	0.04
R246Q01	1	0.95	0.07
R246Q02	1	0.61	0.04
Factor variance (*SE*)	0.423 (0.012)	1.0	
DIC (PD)	89,793.4 (5,478)	89,129.7 (5,580)	

Note: DIC = deviance information criterion; PD = effective number of parameters.

parameters is poor using the Gibbs sampling approach but is satisfactory when Metropolis-Hastings sampling is used (see appendix, Step 1, Ordered Responses of the algorithm). Factor scores are estimated by computing the mean of the chain values for each student. The variances of these means are also estimated from the chain, and the inverses of these variance estimates are used as weights in Table 3 to provide a valid analysis comparing the French and English mean scores; standard errors of parameters are computed using sandwich estimators. This analysis and that in Table 4 are thus just weighted least squares regression models where we have followed the PISA procedure of computing a factor score for each student and then using these in subsequent analyses.

It is evident from inspecting the estimates and their standard errors that there is strong evidence for differential loadings. This is confirmed by comparing the values of the DIC. In the present case, there is a reduction of 664 for the model with unconstrained loadings. The two assumptions for the loadings also provide somewhat different fixed-part parameter estimates for the difference between France and England, although in both cases the 95% interval includes zero. The point estimates for the equal loadings model have a variance about two thirds that for the unequal loadings model, although the results are similar in terms of effect sizes to those obtained using PISA procedures (ONS, 2002). Table 4 extends the comparison by including a gender effect and an interaction between gender and country. This shows a slightly smaller difference in favor of girls in England than in France (although not reaching the formal 5% significance level) with similar results in terms of effect sizes for both models.

We now compare this procedure with a single-model approach where the explanatory variables are included in the factor model. In this and subsequent analyses, we use the model with unconstrained loadings. Thus, for a model with the single additional explanatory variable country (denoted by the dummy variable x_1), we have the single-level model

$$y_{rij} = \beta_{0r} + \beta_1 x_{1ij} + \lambda_r^{(1)} v_{ij}^{(1)} + e_{rij}. \tag{4}$$

Note that if we were to adopt the item cluster effect model of Scott and Ip (2002), Model 4 would become, for cluster k,

$$\begin{aligned} y_{rij} &= \beta_{0r} + \beta_1 x_{1ij} + \lambda_r^{(1)} (v_{ij}^{(1)} + \gamma_{kij}) + e_{rij} \\ \gamma_{kij} &\sim N(0, \sigma_k^2). \end{aligned}$$

Instead of a constant β_1, we could allow a different country coefficient for each response (β_{1r}), but interest here lies in an overall comparison between countries so that these are constrained to be equal. We discuss how to interpret this parameter below.

TABLE 3
Single-Level Weighted Least Squares Models for Country Comparisons

	Estimate (*SE*)	
Parameter	Equal Loadings (Rasch Model)	Unequal Loadings
Intercept	−0.05 (0.008)	−0.007 (0.012)
Country (England–France)	−0.002 (0.011)	0.021 (0.018)

TABLE 4
Single-Level Weighted Least Squares Models for Country and Gender Comparisons

	Estimate (*SE*)	
Parameter	Equal Loadings (Rasch Model)	Unequal Loadings
Intercept	0.040 (0.011)	−0.041 (0.012)
Country (England–France)	−0.006 (0.016)	−0.018 (0.024)
Gender (male–female)	−0.084 (0.016)	−0.112 (0.023)
Country × Gender	0.034 (0.023)	0.042 (0.034)

The OECD analyses essentially fit the equivalent of the following model:

$$\begin{aligned} &y_{rij} = \beta_{0r} + (X^{(c)}\beta^{(c)})_{ij} + \lambda_r^{(1)} v_{ij}^{(1)} + e_{rij},\ v_{ij}^{(1)} = \alpha_1 x_{1ij} + v_{ij}^{*} \\ &v_{ij}^{(1)} \sim N(0,1),\ v_{ij}^{*} \sim N(0, \sigma_v^2), \end{aligned} \tag{5}$$

albeit with equal loadings, where the term $(X^{(c)}\beta^{(c)})_{ij}$ represents the effect of the conditioning variables. However, the two distributional assumptions in Model 5 will not in general both be satisfied; one case where they are satisfied is where $\{v, x, v^{*}\}$ have a joint multivariate normal distribution. Instead, the following model avoids this problem:

$$\begin{aligned} &y_{rij} = \beta_{0r} + \lambda_r^{(1)} v_{ij}^{(1)} + e_{rij},\ v_{ij}^{(1)} = \alpha_1 x_{1ij} + \delta_{ij} \\ &\delta_{ij} \sim N(0, \sigma_\delta^2), \end{aligned} \tag{6}$$

which on substitution gives

$$y_{rij} = \beta_{0r} + \alpha_1 \lambda_r^{(1)} x_{1ij} + \lambda_r^{(1)} \delta_{ij} + e_{rij}. \tag{7}$$

This is a structural equation model, and Model 6 is not equivalent to Model 4. It is an example of a Multiple Indicator Multiple Indicator Cause (MIMIC) model (Muthen, 1989). More generally, the structural part of the model will include

further variables of interest at individual and school levels. A choice between Models 6 and 4 in any particular case can be determined by the model that has the better fit to the data. We note, however, that the interpretation of Model 4 is not completely straightforward because the responses have different variances; only the Level 1 residual variances are equal. Thus, although the coefficients are equal across responses in Model 4, the response distributions themselves are not standardized. We therefore use models with structural predictors based on Model 6 in the following analyses.

In the present case, fitting Model 4 to the data gives an estimate for β_1 of 0.009 (0.019); with equal loadings, this is −0.002 (0.011). Fitting Model 6 gives an estimate for α_1 of 0.014 (0.026) with a DIC value of 89,145.5, compared with a rather similar value of 89,142.8 for Model 4.

Note that in the (Rasch) case of equal loadings, Models 4 and 6 do become equivalent. We also note that for the case of more than one factor at a level and also for factors at more than one level, we may wish to model each factor score as a function of explanatory variables, in which case Model 4 is inadequate and we must use models that are an extension of Model 6.

We now look at the case of the two-level model, where Model 4 becomes

$$y_{rij} = \beta_{0r} + \beta_1 x_{1ij} + \lambda_r^{(1)} v_{ij}^{(1)} + \lambda_r^{(2)} v_j^{(2)} + u_{rj} + e_{rij}, \tag{8}$$

and the PISA analyses fit essentially the (variance components) model,

$$\begin{aligned} &y_{rij} = \beta_{0r} + \lambda_r^{(1)} v_{ij}^{(1)} + e_{rij},\ v_{ij}^{(1)} = \alpha_1 x_{1ij} + u_j + v_{ij}^{*} \\ &v_{ij} \sim N(0,1),\ v_{ij}^{*} \sim N(0, \sigma_v^2),\ u_j \sim N(0, \sigma_u^2), \end{aligned} \tag{9}$$

where again we have the problem that the distributional assumptions cannot in general simultaneously be satisfied. An alternative two-level model is

$$\begin{aligned} &y_{rij} = \beta_{0r} + \lambda_r^{(1)} v_{ij}^{(1)} + e_{rij},\ v_{ij}^{(1)} = \alpha_1 x_{1ij} + u_j + \delta_{ij} \\ &\delta_{ij} \sim N(0, \sigma_v^2),\ u_j \sim N(0, \sigma_u^2), \end{aligned}$$

which becomes, on substitution,

$$y_{rij} = \beta_{0r} + \alpha_1 \lambda_r^{(1)} x_{1ij} + \lambda_r^{(1)} \delta_{ij} + \lambda_r^{(1)} u_j + e_{rij}. \tag{10}$$

Comparing Model 10 with Model 4, we see that effectively the Level 1 and Level 2 loading vectors are constrained to be equal and there is no additional Level 2 residual term in Model 10. Thus, Model 10 becomes a highly constrained model. We show below, however, that an extension of Model 10 does provide a useful model. Thus, Table 5 shows a simple two-level model fit from which it is clear that the Level 1 and Level 2 loading vectors are very different. Furthermore, even with equal loadings, Models 8 and 10 are not equivalent.

