School Effectiveness and School Improvement
2004, Vol. 15, Nos. 3–4, pp. 261–280

Taylor & Francis
Taylor & Francis Group

Applications of Multilevel IRT Modeling

Jean-Paul Fox
Department of Research Methodology, Measurement and Data Analysis,
University of Twente, Enschede, The Netherlands

ABSTRACT

The recent development of multilevel IRT models (Fox & Glas, 2001, 2003) has been shown to be very useful for analyzing relationships between observed variables on different levels containing measurement error. Model parameter estimates and their standard deviations are concurrently estimated taking account of measurement error in observed variables. The multilevel IRT models are, in particular, useful in the analysis of school effectiveness research data since hierarchical structured educational data are subject to error. By re-examining some school effectiveness studies, the basic aspects of this new model and consequences of measurement error are shown.

INTRODUCTION

Over the last decade, there has been an increasing interest in the accountability of educational institutions. Research has focused on measuring the ''quality'' of schools and making quantitative comparisons between schools. Most of the debate in the literature concerns the best choice of indicator measures, as a statistical measurement on a school which is intended to be related to the quality of its functioning. These so-called ''performance indicators'' can be used to judge school effectiveness and to determine what measures can be taken for improvement, obtaining knowledge about the relative size of school differences, and to what extent other indicators may explain differences.

Address correspondence to: Jean-Paul Fox, Department of Research Methodology, Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: fox@edte.utwente.nl

Manuscript submitted: September 30, 2003
Accepted for publication: December 11, 2003

10.1080/0924345051233138321 2$16.00

Within this context, attention has focused on contextual differences and the appropriate specification of a statistical model.

The object in school effectiveness research is to explore differences within and between schools, by investigating the relationship between explanatory and outcome factors. This involves choosing an outcome variable, such as examination achievement, and studying differences among schools after adjusting for relevant background variables. A general acceptable statistical model in the assessment of school effectiveness requires the deployment of multilevel analysis techniques. A multilevel model describes the relationships between one or more ''outcome'' variables (examination results, attitudes), school and teacher characteristics (teachers' attitudes, financial resources, class size), and students' characteristics (achievements, social background). The model is appropriate for the analysis of multistage nested designs. The sampling design reflects the structure with students nested in classes, and classes within schools. A multilevel analysis takes this structure into account, and variance components are modeled at each sampling level, that is schools and classes are regarded as random effects. Furthermore, the model takes into account the homogeneity of results of individual pupils in the same school, since pupils in the same school share common experiences. The appropriateness of multilevel models in the assessment of school effectiveness was shown by Aitkin and Longford (1986). From that time, most of the research has focused on multilevel modeling of hierarchical structured educational data and the assessment of relevant input and output indicators (e.g., Goldstein, 1995; Longford, 1993; Scheerens & Bosker, 1997; Snijders & Bosker, 1999). Aitkin and Longford (1986) and Goldstein (1979) pointed out the importance of reliable factors that reflect characteristics of individuals and schools. However, the ''traditional'' multilevel analysis techniques assume that the indicators are measured without error. Quantifiable uncertainties place limitations on the precision and the correctness with which schools are compared. Educational systems can be compared with each other regarding their performances, and adjusted for context, given accurate measures of all indicators. The importance of reliable measurements is well known (Goldstein, 1979), but obtaining and handling measurement error variances in the analysis was problematic. Therefore, Fox and Glas (2001, 2003) developed a new approach to multilevel modeling that takes into account the uncertainty regarding the performance indicators on different levels.

The present article illustrates the importance of taking measurement error into account in multilevel analysis. Therefore, three different interesting

applications within educational research are presented. They show the model specification options and the effects of modeling explicitly measurement errors. After this introduction, a description of measurement error models will be given. In the next section, a new approach to multilevel modeling is described. Then, three real data examples will be given. The last section contains a discussion and suggestions for further research.

MEASUREMENT ERROR

Most measurements in educational research are subject to error. That is, independently and repeatedly measuring does not produce an identical result, but different observed values would be obtained. Scores obtained from a measurement instrument can be affected by item inconsistency, fluctuations within individuals, and the environment where the test took place. The effects of measurement error have been studied (e.g., Goldstein, 1979) and it has been shown that the use of unreliable explanatory variables produces bias in the estimation of the regression coefficients. Differences between schools and individuals are blurred due to measurement error in the observed variables. Moreover, estimates of the variance components are too low since the observed variables are treated as measured without error. Therefore, factors or school effects with no influence, when taking the measurement error in the observed variables into account, may appear to be present when ignoring the measurement error. Or school effects appear to be significantly different but may be comparable when taking measurement error into account.

Item Response Theory

Measuring individual differences in ability is difficult, since examinees' abilities cannot be observed directly. Tests or questionnaires are needed to measure the individual abilities. In a conventional testing procedure, the examinees answer items scored as correct or incorrect, that is, dichotomously scored items. The easiest way to do this is to consider the number of correct answers as a test score representing ability. An important aspect is the accuracy of the test score. Test scores based on a few items are unreliable. Obviously, the accuracy of the test score increases as the number of answered items increases. This test score is based on the assumption that all items provide equal information in estimating ability. A more advanced score consists of a weighted score of the number of correct items. But there are other

important aspects that should be taken into account when measuring examinees' abilities. The characteristics of the test and of the examinees should be separated. This means that the test score is not test dependent, and makes it possible to compare examinees who take different tests. The errors of measurement differ for examinees of different ability. Consider an examinee who answers all items incorrectly. The test score indicates a low ability but provides no information about exactly how low. More precise information is obtained when an examinee gets some items right and some wrong.

These and other features can be obtained within the framework of item response theory (IRT). The class of item response theory (IRT) models describe the relationship between an examinee's ability and responses based on the characteristics of the items of the test. The dependence of the observed responses to dichotomous or polytomous scored items on an ability is fully specified by the item characteristic function, which is the regression of item scores on latent ability. The item response function is used to make inferences about latent ability from the observed item responses. The item characteristic functions cannot be observed directly because the ability parameter is not observed. But under certain assumptions it is possible to infer the information of interest from the examinee's responses to the test items (e.g., Lord, 1980; Lord & Novick, 1968).

One of the forms of the item response function for a dichotomous item is the two-parameter model. The probability that an examinee answers an item correctly depends on his ability, the difficulty of the item, and the discriminating behavior of the item. The difficulty parameter is the point on the ability scale where the probability of a correct response is 50%. The greater the value of the difficulty parameter, the greater the ability that is required to have a 50% chance of getting the item correct. The discrimination parameter is proportional to the slope of the item response function at the point of the difficulty parameter on the ability scale. Items with high discrimination parameter values are useful for separating examinees into different ability levels. In Figure 1, two different item characteristic curves are plotted. The curves differ by location on the ability scale and by the slopes. Item 2 is more difficult and shifted to the right on the ability scale. The item characteristic curve of Item 1 corresponds with a higher discrimination parameter. As a result, a small increase in ability leads to a higher increase in probability of scoring correct in comparison to item 2.

An assumption of the two-parameter IRT model is that only one ability is measured by a set of items in a test. The ability of an examinee is the only

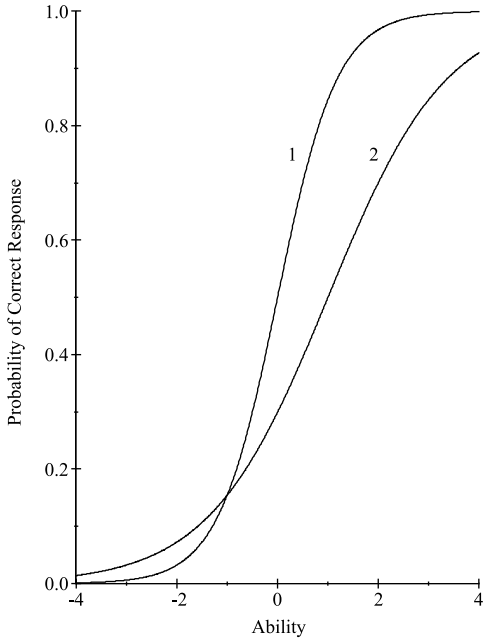

Fig. 1. Two-parameter item characteristic curves for two different items.

factor influencing its responses to test items. There exists no relationship between the examinee's responses given the examinee's ability. An IRT model may provide an adequate description of the test data. It is essential to test the fit of the model to the data. The examinees' abilities are unobservable but can be estimated. An IRT model provides a framework for the uncertainty regarding the estimate of the ability. So, an IRT model can be used to measure the abilities of the examinees, and quantifies the uncertainty regarding the estimate. For an overview of different IRT models, to handle the effect of guessing or polytomously scored items, see, for example, Hambleton and Swaminathan (1985) or Van der Linden and Hambleton (1997).

MULTILEVEL ITEM RESPONSE THEORY (IRT) MODEL

In most educational research, measurements are needed at the individual and the group levels. For example, students' pretest scores, socioeconomic status

(SES), or intelligence are often used as explanatory variables in predicting students' examination results. Students' examination results can be seen as an indicator for the students' abilities. They are measured subject to error. The explanatory variables, SES and intelligence, cannot be observed directly and test data are needed to estimate these predictors. In summary, observed test data can be seen as indicators for the latent variables, say, abilities or characteristics. The latent variables can be incorporated in a multilevel model.