TABLE 5
Two-Level Model With One Factor at Each Level

		Level 1 Loading		Level 2 Loading	
Question	Intercept (Threshold)	Estimate	*SE*	Estimate	*SE*
R040Q02	0.61	0.41	0.04	0.39	0.05
R040Q03A	0.32	0.47	0.05	0.57	0.06
R070Q02	0.31	0.55	0.05	0.49	0.04
R070Q03	0.88	0.45	0.05	0.41	0.04
R076Q05	0.05	0.63	0.06	0.56	0.05
R077Q02	0.82	0.53	0.05	0.35	0.04
R083Q02	1.26	0.37	0.05	0.25	0.05
R083Q03	1.16	0.57	0.06	0.31	0.06
R088Q03	0.92 (1.33)	0.43	0.06	0.48	0.05
R091Q05	2.03	0.33	0.07	0.12	0.07
R100Q04	0.32	0.57	0.05	0.56	0.03
R104Q01	1.62	1.01	0.08	0.63	0.05
R104Q02	−0.11	0.31	0.04	0.18	0.04
R104Q05	−0.11 (1.94)	0.77	0.09	0.54	0.06
R104Q06	0.90	0.47	0.05	0.31	0.06
R110Q04	1.44	0.74	0.07	0.61	0.05
R110Q05	1.47	0.76	0.08	0.43	0.05
R111Q04	0.97	0.36	0.05	0.45	0.05
R119Q06	1.32	0.49	0.05	0.29	0.05
R122Q03T	0.60 (0.92)	0.63	0.09	0.48	0.05
R216Q04	−0.10	0.49	0.05	0.53	0.05
R219Q01E	1.33	0.37	0.05	0.33	0.05
R220Q01	0.32	0.42	0.05	0.38	0.04
R225Q03	1.88	0.64	0.08	0.37	0.06
R225Q04	1.21	0.61	0.06	0.54	0.05
R227Q02T	0.46 (0.89)	0.36	0.05	0.37	0.03
R227Q06	1.34	0.71	0.07	0.60	0.06
R234Q01	1.46	0.63	0.07	0.48	0.06
R234Q02	−0.72	0.42	0.05	0.32	0.04
R237Q01	0.81	0.73	0.06	0.50	0.05
R238Q01	0.41	0.53	0.05	0.45	0.04
R239Q02	0.33	0.53	0.04	0.32	0.04
R245Q01	0.81	0.34	0.05	0.29	0.04
R246Q01	1.04	0.85	0.07	0.57	0.06
R246Q02	−0.23	0.56	0.05	0.32	0.04
DIC (PD)	88,579.5 (6,148)				

Note: Burn in = 1,000; sample = 5,000. DIC = deviance information criterion; PD = effective number of parameters.

In the next section, we develop Model 7, introducing further explanatory variables, and in a later section, we look at the dimensionality structure of the data.

One-Dimensional Models With Several Explanatory Variables

In view of the problems of differential lengths of schooling discussed in the introduction, we fit in our initial models the interactions between age and country. Age is categorized as a dummy variable January to August versus September to December births. The additional use of grade is problematic. For example, if we compare the September to December births in the two countries, all the English are in Grade 10, whereas the repeating French are in Grade 9. If we only compare Grade 10 pupils, then the French will tend to do relatively better because of the strong negative association between performance and repetition. For the children born in January to August, all the English are in Grade 11, whereas the French nonrepeaters are in Grade 10, and the repeaters are in Grade 9. For these reasons, we have not used grade in our comparisons, but a special analysis of the effect of grade in the 2003 PISA survey in France will be reported elsewhere.

Table 6 extends Model 6 by including age-group, gender, and the first-order interactions of age-group, gender, and country; interactions between age trends and country are negligible and not displayed.

We see that there is an advantage to the older pupils and a negative trend with month of birth for the period January to August, with the older children scoring higher and a gender effect in favor of girls. There is little evidence for a trend for the period September to December, or for a country difference, or for any interactions, except possibly for country by gender.

Two-Level Models

We now fit the full two-level factor model with structural predictors and a single factor at each level together with just gender and age terms. No interactions are significant, and fitting a coefficient for country also gave a high standard error for the England–France difference, as well as a rather badly mixing chain with very high serially correlated values. Running the chain for longer provided no evidence that the coefficient was significant. The results are presented in Table 7.

There remains a large gender difference in favor of girls and an advantage to the older pupils. There is a negative trend with month of birth for the period January to August and also evidence for a positive trend for the period September to December. The latter seems difficult to explain.

Exploring Dimensionality

We now explore the dimensionality structure of the data. We have performed a series of analyses at a single level that establishes the existence of at least two dimensions. In the PISA analyses, items were a priori selected for membership of the three separate proficiencies, with each item identified with just one proficiency.

TABLE 6
Single-Level Factor Model for Country and Age-Group Comparisons

Parameter	Estimate (*SE*)
Country (England–France)	−0.036 (0.031)
Gender (male–female)	−0.187 (0.042)
Age group (September–December = 1)	−0.264 (0.054)
Country × Age Group	0.052 (0.047)
Country × Gender	0.081 (0.044)
Age Group × Gender	−0.013 (0.056)
Age coefficient (January–August)	−0.033 (0.007)
Age coefficient (September–December)	0.037 (0.020)
DIC (PD)	89,136.9 (5,570)

Note: Structural model. Intercepts are not shown. Burn in = 1,000; sample = 5,000. Unequal factor loadings. Age (January–August) is defined as born in January (0), February (1), ... August (7), September–December (0). Age (September–December) is defined as born in September (0), October (1), ... December (3), January–August (0). Age is measured in months.

TABLE 7
Two-Level Factor Model for Gender and Age-Group Comparisons

Parameter	Estimate (*SE*)
Gender (male–female)	−0.145 (0.031)
Age group (September–December = 1)	−0.207 (0.055)
Age coefficient (January–August)	−0.023 (0.009)
Age coefficient (September–December)	0.059 (0.020)
DIC (PD)	88,592.2 (6,223)

Note: Structural model. Intercepts not shown. Burn in = 1,000; sample = 5,000. Unequal loadings. DIC = deviance information criterion; PD = effective number of parameters.

The items used in our analyses of the Retrieving Information proficiency subscale are therefore unique to that scale. In the present analyses, we fit orthogonal factors so that we can detect dimensions along which countries may differ meaningfully (see Steele & Goldstein, 2006, for an example with correlated factors).