The idea behind the multilevel IRT model is to measure each latent variable incorporated in the multilevel model with an IRT model. Let y_{ijk} denote the observed item response of the i^{th} student in the j^{th} school to item k. Let a two-parameter IRT model relate the observed dichotomous item response with the students' latent ability, θ_{ij}, that is,

$$P(Y_{ijk} = 1|\theta_{ij}, a_k, b_k) = \Phi(a_k\theta_{ij} - b_k) \tag{1}$$

where Φ is the cumulative normal distribution, a_k and b_k are the discrimination and difficulty parameter of item k, respectively. The latent variable can be incorporated in a multilevel model as a dependent variable to explore differences among students' abilities,

$$\theta_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \cdots + \beta_{Qj}X_{Qij} + e_{ij} \tag{2}$$

and

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}W_{1qj} + \cdots + \gamma_{qS}W_{Sqj} + u_{qj} \tag{3}$$

for $q = 0, \ldots, Q$. Both residuals, e_{ij} and u_{qj}, are assumed to be normally distributed. The other assumptions of the multilevel model can be found in Snijders and Bosker (1999). The Equations (1) to (3) define a multilevel IRT model, with a latent dependent variable measured by a two-parameter IRT model. Here, a latent variable is used as a dependent variable. Then, the IRT model can be seen as a level within the multilevel model. The item responses are Level-1 units, and the examinees are nested within the item responses. The multilevel IRT model presented is just one form of the various possibilities. Latent variables can also be incorporated as an explanatory variable at the student- or school-level. A multilevel IRT model can also consist of a multilevel model that defines the relation between different latent variables and various IRT models for measuring the latent variables. Furthermore, different IRT models can be used to measure a latent variable. For example, a graded

response IRT model can be used to handle polytomous response data. This measurement model assumes that the available item categories to which a student responds can be ordered. The probability that the student with latent ability θ_{ij} obtains a grade c, or gives a response falling into category c on item k is defined by

$$P(Y_{ijk} = c|\theta_{ij}, a_k, \mathbf{b}_k) = \Phi(b_{kc} - a_k\theta_{ij}) - \Phi(b_{kc-1} - a_k\theta_{ij}) \quad (4)$$

where b_{kc} and b_{kc-1} are the threshold parameters of category c. This graded response IRT model is, in particular, useful for responses to Likert-type items, whose responses are scored in graded fashion. These type of items were used in the third application where the graded response model is used as a measurement model.

A second characteristic of multilevel IRT modeling is that it can simultaneously estimate all model parameters. This means that the uncertainty in the measurements of the latent variables is taken into account in the estimation of the other model parameters. In most cases, individual abilities or group characteristics are estimated and imputed in the multilevel analysis. Subsequently, the measurements are assumed to be observed without an error. As a result, the estimated regression coefficients are biased and their standard deviations are too small.

Bayes Factor (BF) can be used when choosing between a set of competing multilevel IRT models. Suppose that there are two models M_1 and M_2 and let $\mathbf{y}$ denote the observed data. The quantity

$$BF = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} \quad (5)$$

is the Bayes factor and defines the ratio of the posterior odds of M_1 over M_2. In practice, values of the BF are evaluated on a log scale. If $\log_{10}(BF) > 2$, then it can be stated that model M_1 is more likely than model M_2.

The standard multilevel software (MLwiN, Goldstein et al., 1998; HLM, Raudenbush, Bryk, Cheong, & Congdon, 2000) cannot be used to estimate simultaneously all parameters of a multilevel IRT model. Therefore, software, which is freely available via the internet (Fox, 2003), has been developed to make multilevel IRT modeling accessible for other researchers. Details about the estimation method can be found in Fox (2001) and Fox and Glas (2001, 2003). The computations for the applications below are done using the multilevel IRT software.

THREE APPLICATIONS

The applications focus on school effectiveness research with fundamental interest in the development of knowledge and skills of individual students in relation to school characteristics. Data are analyzed at the individual level and it is assumed that classrooms, schools, and experimental interventions have an effect on all students exposed to them. In school or teacher effectiveness research, both levels of the multilevel model are of importance because the objects of interest are schools and teachers as well as students. Interest may exist in the effect on student learning of the organizational structure of the school, characteristics of a teacher, and the characteristics of the student.

A Dutch Primary School Mathematics Test

The mathematical skills of grade-8 students across schools were compared (Doolaard, 2002). One of the research questions in the study was whether schools that participate in the central primary school leaving test in the Netherlands on a regular basis perform better than schools that do not participate on a regular basis. This dataset was also analyzed by Fox and Glas (2001). Students of grade 8 in 97 schools were given a mathematics test consisting of 18 mathematics items (correct/incorrect) taken from the school-leaving examination developed by the National Institute for Educational Measurement (Cito). Of the 97 schools sampled, 72 schools regularly participated in the school-leaving examination (Cito; $0 = \text{no}$, $1 = \text{yes}$). The total number of students for which data were available was 2,156. Three student characteristics were used as predictors of student achievement: socioeconomic status (SES), non-verbal intelligence test (ISI), and Gender ($0 = \text{male}$, $1 = \text{female}$).

In the multilevel IRT analysis, a two-parameter IRT model was used to measure the math abilities of the students. A multilevel model was used to explain differences in the math abilities. The multilevel IRT analysis was compared with a multilevel analysis using observed sum scores, the sum of the number of correct answers, as a measurement for the math abilities. Further on, a multilevel analysis using observed sum scores will be denoted as a standard multilevel analysis, and a multilevel model including observed sum scores will be denoted as a standard multilevel model. The reliability of the scale scores (Cronbach's α) equals .88. Notice that the reliability coefficient relates to the observed sum scores and only makes sense in the standard multilevel analysis. For comparative purposes, the dependent variable in the

Table 1. A Dutch Primary School Mathematics Test. Parameter Estimates of the Standard Multilevel and Multilevel IRT Empty Model.

	Multilevel model		Multilevel IRT model	
Parameter	Estimate	S.E.	Estimate	S.E.
Fixed				
Intercept	−.060*	.051	−.064*	.059
Random				
Within schools	.812	.025	.767	.027
Between schools	.210	.036	.292	.051

Note. *Not significant at the .05 level.

standard multilevel and multilevel IRT model and the Level-1 predictors, ISI and SES, were normally standardized. Table 1 reports the parameter estimates of both empty models.

The grouping of students in schools explained a lot of variance in the math abilities. The multilevel IRT parameter estimates show that 28% of the individual variance is explained at the school-level. According to the standard multilevel analysis only 21% is explained at the school-level. Thus, a higher proportion of variance is explained at the school-level according to the multilevel IRT model. Notice that the dependent variables in both models are normally standardized. The standard errors are somewhat larger in the multilevel IRT model, since the measurement error in the dependent variable is taken into account in the estimation of the other model parameters.

To explain variance at the individual level, three Level-1 explanatory variables were introduced. Further, the Level-2 variable Cito was incorporated. The item responses for measuring SES and ISI were not available. Therefore, the computed scores were directly used in the analysis. The main result of the analysis (Table 2) was that conditionally on SES, ISI, and Gender, the Cito schools performed better than the non-Cito schools. That is, a significant positive effect on the students' math abilities was found from schools that participate on a regular basis in the central primary school-leaving test. The positive effects of the Level-1 predictors ISI and SES show that students with a high socioeconomic status and high score on the nonverbal intelligence test performed better on the math test. The effect of Gender was also significant and negative, meaning that the boys outperformed girls on the math test.

Table 2. A Dutch Primary School Mathematics Test. Parameter Estimates of the Standard Multilevel and Multilevel IRT Model, Including Level-1 and Level-2 Predictors.

	Multilevel model		Multilevel IRT model	
Parameter	Estimate	S.E.	Estimate	S.E.
Fixed				
Intercept	−.311	.082	−.322	.095
Student variables				
ISI	.411	.016	.460	.019
SES	.217	.021	.251	.021
Gender	−.159	.037	−.185	.038
School variables				
Cito	.462	.091	.488	.107
Random				
Within schools	.577	.018	.460	.019
Between schools	.115	.020	.173	.031

From Table 2 it can be seen that the standard multilevel analyses led to different outcomes. The estimated effects of the predictors corresponding to the multilevel IRT model are all higher than the effects of the multilevel model with an observed sum score as the dependent variable. Differences between students' mathematics abilities and the effects of the individual characteristics become more apparent. The measurement error in the observed sum scores causes that the effects of the Level-1 and Level-2 predictors are attenuated towards zero. The measurement based on the two-parameter IRT model resulted in a sharper distinction between students' outcomes than the observed sum scores. As a result, more variance is unexplained at Level-1 and the variance explained due to the grouping is considerably lower in the standard multilevel analysis. A substantial proportion of variance is explained by the explanatory variables, in the multilevel IRT analysis, 40% at the student level and 41% at the school-level, in the standard multilevel analysis, 29% and 45%, respectively. Thus, from the multilevel IRT analysis follows a much higher importance of the Level-1 explanatory variables.