In common with all factor models, we have choices to make to ensure identifiability. In the models of this article where Ω_1, Ω_2 are identity matrices, a simple procedure at any given level is to set, for the jth factor ($j > 1$), $\lambda_k = 0$, $k = 1, \ldots j - 1$ (Goldstein & Browne, 2004). In Table 8, we do this to fit two Level 1 factors, setting the first item of the second factor to zero. We start with a model including just the intercepts and a fixed effect for country varying across responses. This model (DIC = 87,028.9) provides a better fit and somewhat different loading estimates compared with a model fitting intercepts only (DIC = 87,687.3) and a much better fit than the basic model for one factor (DIC = 88,471.3).

TABLE 8
Single-Level Factor Model Loading Estimates With Two Uncorrelated Factors at Level 1

	Level 1 Factor 1		Level 1 Factor 2	
Question	Estimate	*SE*	Estimate	*SE*
R040Q02	0.65	0.06	0.00	0.00
R040Q03A	0.95	0.10	−0.01	0.09
R070Q02	0.85	0.07	0.17	0.08
R070Q03	0.65	0.06	0.16	0.07
R076Q05	1.09	0.08	0.08	0.10
R077Q02	0.51	0.05	0.34	0.07
R083Q02	0.32	0.06	0.27	0.06
R083Q03	0.43	0.07	0.42	0.06
R088Q03	0.78	0.10	0.05	0.07
R091Q05	0.37	0.10	0.10	0.14
R100Q04	0.50	0.07	0.55	0.06
R104Q01	0.62	0.14	1.09	0.13
R104Q02	−0.07	0.06	0.58	0.05
R104Q05	0.49	0.09	0.78	0.11
R104Q06	−0.09	0.11	1.07	0.15
R110Q04	0.70	0.05	0.56	0.09
R110Q05	0.61	0.06	0.57	0.07
R111Q04	0.46	0.05	0.35	0.06
R119Q06	0.48	0.06	0.29	0.07
R122Q03T	0.93	0.09	0.12	0.09
R216Q04	0.79	0.06	0.14	0.06
R219Q01E	0.66	0.07	0.04	0.06
R220Q01	0.40	0.06	0.43	0.07
R225Q03	0.67	0.08	0.41	0.10
R225Q04	0.73	0.07	0.40	0.10
R227Q02T	0.43	0.05	0.30	0.05
R227Q06	0.65	0.08	0.67	0.07
R234Q01	0.87	0.12	0.32	0.17
R234Q02	0.43	0.06	0.27	0.07
R237Q01	0.67	0.06	0.48	0.07
R238Q01	0.54	0.07	0.49	0.10
R239Q02	0.41	0.05	0.52	0.05
R245Q01	0.35	0.06	0.28	0.07
R246Q01	0.68	0.06	0.99	0.09
R246Q02	0.38	0.05	0.65	0.07
DIC	87,028.9			

Note: Intercept and country fixed-part predictors without equality constraints. DIC = deviance information criterion.

We see clear evidence in Table 8 for at least two factors. If we fix all the loadings below 0.2 to 0 and reestimate, we obtain the results in Table 9, with a somewhat higher value of DIC (87,312.6).

The interpretation of factors estimated in this way is problematic because a different choice of zero loading will, in general, lead to different loading patterns. In fact, using different starting values, we find that the loadings are not stable, moving from one factor to the other. To explore the various possibilities is time-consuming, and we have not done this because our principal aim is to see whether more than one dimension exists; further factors can be fitted in similar ways, however. Another approach would be to fit simple-structure models where each item loads on only one factor at each level, but the factors are allowed to be correlated. This involves choosing appropriate subsets of items, and we have not pursued this. An exploration of the factor space will need to make choices about the loadings to be fixed based on substantive considerations of item formats, positioning, and content. In addition, when carrying out such an exploration, we should fit a two-level model and fit explanatory variables such as age and country and also allow for the possibility that factor structures may vary across countries. It may also be useful to carry out exploratory analyses separately at each level based on a separate modeling of estimated Level 1 and Level 2 residual covariance matrices (see, e.g., Rowe, 2003).

A Constrained Multilevel Structural Model

Rather than fitting the following Model 11 and estimating the loadings for each change in model parameters, we can consider fitting a standardizing model such as Model 6 and subsequently treating the posterior mean estimates of the loadings as fixed in further analyses. The advantage of this approach is that we are dealing with essentially the same factors, as defined by the loadings, in each analysis. Reestimating the loadings for each fitted model will complicate interpretation.

Clearly, various choices for the standardizing model are possible—for example, fitting a 2-level structure with the loadings for each response constrained to be equal across levels. In practical applications, sensitivity analyses can be performed to see whether inferences are strongly affected by different choices.

We present here only the results from fitting a single factor, but the model can be extended by fitting structural parameters in the case of more than one factor.

$$
\begin{aligned}
y_{rij} &= \beta_{0r} + \sum_{h=1}^{q} \beta_h x_{hij} + \lambda_r^{(1)} v_{ij}^{(1)} + e_{rij} \\
v_{ij}^{(1)} &= \sum_{k=1}^{p} \alpha_k z_{kij} + v_{ij}^{*} + u_j^{*},
\end{aligned}
\tag{11}
$$

TABLE 9
Single-Level Factor Model Loading Estimates With Two Uncorrelated Factors at Level 1, Setting Loadings < 0.2 in 7 to Zero

	Level 1 Factor 1	Level 1 Factor 2
Question	Estimate	Estimate
R040Q02	0.64	0
R040Q03A	0.93	0
R070Q02	0.95	0
R070Q03	0.72	0
R076Q05	1.09	0
R077Q02	0.53	0.32
R083Q02	0.34	0.23
R083Q03	0.47	0.36
R088Q03	0.81	0
R091Q05	0.40	0
R100Q04	0.57	0.49
R104Q01	0.76	1.02
R104Q02	0	0.57
R104Q05	0.63	0.79
R104Q06	0	1.15
R110Q04	0.74	0.59
R110Q05	0.64	0.59
R111Q04	0.49	0.29
R119Q06	0.49	0.28
R122Q03T	0.94	0
R216Q04	0.84	0
R219Q01E	0.62	0
R220Q01	0.45	0.39
R225Q03	0.63	0.34
R225Q04	0.72	0.39
R227Q02T	0.45	0.25
R227Q06	0.72	0.53
R234Q01	0.80	0.31
R234Q02	0.48	0.19
R237Q01	0.73	0.45
R238Q01	0.57	0.43
R239Q02	0.45	0.47
R245Q01	0.40	0.22
R246Q01	0.77	0.93
R246Q02	0.44	0.61
DIC	87,312.6	

Note: Intercept and country fixed-part predictors without equality constraints. DIC = deviance information criterion.

where the structural predictors z_k are distinct from the fixed-part predictors x_h and the Level 2 random effect is incorporated into the Level 1 structural model. We may allow different variances for different groups at both Level 1 and Level 2, and in the present case, we fit different country variances at both Level 1 and Level 2. The second line of Model 11 becomes

$$\begin{aligned} v_{ij}^{(1)} &= \sum_{k=1}^{p} \alpha_k z_{kij} + v_{1ij}^{*} z_{1j} + v_{2ij}^{*} z_{2j} + u_{1j}^{*} z_{1j} + u_{2j}^{*} z_{2j} \\ z_{1j} &= 1 \text{ if England, 0 if France, } z_{2j} = 1 - z_{1j} \\ v_1^{*} &\sim N(0, \sigma_{v1}^2),\ v_2^{*} \sim N(0, \sigma_{v2}^2),\ u_1^{*} \sim N(0, \sigma_{u1}^2),\ u_2^{*} \sim N(0, \sigma_{u2}^2). \end{aligned} \tag{12}$$

When Models 11 and 12 are combined, because the Level 1 loadings are assumed known, we have the random coefficient factor model

$$\begin{aligned} y_{rij} = \beta_{0r} + \sum_{h=1}^{q} \beta_h x_{hij} + \sum_{k=1}^{p} \alpha_k z_{kij} \lambda_r^{(1)} \\ + v_{1ij}^{*} \lambda_r^{(1)} z_{1j} + v_{2ij}^{*} \lambda_r^{(1)} z_{2j} + u_{1j}^{*} \lambda_r^{(1)} z_{1j} + u_{2j}^{*} \lambda_r^{(1)} z_{2j} + e_{rij}. \end{aligned} \tag{13}$$

We fit Model 13 with the Level 1 loadings of Table 5 and the intercept and structural predictors of Table 7 without the gender and country interactions (which are not significant at the 5% level), and we obtain the results in Table 10.