In the multilevel IRT analysis, the measurement error in the estimated math abilities is taken into account in the estimation of the other parameters. Therefore, the standard errors of the parameter estimates corresponding to the multilevel IRT model are somewhat higher. These standard errors provide a more reliable basis for testing the importance of the predictor variables, since

the inaccuracy of the measurement of the math achievements is taken into account.

Contribution of Adaptive Instruction to Pupils' Performances

A mathematics test consisting of 40 items (correct/incorrect) was administered to 3,713 pupils of grade 4 in 198 regular primary schools (Bosker, Blatchford, & Meijnen, 1999). Interest was focused on the relation between achievement in mathematics, educational provision at the school-level and adaptive instruction by teachers. A test of 23 items (correct/incorrect) measuring the willingness and capability to introduce adaptive instruction (AI) was taken by teachers. Adaptive education, education that is adapted to differences in the classroom, is a topic that has received a lot of attention in primary education recently. The idea is that differences between pupils are taken into account in such a way that pupils learn in different ways and at a different place. Therefore, the education should consist of a variety of instructional approaches and learning methods in order to adapt education to the pupils. Teachers should be aware of differences between pupils and adapt instruction accordingly. The expectations about the contribution of adaptive instruction to pupils' performance are high. However, effects of adaptive instruction on pupils' performances in Dutch primary schools were not found in Fletcher-Campbell, Meijer, and Pijl's (1999) study.

An intelligence test consisting of 37 items (IQ, correct/incorrect) was administered to the pupils, and IQ was used as a student-level explanatory variable. The multilevel model was used to explain variance in the students' math abilities (the dependent variable) that was due to grouping. Furthermore, we investigated whether or not AI and IQ had a significant effect on pupils' math achievements. Therefore, a measurement was needed for math abilities, pupils' intelligence, and teachers' capability in adaptive instruction. In the ''standard'' multilevel modeling approach, sum scores were used as measurements for these latent variables. The measurements for the math abilities, intelligence, and adaptive instruction had a reliability of 0.88, 0.84, and 0.78, respectively. In the multilevel IRT modeling approach, the two-parameter IRT model was used to measure these latent variables. Again, in both multilevel analyses, all latent variables were normally standardized to make the outcomes comparable.

The standard multilevel analysis of the empty model (Table 3) reveals that 15% of the variance in mathematics achievement is explained at the school-level, the other 85% at the individual level. This result is comparable to results

Table 3. Contribution of Adaptive Instruction. Parameter Estimates of the Standard Multilevel and Multilevel IRT Empty Model.

	Multilevel model		Multilevel IRT model	
Parameter	Estimate	S.E.	Estimate	S.E.
Fixed				
Intercept	−.024*	.032	−.031*	.035
Random				
Within schools	.850	.020	.810	.021
Between schools	.150	.020	.190	.026

Note. *Not significant at the .05 level.

of Bosker and Witziers (1996). They found in their meta-analysis that 18% of the variance in achievement can be attributed to differences between schools in western countries. The multilevel IRT analysis shows that 19% of the variance is explained at the school-level. This example also shows that more variance is explained at the school-level in the multilevel IRT analysis, meaning that schools differ more than expected on the basis of a standard multilevel analysis.

The empty models (Table 3) assume that the mathematical skill of the students can be broken down in a school contribution, $\gamma_{00} + u_{0j}$, and a deviation, e_{ij}, for each student from their school's contribution,

$$\theta_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \tag{6}$$

The multilevel IRT model used the item responses and an IRT model, Equation (1), and the standard multilevel model used the student's observed sum score to estimate the latent ability of student ij. Both models do not differ much in the estimate of γ_{00}, but differ in their school's contribution. Table 4 gives the school effects, u_{0j}, and rankings of the first 10 schools in the sample for both models. It can be seen that the estimated school effects differ and this may influence the statistical inference when adjusting the school effects for any school characteristics. The lowest and highest school effects under both models correspond with the same schools. However, the effects are smaller for the standard multilevel model since the observed sum score distribution is negatively skewed.

To explain differences in mathematics achievement, the explanatory variables IQ and AI were introduced. The standard multilevel analysis is based on three observed sum scores for the dependent and explanatory variables

Table 4. Contribution of Adaptive Instruction. Realized Values for the School Effects, Rankings of the First 10 Schools, and the School With the Lowest and Highest Rank Number.

	Multilevel model		Multilevel IRT model	
School	School effect	Rank	School effect	Rank
1	.341	23	.419	16
2	.352	21	.332	31
3	.414	14	.413	18
4	−.013	111	.014	88
5	−.066	128	−.096	112
6	−.225	160	−.282	154
7	−.467	183	−.374	167
8	−.182	153	−.264	148
9	−.251	163	−.312	159
10	−.220	159	−.361	165
173	.645	1	.977	1
158	−1.408	198	−1.115	198

given the observed item responses. The multilevel IRT analysis comprehends the estimation of three two-parameter IRT models and the parameters of the multilevel model containing the corresponding latent variables. The multilevel IRT software enables the researcher to estimate simultaneously all parameters. So, the sources of the measurement errors are modeled with IRT models and the measurement errors are taken into account in the estimation of the multilevel model parameters.

Table 5 presents the parameter estimates of both analyses. The item parameter estimates corresponding to the multilevel IRT model are not given, since they do not contain relevant information for this example. They merely show the characteristics of the items and are of importance when one is interested in the quality and characteristics of the measurement.

The main result of the standard multilevel analysis, using observed sum scores, was that, conditionally on IQ, adaptive instruction has a small positive effect on students' mathematics achievement. This means that the performance of the students improves when education is adapted to their needs. Further, a high IQ score has a positive effect on the math ability. The Level-1 and -2 predictors together account for a substantial proportion of variation in student achievement: 14% of the student-level and 35% of the school-level variance. The outcomes from the multilevel IRT analysis are slightly different. Both effects of IQ and AI are positive but the effect of IQ on math abilities is

Table 5. Contribution of Adaptive Instruction. Parameter Estimates of the Standard Multilevel and Multilevel IRT Model With Explanatory Variables IQ and AI.

	Multilevel model		Multilevel IRT model	
Parameter	Estimate	S.E.	Estimate	S.E.
Fixed				
Intercept	−.016*	.026	−.021*	.029
Student variables				
IQ	.384	.015	.504	.018
School variables				
AI	.064	.027	.057*	.033
Random				
Within schools	.732	.014	.602	.020
Between schools	.097	.017	.127	.018

Note. *Not significant at the .05 level.

much higher, and the effect of AI is not significantly different from zero. Almost 24% of the proportion of Level-1 variance is explained by the IQ variable. The multilevel IRT analysis reveals that intelligence is a more important predictor for mathematics achievement than was expected from the standard multilevel analysis. In the multilevel IRT analysis, the positive effect of predictor AI is slightly lower than in the standard multilevel analysis but not significant at the .05 level. It follows from both analyses that the Level-2 predictor explains almost the same proportion of variance, 33%. This shows that the measurement error concerning the estimate of the latent variable AI causes that the effect is no longer significant. The underlying characteristic, teachers' capability of adaptive instruction, was measured by 198 teachers using a test of 23 items. The relatively small sample size causes relatively large standard errors of the estimates resulting in a nonsignificant effect of the predictor AI. It must be concluded that based on the multilevel IRT analysis doubts remain as to whether adaptive instruction improves the performances of the students on a math test. This is in contrast with the conclusion from the standard multilevel analysis that ignores the measurement error.

School Effects in the West Bank

A mathematics test consisting of 50 dichotomously scored items (correct/incorrect) was administered to 3,500 grade-7 students in 119 schools located

in the West Bank. Interest was focused on exploring differences within and between schools in the West Bank and establishing factors which explain these differences with respect to students' mathematics abilities. Characteristics of students, teachers, and schools were administered. An intelligence test (IQ) was administered, gender was recorded as zero for male and one for female, and socioeconomic status (SES) was measured by the educational level of the parents. Tests were taken by teachers and school principles to measure aspects such as the school climate (Climate) and leadership (Leader) of the principle. The item responses of IQ and SES were not available. Therefore, normally standardized observed scores were used.

From the teacher's perspective, the school climate was measured by 23 five-point Likert items, and leadership was measured by 25 five-point Likert items. In the sampling design, only one class was selected from each school, so the data consisted of variance at the student- (Level-1) and school-level (Level-2). A stratified sample of schools ensured that all school types and all geographical districts were represented. The average number of students per class is 28, with a minimum of 10 and a maximum of 46 students. A complete description of the data, including the data collection procedure and the different questionnaires, can be found in Shalabi (2002).

In the multilevel IRT analysis, a two-parameter IRT model was used to measure the latent mathematics abilities given the dichotomous responses. The latent variables at Level 2, Climate and Leader, were each measured using a graded response IRT model to handle the polytomous response data from Likert-type items.

The latent variables in the multilevel IRT model were normally standardized. In the standard multilevel analysis, normally standardized observed sum scores were used as measurements for the latent variables Climate and Leader. The reliabilities of the measurements of Climate and Leader were 0.93 and 0.95, respectively.

The variation in the individuals' mathematics abilities and heterogeneity across schools was measured with an empty standard multilevel and an empty multilevel IRT model, that is, only an intercept at Level-1 varying across schools. Table 6 presents the results of the parameter estimates. Again, in the multilevel IRT analysis more variance is explained at the school-level. That is, about 50% of the variance between students' mathematics abilities is explained by school differences, compared to 43% in the standard multilevel analysis.