We have also fitted Model 13 where the predictors are in the fixed part of the model rather than the structural part (Table 11).

We note that for the structural model, the coefficients tend to be smaller compared to their standard errors than for the fixed-part predictor coefficient model, and the latter is also a better fit with a DIC of 88,559.2, compared with 88,573.1 in the structural model.

We see that the ratio of Level 2 to Level 1 plus Level 2 factor variances, the variance partition coefficient (VPC; Goldstein, Browne, & Rasbash, 2002), is 21% for England and 49% for France. These values are similar to those presented in PISA (Adams & Wu, 2002). Goldstein (2004) suggested that the explanation for the high value for France is that the data contain a mixture of pupils from Grades 9 and 10. As pointed out above, repetition implies a greater variation among schools. We have therefore conducted an analysis for France only using Model 13 and fitting only intercept terms, and we find that for Grade 9 (collège) pupils, the Level 2 variance estimate is 0.19, and for Level 1, the variance estimate is 0.79, giving a VPC of 19%. For lycée pupils, the variances are 0.14 and 0.54 with a VPC of 21%; thus, the VPC estimate for each school type is close to the English estimate. This explanation for the apparently high between-school variation also accounts for results from the Trends in International Mathematics and Science Study (TIMSS; Mullis et al., 2001), which show similar values for the two countries and where the sampling for France was carried out only in collège.

TABLE 10
Two-Level (Structural) Factor Model for Country, Gender, and Age-Group Comparisons

Parameter	Estimate (*SE*)
Country (England–France)	0.112 (0.072)
Gender (male–female)	−0.147 (0.035)
Age group (September–December = 1)	−0.195 (0.044)
Age coefficient (January–August)	−0.022 (0.008)
Age coefficient (September–December)	0.056 (0.021)
σ^2_{v1} (France Level 2 variance)	0.70 (0.08)
σ^2_{v2} (France Level 1 variance)	0.74 (0.04)
σ^2_{u1} (England Level 2 variance)	0.35 (0.05)
σ^2_{u2} (England Level 1 variance)	1.30 (0.05)
DIC (PD)	88,573.1 (6,146)

Note: Intercepts not shown. Burn in = 1,000; sample = 5,000. Loadings fixed to Level 1 loadings in Table 5. DIC = deviance information criterion; PD = effective number of parameters.

TABLE 11
Two-Level Factor Model for Country, Gender, and Age-Group Comparisons

Parameter	Estimate (*SE*)
Country (England–France)	0.001 (0.032)
Gender (male–female)	−0.059 (0.017)
Age group (September–December = 1)	−0.089 (0.026)
Age coefficient (January–August)	−0.010 (0.004)
Age coefficient (September–December)	0.019 (0.011)
σ^2_{v1} (France Level 2 variance)	0.70 (0.08)
σ^2_{v2} (France Level 1 variance)	0.75 (0.04)
σ^2_{u1} (England Level 2 variance)	0.36 (0.05)
σ^2_{u2} (England Level 1 variance)	1.29 (0.05)
DIC (PD)	88,559.2 (6,196)

Note: Explanatory variables in fixed part. Intercepts are not shown. Burn in = 1,000; sample = 5,000. Loadings fixed to Level 1 loadings in Table 5. DIC = deviance information criterion; PD = effective number of parameters.

Discussion

Our analyses and discussion have shown that comparisons between two educational systems with different pupil progression structures are complex. The combination of different ages of starting school and different allocation to year groups on the basis of birth date and repetition of grades makes any meaningful

comparison extremely difficult. Although we have here compared only England and France, our view is that the same problems occur when nonrepetition systems are compared with those that have important percentages of repetition, such as those of Spain, Portugal, and Belgium.

We have demonstrated that, even within a single proficiency domain, the data structure appears to contain at least two dimensions, although we have not conducted a full multilevel analysis of the dimensionality structure, nor have we attempted to identify and label factors as such. Nevertheless, even in the one-dimensional case, the (Rasch) assumption of equal item loadings is not supported by the data.

We have shown how a valid multilevel factor model can be fitted and, in particular, how to structure the factor variances at each level in order to properly study between-school variability. Model 13 is an example of a random coefficient factor model that can readily be extended to include further explanatory variables such as gender or age and also, for example, to cross-classifications. Thus, the full range of multilevel modeling procedures can be incorporated into these analyses, and such analyses will often lead to inferences that differ from those based on single-level models. The procedure is also much simpler than the plausible value procedure proposed by the OECD because it requires only a single fitting of a multilevel model. An issue with this approach is the choice of loadings to use. In our case, we have chosen a set of loadings from a two-level model with a single factor at each level. Other choices are possible, such as including fixed predictors in the initial model. In general, it would be useful to perform sensitivity analyses to determine whether such choices substantially affect inferences. Once the loadings are chosen, they effectively define the latent factors, and it is meaningful to make comparisons across subgroups only if we then use the same set of loadings in all analyses. We have not taken into account the uncertainty in the estimates of the loadings. Rather, we take the view that the first stage that determines the values of these loadings provides a practically useful metric for further analysis. Nevertheless, it is important to have a suitably large sample to ensure that sampling variability is small. If necessary, we can incorporate prior information, for example, from previous studies, into the estimation of these parameters.

Finally, although the main thrust of this article is to present a methodology for handling complex multilevel data in comparative studies, we should not ignore the serious drawback of a lack of longitudinal data in surveys such as PISA and other similar surveys such as TIMSS. Without such measures of prior performance on the same sample of students, it is not possible to overcome the comparability problems that arise from the different ways in which educational systems are organized, as we have described. Likewise, without such prior measures, it is not possible to attribute any observed differences between systems or subgroups to the education systems per se rather than, for example, social, cultural, or economic factors. Goldstein (2004) discussed this issue in more detail in the context of the stated aims of the PISA study.

Appendix
A Markov Chain Monte Carlo Algorithm for Two-Level Factor Analysis With Extension to a Structural Equation Model

The basic steps of this algorithm are given by Goldstein and Browne (2004): They are extended here to include ordered categorical responses, constraints among fixed parameters, and structural model predictors.

We write a basic two-level factor model for normal responses as

$$\begin{aligned} y_{rij} &= \sum_h \beta_{hr} x_{hrij} + \sum_{f=1}^{F} \lambda_{fr}^{(2)} v_{fj}^{(2)} + \sum_{g=1}^{G} \lambda_{gr}^{(1)} v_{gij}^{(1)} + u_{rj} + e_{rij} \\ u_{rj} &\sim N(0, \sigma_{ur}^2), e_{rij} \sim N(0, \sigma_{er}^2), v_j^{(2)} \sim MVN_F(0, \Omega_2), v_{ij}^{(1)} \sim MVN_G(0, \Omega_1) \\ r &= 1, \ldots, R, i = 1, \ldots, n_j, j = 1, \ldots, J, \sum_{j=1}^{J} n_j = N, \end{aligned} \tag{A1}$$

where the data structure is that for a multivariate two-level model with responses nested within individuals within schools. The subscript r indexes the responses. We have F factors at Level 2 and G factors at Level 1 with corresponding coefficients or loadings. Where F or G is > 1, we must introduce constraints on the loadings for second and subsequent factors. A common choice is to set, for the jth factor $(j > 1)$, $\lambda_k = 0$, $k = 1, \ldots j - 1$.

In the standard implementation, we assume independent factors with known variance matrices $= I$. The following steps generalize this to allow factor variances and covariances to be estimated. Gibbs sampling is used, except for factor covariances where Metropolis-Hastings sampling is used. The response variables can be normally distributed, binary, or ordered categorical, with any mixture of these. In addition, we allow a structural equation model of the following type to be fitted. For a set of factors, say $\nu = \{\nu_1, \ldots \nu_G\}$ at Level 1 (dropping the superscript), we can write the following model for a set of structural explanatory variables $\{Z_k\}$:

$$v_{gij} = \sum_k \gamma_{gk} z_{gkij} + v_{gij}^*, \; v_{ij}^* \sim MVN_G(0, \Omega_1), \tag{A2}$$

where we refer to the coefficients λ_{gk} as structural parameters. After substitution, the Level 1 component of the first line of Model A1 becomes

$$\begin{aligned} \widehat{y}_{rij} &= \sum_h \beta_{hr} x_{hrij} + \sum_k \sum_g \lambda_{gr} \gamma_{gk} z_{gkij} + \sum_g \lambda_{gr} v_{gij}^* \\ y_{rij} &= \widehat{y}_{rij} + e_{rij}, \; v_{ij}^* \sim MVN_G(0, \Omega_1). \end{aligned} \tag{A3}$$

In the following steps, we give details of how to implement the algorithm. The basic code is written in MATLAB (Mathworks, 2004) and is being incorporated into MLwiN (Browne, 2004; Rasbash et al., 2004) by extending the existing factor-fitting procedures. Default diffuse priors are assumed throughout (Browne, 2004).

From suitable starting values, the following steps are carried out. Default starting values are to set factor scores and factor loadings to 1. Fixed coefficient starting values are estimated from overall response proportions assuming a model with intercept terms only.

Step 1

In this step, any binary or ordered responses are replaced by a value sampled from a normal distribution, conditional on current parameter values as follows.

Binary Response

For each binary response, we sample from a standard normal distribution. Where a binary response is missing, the normal value is imputed in the next step.

1. Compute the current predicted value for binary response variable r

$$\hat{y}_{rij} = X_r^T \hat{\beta}_r + \sum_{f=1}^{F} \hat{\lambda}_{fr}^{(2)} \hat{v}_{fj}^{(2)} + \sum_{g=1}^{G} \hat{\lambda}_{gr}^{(1)} \hat{v}_{gij}^{(1)} + \hat{u}_{rj},$$

where ^ denotes the current value.

2. Compute, for all i, j,

$$P^* = \int_{-\infty}^{-\hat{y}} \phi(t)dt.$$

3. Generate N uniform random numbers (0, 1) into R^*, where N is the number of Level 1 units.
4. Calculate $T^* = Y((J - P^*)R^* + P^*) + (J - Y)P^*R^*$, where J is an $(N \times 1)$ vector of ones. This provides a set of uniform random numbers from $(0, P^*)$ or $(P^*, 1)$, depending on Y, the vector of observed responses.
5. Choose e^*, the required draw from N(0, 1) and hence $Y^* = \hat{y} + e^*$, from the inverse normal distribution, given T^*.

Note that this constrains the Level 1 variance to be equal to 1.

Ordered Responses

Suppose we have a p-category response, numbered $1, \ldots p$. As above, we consider the probit link proportional odds model:

$$\gamma^{(j)} = \int_{-\infty}^{\alpha_j - \hat{y}} \varphi(t)dt$$

$$\gamma^{(j)} = \sum_{f=1}^{j} \pi_f \text{ categories } j = 1, \ldots p-1,$$

where α_j is the "threshold" parameter defining the jth category and where $\hat{y}$ is the current predicted value, and we assume that the intercept term is incorporated in the fixed-part predictor so that $\alpha_1 = 0$.

Albert and Chib (1993) have shown that we can convert this to a standard normal model by sampling to obtain e^* (and hence Y^* as above), as follows.

> For a Category 1 response, we sample from the standard normal distribution $[-\infty, -\hat{y}]$.
> For a Category p response, we sample from the standard normal distribution $[\alpha_{p-1} - \hat{y}, \infty]$.
> For every other Category j, we sample from the standard normal distribution $[\alpha_{j-1} - \hat{y}, \alpha_j - \hat{y}]$.

Note that this sampling becomes equivalent to that for the binary case for $p = 2$.

For the $\{\alpha_j\}$, conditional on current values of $Y^* = \hat{y} + e^*$, we must select a new $\alpha_j (j > 1)$ so that the order relationships among the threshold parameters are preserved. This implies that each new value α_j must satisfy in turn $\alpha_j \geq \alpha_{j-1}, \alpha_j \geq \max(Y^*$ for all responses in Category $j)$, and hence a lower bound is given by $\alpha_j = \max\{\alpha_{j-1}, \max(Y^*$ for all responses in Category $j)\}$.

Likewise, an upper bound is given by $\alpha_j = \min\{\alpha_{j+1}, \min(Y^*$ for all responses in Category $j+1)\}$, where for α_{p-1} the upper bound is just $\alpha_j = \min(Y^*$ for all responses in Category $p)$.

Albert and Chib (1993) show that sampling of each α_j is from a uniform distribution on the intervals defined by these lower and upper bounds.

For the α_j, good starting values are important. Where the number of Level 1 units is large, the intervals for sampling the α_j will tend to be very small. With poor starting values, this implies that the α_j values and the intercept fixed term will converge very slowly. To obtain good starting values, we can consider randomly subsampling the data and fitting to a small subset. An alternative is to randomly subsample the Y^* in calculating the α_j, which will generally provide more accurate starting parameter values. The choice of sampling fraction is an area for further research—a choice that yields between 500 and 1,500 Level 1 units has been found to be satisfactory.

An alternative to using the Albert and Chib (1993) approach is to adopt Metropolis-Hastings sampling (Cowles, 1996), and this typically results in much faster convergence.