The model was elaborated to include SES, Gender, and IQ as student characteristics, see Table 7. The three Level-1 predictors all have a significant

Table 6. Effective Schooling in the West Bank. Parameter Estimates of the Standard Multilevel and Multilevel IRT Empty Model.

	Multilevel model		Multilevel IRT model	
Parameter	Estimate	S.E.	Estimate	S.E.
Fixed				
Intercept	.005*	.061	.005*	.066
Random				
Within schools	.569	.014	.515	.014
Between schools	.427	.058	.507	.069

Note. *Not significant at the .05 level.

Table 7. Effective Schooling in the West Bank. Parameter Estimates of the Standard Multilevel and Multilevel IRT Model Including the Level-1 Predictors.

	Multilevel model		Multilevel IRT model	
Parameter	Estimate	S.E.	Estimate	S.E.
Fixed				
Intercept	−.088*	.060	−.097*	.064
Student variables				
SES	.111	.017	.124	.015
Gender	.197	.054	.213	.061
IQ	.344	.020	.351	.015
Random				
Within schools	.470	.011	.408	.012
Between schools	.313	.043	.370	.052

Note. *Not significant at the .05 level.

positive effect on students' mathematics achievement. Female students performed better than male students. The three Level-1 variables together account for a substantial proportion of variation in students' achievement: 21% of the student-level and 27% of the school-level variance in the multilevel IRT model, and 17% and 27% in the standard multilevel analysis. Thus, in the multilevel IRT analysis, the Level-1 predictors discriminate better between students' abilities, which manifests itself in higher effects, and a higher reduction in Level-1 variance.

In the next model, Level-2 explanatory variables were incorporated to explain differences between schools. In the multilevel IRT analysis, IRT

Table 8. Effective Schooling in the West Bank. Parameter Estimates of the Standard Multilevel and Multilevel IRT Model Including the Level-1 and Level-2 Predictors.

	Multilevel model		Multilevel IRT model	
Parameter	Estimate	S.E.	Estimate	S.E.
Fixed				
Intercept	−.088*	.059	−.106*	.062
Student variables				
SES	.112	.017	.125	.024
Gender	.192	.054	.211	.060
IQ	.343	.019	.355	.015
School variables				
Leader	.205	.065	.228	.083
Climate	−.119*	.067	−.130*	.084
Random				
Within schools	.463	.012	.396	.012
Between schools	.287	.040	.326	.049
SES	.010	.009	.039	.007

Note. *Not significant at the .05 level.

models were used to estimate Climate and Leadership given the polytomous outcomes, and observed sum scores were used in the standard multilevel analysis. It was also investigated whether the effect of SES varied from school to school.

In Table 8 we can see that the variable Leader has a positive significant effect on the math abilities, but the effect of variable Climate is negative and nonsignificant. The multilevel IRT model including these Level-2 variables (model 2), can be compared to the model in Table 7 (model 1) using Bayes Factor. The Bayes Factor[1] in favor of model 2 is $\log_{10}(-96771 + 96773) = .301$. Thus, the latent variables at Level-2 did not result in a better model fit. Climate and Leadership proved to have not much of an influence on the students' mathematics abilities but they are rarely investigated in developing countries. The estimated effects of Leader and Climate are higher in the multilevel IRT analysis, and they have higher standard errors. Taking account of the measurement error in the measurements of the Level-2 variables did not

[1]In fact a pseudo-Bayes Factor, which is an approximation of the BF, is computed, since it requires less intensive computations.

result in a different conclusion regarding the significance of the predictors. Both analyses show that the effect of SES varied significantly from school to school. In the multilevel IRT analysis more variance is explained by the differential school effect of SES.

The primary schools in the West Bank differ a lot with regards to the mathematics abilities of the students, but the process factors, measured by Climate and Leader, did not explain much variation at Level-2. The Level-1 characteristics SES, IQ, and gender explained a lot of variation at the student-level. The multilevel IRT model, that is, modeling measurement error in the latent dependent and independent explanatory variables, resulted in larger effects and more explained variance at both levels. The effects were attenuated when the traditional methods (the standard multilevel model) were used which ignored the measurement error by, for example, using observed sum scores as an estimate for the latent variables.

DISCUSSION AND CONCLUSION

Fox and Glas (2001, 2003) introduced a multilevel IRT model for handling hierarchical structured data and handling measurement error in performance indicators using IRT models. IRT models describe the relationship between an underlying characteristic, for example, an examinee's ability, and observed item responses based on characteristics of the items. A sharper distinction between estimates of the underlying abilities can be obtained since they are based on complete response patterns instead of a sum of the observed item scores, or number correct scores, as in the classical test theory model (see e.g., Lord, 1980; Lord & Novick, 1968).