Thus, conditional on the current parameters, the component of the likelihood associated with a particular ordered category response is given by

$$P_\alpha = \prod_{i=1}^{N} \prod_{k=1}^{p} \pi_{\alpha,k}^{w_{i,k}}$$

for given α, where $w_{i,k} = 1$ *iff* response for unit i is in category k. We have

$$\pi_k = \int_{\alpha_{k-1} - (X_1\beta_1 + ZU)}^{\alpha_k - (X_1\beta_1 + ZU)} \varphi(t)dt,\ 1 < k < p$$

$$\pi_1 = \int_{-\infty}^{-(X_1\beta_1 + ZU)} \varphi(t)dt,$$

$$\pi_p = \int_{\alpha_{p-1} - (X_1\beta_1 + ZU)}^{\infty} \varphi(t)dt.$$

We select a new set of values α^* (one at a time) using a suitable (e.g., normal) proposal distribution and set new threshold parameters $= \alpha^*$ with probability $\min(1, P_{\alpha^*}/P_\alpha)$. The choice of proposal distribution variance may be derived adaptively but is not crucial, and in practice, using $5.8/N$ has been found to be suitable.

Step 2: Missing Data

In general, not all individuals will have every response observed. If we assume missing conditionally at random (or uninformatively through the study design, as in rotation or matrix sampling typically used in large-scale educational assessment surveys), then we can assume a uniform prior for the missing values. If we have an individual with missing response r, then update y_{rij} $(r = 1, \ldots, R,\ \ i = 1, \ldots, n_j,\ \ j = 1, \ldots, J\ \ \forall\ y_{rij}$ that are missing) from the following distribution, given the current values,

$$y_{rij} \sim N(\hat{y}_{rij}, \sigma^2_{er}).$$

Step 3: Fixed Coefficients

No Structural Parameters

Update the current value of $\beta_r = \{\beta_{hr}\}^T (r = 1, \ldots, R)$ from the following distribution:

$$p(\beta_r) \sim N\left(D_r \sigma_{er}^{-2} X_r^T \tilde{y}_r, D_r\right),$$

where

$$D_r = \sigma_{er}^2 \left(X_r^T X_r\right)^{-1}$$

and

$$\tilde{y}_r = \{\tilde{y}_{rij}\},\ \tilde{y}_{rij} = e_{rij} + X_r \beta_r,$$

where the Level 1 residuals, e_{rij}, are recomputed at each step by subtraction using the current predicted values $\hat{y}_{rij}$.

For some models, we require equality of the coefficients for a given predictor variable across responses—for example, to fit a model to detect an overall difference between, say, men and women. That is, for variable x_h, we require $\beta_{hr} = \beta_h, \forall r$. The exponent of the likelihood for a given Level 1 unit, omitting identification subscripts, for this set of parameters can be written as

$$-\frac{1}{2} \sum_r \sigma_{er}^{-2} (\tilde{y}_r - x_h \beta_h)^2,$$

which leads us to sample β_h from a normal distribution with mean

$$(x_h^T x_h)^{-1} x_h^T \tilde{y}^*,\ \tilde{y}^* = \left(\sum_r \tilde{y}_r \sigma_{er}^{-2}\right)\left(\sum_r \sigma_{er}^{-2}\right)^{-1}$$

and variance

$$(x_h^T x_h)^{-1} \left(\sum_r \sigma_{er}^{-2}\right)^{-1}.$$

Note that we can constrain any subset of the coefficients for an explanatory variable by replacing the variable with a separate predictor with zeros for the nonconstrained elements and a further predictor with zeros for the constrained elements. Note also that separate updating for each coefficient β_h is now required rather than the block updating for each response variable as above, where there are no constraints. In the general case where the explanatory variables are not constant over responses, we sample β_h from a normal distribution with mean

$$\left\{\sum_r \sigma_{er}^{-2} (x_{hr}^T x_{hr})\right\}^{-1} \left\{\sum_r \sigma_{er}^{-2} x_{hr}^T \tilde{y}_r\right\}$$

and variance

$$\left\{\sum_r \sigma_{er}^{-2}(x_h^T x_h)\right\}^{-1}.$$

Structural Parameters

For the structural Model A3, we perform the following. We consider the Level 1 case; Level 2 follows straightforwardly.

When sampling γ_{gk}, we treat this in the same way as the other fixed coefficients with explanatory variable $\lambda_{gr}z_{gkij}$ and coefficients constrained to be equal across responses. For the factor scores and variances, we have the same steps as below. The loadings are updated as below, using the v_{gij} in Model A2.

Step 4: Level 2 Loadings

Update $\lambda_{fr}^{(2)}$ $(r = 1, \ldots, R, f = 1, \ldots, F)$ from the following distribution:

$$p(\beta_r) \sim N\left(D_r^{(2)}\sigma_{er}^{-2}X_r^T\tilde{y}_r, D_r^{(2)}\right),$$

where

$$D_r^{(2)} = \sigma_{er}^2\left(v^{(2)^T}v^{(2)}\right)^{-1},\ \tilde{y}_r = \{\tilde{y}_{rij}\},\ \tilde{y}_{rij} = e_{rij} + \{v_{rj}^{(2)}\lambda_r^{(2)}\},$$
$$\lambda_r^{(2)} = \{\lambda_{1r}^{(2)}, \ldots, \lambda_{Fr}^{(2)}\},\ \text{and}\ \{v_{rj}^{(2)}\lambda_r^{(2)}\}, v^{(2)} = \{v_{rj}^{(2)}\}$$

are expanded to the length of the data set. Where loadings are constrained, the remaining lambdas are sampled by removing columns corresponding to the constrained loadings from $v^{(2)}$ or, correspondingly, by setting the corresponding columns of $v^{(2)}$ to zero and using a generalized (Pseudo or Moore-Penrose) inverse.

Step 5: Level 1 Loadings

Update $\lambda_{gr}^{(1)}$ $(r = 1, \ldots, R,\ g = 1, \ldots, G)$ from the following distribution:

$$p(\lambda_r^{(1)}) \sim N\left(D_r\sigma_{er}^{-2}v^{(1)^T}\tilde{y}_r, D_r\right),$$

where

$$D_r^{(1)} = \sigma_{er}^2\left(v^{(1)^T}v^{(1)}\right)^{-1}, \tilde{y}_r = \{\tilde{y}_{rij}\},\ \tilde{y}_{rij} = e_{rij} + \{v_{rij}^{(1)}\lambda_r^{(1)}\}, \lambda_r^{(1)} = \{\lambda_{1r}^{(1)}, \ldots, \lambda_{Gr}^{(1)}\}.$$

This is similar to Step 4 with the same sampling for constrained loadings.

Step 6: Level 2 Factor Scores and Covariance Matrix

Factor Scores

Update $v_j^{(2)}$ $(j = 1, \ldots, J)$ from the following distribution:

$$p(v_j^{(2)}) \sim MVN_F\left(D_j^{(2)}\left(\sum_r \sum_{i=1}^{n_j} \frac{\lambda_r^{(2)} d_{rij}^{(2)}}{\sigma_{er}^2}\right), D_j^{(2)}\right),$$

where

$$D_j^{(2)} = \left(\sum_r \frac{n_j \lambda_r^{(2)} (\lambda_r^{(2)})^T}{\sigma_{er}^2} + \Omega_2^{-1}\right)^{-1}$$

and

$$d_{rij}^{(2)} = e_{rij} + \sum_{f=1}^{F} \lambda_{fr}^{(2)} v_{fj}^{(2)},\ \lambda_r^{(2)} = (\lambda_{1r}^{(2)}, \ldots, \lambda_{Fr}^{(2)})^T,\ v_j^{(2)} = (v_{1j}^{(2)}, \ldots, v_{Fj}^{(2)})^T.$$

In the standard implementation, the factor variance matrix is assumed known, $\Omega_2 = I$. Where the variances are fixed to unity and nonzero covariances are to be estimated, see Step 8.

If any loadings are fixed, then the factor covariance matrix is estimated.

Estimated Factor Covariance Matrix—Diagonal

Assuming a diagonal matrix $\Omega_2 = diag(\sigma_{21}^2, \ldots, \sigma_{2F}^2)$, we sample σ_{2f}^2 from

$$f(\sigma_{2f}^2) \sim \Gamma^{-1}(\hat{a}_{2f}, \hat{b}_{2f}),$$

where

$$\hat{a}_{2f} = J/2 + \varepsilon,\ \hat{b}_{2f} = \sum_j v_{fj}^2/2 + \varepsilon$$

and where we assume a prior distribution

$$p(\sigma_{2f}^2) \sim \Gamma^{-1}(\varepsilon, \varepsilon).$$

Estimated Factor Covariance Matrix—General

In this case, we sample from

$$\Omega_2^{-1} \sim Wishart(w_2, S_2)$$
$$w_2 = m + \delta_p,\ S_2 = \left(\sum_{j=1}^{m} \nu_j \nu_j^T + S_p\right)^{-1},$$

where ν_j is the vector of Level 2 factor scores for the jth Level 2 unit and we assume a prior $p(\Omega_2^{-1}) \sim Wishart(\delta_p, S_p)$ and where w_2 is the degrees of freedom—the sum of the number of Level 2 units (m) and degrees of freedom associated with the prior. A minimally informative or maximally diffuse choice for the prior would be to take δ_p equal to the order of Ω_2 and S_p equal to a value chosen to be close to the final estimate multiplied by δ_p. Because this is generally unknown, an alternative is to choose $\delta_p = -3,\ S_p = 0$, which is equivalent to choosing a uniform prior for Ω_2.

Step 7: Level 1 Factor Scores and Covariance Matrix

Factor Scores

Update $v_{ij}^{(1)}$ $(i = 1, \ldots, n_j, j = 1, \ldots, J)$ from the following distribution:

$$p(v_{ij}^{(1)}) \sim MVN_G\left(D_{ij}^{(1)}\left(\sum_r \frac{\lambda_r^{(1)} d_{rij}^{(1)}}{\sigma_{er}^2}\right), D_{ij}^{(1)}\right),$$

where

$$D_{ij}^{(1)} = \left(\sum_r \frac{\lambda_r^{(1)} (\lambda_r^{(1)})^T}{\sigma_{er}^2} + \Omega_1^{-1}\right)^{-1}$$

and

$$d_{rij}^{(1)} = e_{rij} + \sum_{g=1}^{G} \lambda_{gr}^{(1)} v_{gij}^{(1)},\ \lambda_r^{(1)} = (\lambda_{1r}^{(1)}, \ldots, \lambda_{Gr}^{(1)})^T,\ v_{ij}^{(1)} = (v_{1ij}^{(1)}, \ldots, v_{Gij}^{(1)})^T.$$

In the standard implementation, the factor variance matrix is assumed known, $\Omega_1 = I$. Where the variances are fixed to unity and nonzero covariances are to be estimated, see Step 8.

If any loadings are fixed, then the factor covariance matrix is estimated.

We note that for the item cluster effect model, where $v_{ij}^{(1)}$ is replaced by $v_{ij}^{(1)} + \gamma_{kij}$ for cluster k, the sampling of γ_{kij} has an analogous form.

Estimated Factor Covariance Matrix—Diagonal

We assume a diagonal matrix $\Omega_1 = diag(\sigma_{11}^2, \ldots, \sigma_{1G}^2)$ and update σ_{1g}^2 from the following distribution:

$$f(\sigma_{1g}^2) \sim \Gamma^{-1}(\hat{a}_{1g}, \hat{b}_{1g}),$$

where

$$\hat{a}_{1g} = N/2 + \varepsilon,\ \hat{b}_{1g} = \sum_{ij} e_{gij}^2/2 + \varepsilon$$

and where we assume a prior distribution

$$p(\sigma_{1g}^2) \sim \Gamma^{-1}(\varepsilon, \varepsilon).$$

Estimated Factor Covariance Matrix—General

In this case, we sample from

$$\Omega_1^{-1} \sim Wishart(w_1, S_1)$$
$$w_1 = N + \delta_p,\ S_1 = \left(\sum_{ij} \nu_{ij} \nu_{ij}^T + S_p \right)^{-1},$$

where ν_{ij} is the vector of Level 1 factor scores for the ijth Level 1 unit and we assume a prior $p(\Omega_1^{-1}) \sim Wishart(\delta_p, S_p)$ and where w_1 is the degrees of freedom—the sum of the number of Level 1 units (N) and degrees of freedom associated with the prior. A minimally informative or maximally diffuse choice for the prior would be to take δ_p equal to the order of Ω_1 and S_p equal to a value chosen to be close to the final estimate multiplied by δ_p. Because this is generally unknown, an alternative is to choose $\delta_p = -3,\ S_p = 0$, which is equivalent to choosing a uniform prior for Ω_1.

Step 8: Factor Covariances

If we allow covariances between the factors, with variances known, then we have the following Metropolis step (set out for Level 2—Level 1 is similar). A uniform prior is assumed:

$$p(\Omega_{2,lm}) \sim Uniform(-1, 1) \forall l \neq m$$

Here $\Omega_{2,lm}$ is the l,mth element of the Level 2 factor variance matrix. We update these covariance parameters using a Metropolis step and a normal random walk proposal as follows.

At iteration t, generate

$$\Omega_{2,lm}^* \sim N(\Omega_{2,lm}^{(t-1)}, \sigma_{plm}^2),$$

where σ_{plm}^2 is a proposal distribution variance that has to be set for each covariance. Then, if $\Omega_{2,lm}^* > 1$ or $\Omega_{2,lm}^* < -1$, set $\Omega_{2,lm}^{(t)} = \Omega_{2,lm}^{(t-1)}$ as the proposed covariance is not valid; else, form a proposed new matrix Ω_2^* by replacing the l,mth element of $\Omega_2^{(t-1)}$ by this proposed value. Likewise, if Ω_2^* is not positive definite, then again set $\Omega_{2,lm}^{(t)} = \Omega_{2,lm}^{(t-1)}$, otherwise set $\Omega_{2,lm}^{(t)} = \Omega_{2,lm}^*$ with probability $\min(1, p(\Omega_2^*|\nu_j^{(2)})/p(\Omega_2^{(t-1)}|\nu_j^{(2)})$ and $\Omega_{2,lm}^{(t)} = \Omega_{2,lm}^{(t-1)}$ otherwise. The components of the likelihood ratio are

$$p(\Omega_2^*|\nu_j^{(2)}) = \prod_j |\Omega_2^*|^{-1/2} \exp -((\nu_j^{(2)})^T(\Omega_2^*)^{-1}\nu_j^{(2)})/2$$

and

$$p(\Omega_2^{(t-1)}|\nu_j^{(2)}) = \prod_j |\Omega_2^{(t-1)}|^{-1/2} \exp -((\nu_j^{(2)})^T(\Omega_2^{(t-1)})^{-1}\nu_j^{(2)})/2.$$

This procedure is repeated for each nonzero covariance. An adaptive procedure (Browne, 2004) can be used to select the proposal distribution parameters.