The multilevel IRT model is particularly useful in school effectiveness research, because it takes all sources of variation into account. Since an IRT model measures more variability between examinees' latent characteristics, and parameter estimates are adjusted for measurement error, more realistic and more precise parameter estimates are obtained. In traditional school effectiveness research, educational decisions are made on the basis of observed scores that include measurement errors. The developed multilevel IRT models can be used to assess school effectiveness, after adjusting observed variables for measurement error. The rank ordering of schools will change when taking account of unequal precision of the observed variables. Unreliable observations within a school will cause the school effect to shrink

towards the overall mean. When modeling all sources of uncertainty, differences in the ranking can be taken more seriously. Yet, when accurate measurements are available, differences between the two approaches will disappear. The exact conditions under which it is better to use a multilevel IRT model remains a point of further study.

It was shown in the different applications that with the multilevel IRT model more variance was explained at both levels, and the effects of the predictors were higher. It must be mentioned that these specific datasets were not randomly chosen, they were just available at the level of item responses. In some cases, wrong decisions can be made when ignoring the measurement error, as in example 2. It turned out that the effect of adaptive instruction was not significant when taking the measurement error into account. In most cases, the school effect was underestimated in the standard approach. The multilevel IRT model showed that the school effects were significantly larger than most other studies, based on standard multilevel models, have shown.

The multilevel IRT model is strongly related to the framework of structural equation modeling (SEM), where there is a measurement part and a structural part. The measurement part of a structural equation model consists of observed response variables and latent variables, and the structural part is defined in terms of the latent variables regressed on each other and on some background variables. The main difference between SEM and multilevel IRT modeling is the use of an IRT model as a measurement model.

Multilevel IRT modeling can be used as an alternative to the traditional way of multilevel modeling, by other researchers in the field of educational research. Other research areas may benefit as well from the development of multilevel IRT modeling. For example, a comparison of institutions, like hospitals, is done with hierarchical structured data containing measurement errors and can be analyzed with a multilevel IRT model resulting in measurement error adjusted parameter estimates.

REFERENCES

Aitkin, M., & Longford, N. (1986). Statistical modelling in school effectiveness studies. *Journal of the Royal Statistical Society, Series A*, *149*, 1–43.

Bosker, R.J., Blatchford, P., & Meijnen, G.W. (1999). Enhancing educational excellence, equity and efficiency. In R.J. Bosker, B.P.M. Creemers, & S. Stringfield (Eds.), *Evidence from evaluations of systems and schools in change* (pp. 89–112). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bosker, R.J., & Witziers, B. (1996, April). *The magnitude of school effects: Does it really matter which school a student attends*? Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Doolaard, S. (2002). Stability and change in results of schooling. *British Educational Research Journal*, *28*, 773–787.

Fletcher-Campbell, F., Meijer, C.J.W., & Pijl, Y.J. (1999). Integration policy and practice. In R.J. Bosker, B.P.M. Creemers, & S. Stringfield (Eds.), *Evidence from evaluations of systems and schools in change* (pp. 65–88). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Fox, J.-P. (2001). *Multilevel IRT: A Bayesian perspective on estimating parameters and testing statistical hypotheses.* Ph.D. thesis, Twente University, Enschede, The Netherlands.

Fox, J.-P. (2003). *Multilevel IRT manual.* Computer software and documentation can be retrieved from Twente University, Department of research methodology, measurement and data analysis, website: http://users.edte.utwente.nl/Fox

Fox, J.-P., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 269–286.

Fox, J.-P., & Glas, C.A.W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*, 169–191.

Goldstein, H. (1979). Some models for analysing longitudinal data on educational attainment. *Journal of the Royal Statistical Society, Series A*, *142*, 407–442.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MLwiN.* London: Multilevel Models Project, Institute of Education, University of London.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Nijhoff.

Longford, N.T. (1993). *Random coefficient models.* New York, NY: Oxford University Press.

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

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

Raudenbush, S.W., Bryk, A.S., Cheong, Y.F., & Congdon, R.T., Jr. (2000). *HLM 5. Hierarchical Linear and nonlinear modeling.* Lincolnwood, IL; Scientific Software International.

Scheerens, J., & Bosker, R.J. (1997). *The foundations of educational effectiveness.* Oxford: Pergamon.

Shalabi, F. (2002). *Effective schooling in the West Bank.* Ph.D. thesis, Twente University, Enschede, The Netherlands.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis.* London: Sage.

Van der Linden, W.J., & Hambleton, R.K. (Eds.). (1997). *Handbook of modern item response theory.* New York, NY: Springer.