Step 9: Level 2 Residuals

Update u_{rj} $(r = 1, \ldots, R, j = 1, \ldots, J)$ from the following distribution:

$$p(u_{rj}) \sim N\left(\frac{D_{rj}^{(u)}}{\sigma_{er}^2}\sum_{i=1}^{n_j} d_{rij}^{(u)}, D_{rj}^{(u)}\right),$$

where

$$D_{rj}^{(u)} = \left(\frac{n_j}{\sigma_{er}^2} + \frac{1}{\sigma_{ur}^2}\right)^{-1}$$

and

$$d_{rij}^{(u)} = e_{rij} + u_{rj}.$$

Step 10: Level 2 Residual Variances

Update σ_{ur}^2 from the following distribution:

$$f(\sigma_{ur}^2) \sim \Gamma^{-1}(\hat{a}_{ur}, \hat{b}_{ur}),$$

where

$$\hat{a}_{ur} = J/2 + \varepsilon,\ \hat{b}_{ur} = \sum_j u_{rj}^2/2 + \varepsilon$$

and we assume a prior $p(\sigma_{ur}^2) \sim \Gamma^{-1}(\varepsilon, \varepsilon)$.

Step 11: Level 1 Variances

Update σ_{er}^2 from the following distribution:

$$f(\sigma_{er}^2) \sim \Gamma^{-1}(\hat{a}_{er}, \hat{b}_{er}),$$

where

$$\hat{a}_{er} = N/2 + \varepsilon,\ \hat{b}_{er} = \sum_{ij} e_{rij}^2/2 + \varepsilon$$

and we assume a prior $p(\sigma_{er}^2) \sim \Gamma^{-1}(\varepsilon, \varepsilon)$.

Step 12: Deviance Information Criterion

We suggest computing the deviance information criterion (DIC; Spiegelhalter et al., 2002) at each step, over the observed nonmissing responses. This gives

$$\sum_{rij} (e_{rij}^2/\sigma_{er}^2 + \log_e(2\pi\sigma_{er}^2))$$

over the normal responses and

$$-2\sum_{rij} \log_e[\pi_{rij} w_{rij} + (1 - \pi_{rij})(1 - w_{rij})]$$

over the binary responses, where the predicted probabilities are obtained from the probit (cumulative normal) function, and $w_{rij} = 1$ *iff* response in category r.

For ordered responses, the deviance contribution is given by

$$-2\sum_{rij} w_{rij} \log_e(\pi_{rij}),\ w_{rij} = 1\ \textit{iff}\ \text{response in category } r$$

computed over the ordered responses, where the predicted probabilities are obtained from the probit (cumulative normal) function using the estimated threshold values.

The deviance of the chain means is calculated using the final parameter values to calculate the residuals. We then have $DIC = \bar{D} + p_D$, $p_D = \bar{D} - D(\bar{\theta})$, where p_D is taken as the effective number of parameters.

We note that where the responses are all either binary or ordered, $e_{rij} \sim N(0, 1)$ so that it is only the missing responses that provide a pseudo-normal response with DIC contribution,

$$\sum_{rij} (e_{rij}^2 + \log_e(2\pi)).$$

Thus, the contribution is independent of the model parameters and may be omitted from the DIC computation.

Note

[1]Item details are released by the Organization for Economic Cooperation and Development for only a small sample of the items.

References

Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris: Organization for Economic Cooperation and Development.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Browne, W. J. (2004). *MCMC estimation in MLwiN*. London: Institute of Education.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov Chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, *6*, 101–110.

Fox, J., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 271–288.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education*, *11*, 319–330.

Goldstein, H., & Browne, W. (2004). Multilevel factor analysis models for continuous and discrete data. In A. Olivares (Ed.), *Advanced psychometrics: A Festschrift to Roderick P. McDonald* (pp. 453–475). Mahwah, NJ: Lawrence Erlbaum.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*, 223–231.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, *53*, 455–467.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mathworks. (2004). *MATLAB program* (Version 7.0). Natick, MA: Author.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology*, *42*, 215–232.

Mislevy, R. J. (1991). Randomisation based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., et al. (2001). *TIMSS assessment frameworks and specifications 2003*. Chestnut Hill, MA: Boston College.

Muthen, B. O. (1989). Latent variable modelling in heterogeneous populations. *Psychometrika*, *54*, 557–585.

Muthen, B. O. (1997). Latent variable modelling of longitudinal and multilevel data. In A. E. Raftery (Ed.), *Sociological methodology* (pp. 453–480). Cambridge, MA: Blackwell.

Muthen, B. O. (2002). Beyond SEM: General latent variable modelling. *Behaviormetrika*, *29*, 81–117.

Office for National Statistics. (2002). *Student achievement in England*. London: Author.

Organization for Economic Cooperation and Development. (2001). *Knowledge and skills for life: First results from Programme for International Student Assessment*. Paris: Author.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika*, *69*, 167–190.

Rasbash, J., Browne, W., & Steele, F. (2004). *A user's guide to MLwiN version 2.0*. London: Institute of Education.

Raudenbush, S. W. (1995). Maximum likelihood estimation for unbalanced multilevel covariance structure models via the EM algorithm. *British Journal of Mathematical and Statistical Psychology*, *48*, 359–370.

Rowe, K. J. (2003). Estimating interdependent effects among multilevel composite variables in psychosocial research: An example of the application of multilevel structural equation modeling. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 255–284). Mahwah, NJ: Lawrence Erlbaum.

Scott, S. L., & Ip, E. H. (2002). Empirical Bayes and item clustering effects in a latent variable hierarchical model: A case study from the National Assessment of Educational Progress. *Journal of the American Statistical Association*, *97*, 409–419.

Song, X., & Lee, S. (2004). Bayesian analysis of two-level nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, *57*, 29–52.

Spiegelhalter, D., Best, N., Carlin, B. P., & Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, *B*, *64*, 583–640.

Steele, F., & Goldstein, H. (2006). A multilevel factor model for mixed binary and ordinal indicators of women's status. *Sociological Methods & Research*, *35*, 137–153.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, *1*, 81–97.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*, 247–260.

Zhu, H.-T., & Lee, S.-Y. (1999). Statistical analysis of nonlinear factor analysis models. *British Journal of Mathematical and Statistical Psychology*, *52*, 225–242.

Authors

HARVEY GOLDSTEIN retired in 2005 from the Institute of Education, University of London, and is now a professor of social statistics at the University of Bristol, United Kingdom, Graduate School of Education, 35 Berkeley Square, Bristol BS8 1JA; h.goldstein@bristol.ac.uk. His major research interest is in the methodology and application of multilevel statistical models, and a secondary interest is in statistical models for educational assessment.

GÉRARD BONNET trained as a linguist and started his career at the Queen's University of Belfast. He is currently head of the International Relations Unit of the Schools Directorate General of the French ministry of education. Previously, he represented France at the Programme for International Student Assessment (PISA) Governing

Board, the International Association for the Evaluation of Educational Achievement (IEA) General Assembly, and several expert groups of the European Commission. He was also for a time a Visiting Fellow at the Institute of Education, University of London, and at the Graduate School of Education of the University of Bristol.

THIERRY ROCHER trained as a statistician with the French Statistical Institute (INSEE) and later joined the education ministry, where he works as a specialist in the measurement of pupils' attainments. He is also currently French representative on the Organization for Economic Cooperation and Development's INES Network A and a member of the PISA Technical Advisory Group.

Manuscript received January 11, 2005
Accepted October 28, 2005