Journal of Educational Measurement
Winter 2004, Vol. 41, No. 4, pp. 311–329

The Utility of Third International Mathematics and Science Study Scales in Predicting Students' State Examination Performance

Nick Sofroniou and Thomas Kellaghan
Educational Research Centre, Dublin, Ireland

To examine the predictive utility of three scales provided in the released database of the Third International Mathematics and Science Study (TIMSS) (international plausible values, standardized percent correct score, and national Rasch score), information was obtained on the performance in state examinations in mathematics and science in 1996 (2,969 Grade 8 students) and in 1997 (2,898 Grade 7 students) of students in the Republic of Ireland who had participated in TIMSS in 1995. Performance on TIMSS was related to later performance in the state examinations using normal and nonparametric maximum likelihood (NPML) random effects models. In every case, standardized percent correct scores were found to be the best predictors of later performance, followed by national Rasch scores, and lastly, by international plausible values. The estimates for normal mixing distributions are close to those estimated by the NPML approach, lending support to the validity of estimates.

A considerable increase has occurred in the last decade in the number of countries involved in national and international assessments of the achievements of students in their education systems (Kellaghan & Greaney, 2001a). The increase reflects a growth in interest in assessing the outcomes of education (what students actually learn) to complement the long-standing practice of monitoring inputs to the educational process (e.g., physical facilities, curricular materials, teacher training, and student participation rates). It also signals an extension of the focus of assessment from its traditional role in the appraisal of individual students to its use to appraise the performance of systems of education, or clearly defined parts of those systems (Kellaghan, 2003; Kellaghan & Greaney, 2001b).

Although intended to appraise the performance of an education system, national and international assessments are based on the aggregation of individual student performance data. In many assessments, however, the tests that students take differ from those traditionally used to assess individual student performance. To extend the range of curriculum coverage in assessments, without increasing the burden on individual respondents, each student may take only a fraction of a large number of assessment tasks. For example, in the Third International Mathematics and Science Survey (TIMSS), mathematics and science tests each comprised eight booklets, only one of which was responded to by an individual student. Items in the booklets were rotated so that each one was answered by a representative sample of students.

This assessment procedure gives rise to the question: How is individual student performance best represented? TIMSS provides three sets of scale scores that can be used to address this question. First, a common scale (actually a series of scales) was obtained

using multiple imputation to provide reliable indices of student proficiency, known as "plausible values" (described, e.g., by Mislevy, 1991). The values were drawn at random from the estimated ability distribution of students with similar item response patterns and backgrounds. The scaling procedure used a multidimensional random coefficients logit model, in combination with a multivariate linear model imposed on the distribution of the population (Adams, Wu, & Macaskill, 1997, provide further details). Five plausible values were drawn for each student, with variation between results calculated for each separate value reflecting the error due to imputation. The values were used only to compute total scores for mathematics and science performance, described in the TIMSS international reports. Because of the high reliability of the total scores, as indicated by high intercorrelations among the five plausible values, the imputation error was ignored in the reports and only the first plausible value was used (Gonzalez & Smith, 1997). A variety of reliability indices are reported by Adams et al., including the separation reliability, which was obtained by fitting the scaling model omitting any conditioning variables, drawing five plausible values per student, and taking the median of the 10 correlations between plausible value pairs. The separation reliabilities in the international calibration sample of Grade 7 and 8 students (Population 2) were 0.89 for mathematics and 0.80 for science.

Unlike the analyses for third- and fourth-grade students in TIMSS, in which an extensive range of conditioning variables was used to generate the plausible values, the background data for students in seventh and eighth grades (Population 2, which is the focus of the study reported in this article) had not been fully cleaned and checked at the time of scaling, with the result that grade was the only conditioning variable.

As well as plausible values, TIMSS official datasets include mathematics and science scores based on the number of raw score points obtained on a set of items. These were standardized by booklet to generate a score that could be used in comparisons across booklets in preliminary analysis and for test-curriculum matching analyses (Beaton & Gonzalez, 1997). Scores were computed to have a weighted mean score of 50 within each booklet, with a weighted standard deviation of 10. In this article we refer to them as standardized percent correct scores.

The third type of score available for TIMSS is a national Rasch score, which was computed by standardizing the mathematics and science logits to create logit scores with a weighted mean of 150 and a weighted standard deviation of 10 within each country. These scores were intended for preliminary analysis within countries but not for comparisons at the international level since each country was assigned the same mean by the standardization.

Plausible value scores are designed for use in the estimation of population parameters (see Adams et al., 1997) and neither these, the standardized percent correct, nor the national Rasch score can be regarded as optimal estimates of individual student proficiency. If this view is accepted, severe limitations are placed on the type of within-country analyses that can be carried out. However, perhaps because of the considerable expense involved in collecting data, researchers and policymakers have shown an interest in using the data to describe relationships between individual student achievements and a variety of background variables within countries. This had been done for earlier international studies (see Thorndike, 1974) and reports of studies using individual student data from TIMSS are beginning to appear. Some specify that plausible

value scores were used (O'Dwyer, 2002; Wilkins, Zembylas, & Travers, 2002), while others do not say what scores were used (Pelgrum & Plomp, 2002; Shen, 2002; Shen & Pedulla, 2000).

The use of TIMSS data, despite reservations, in analyses based on individual student performance, gives rise to the question: Which of the three estimates provided in the TIMSS database is most appropriate? One way to attempt to answer this question is to relate the indices to performance on another assessment. In many countries, assessment procedures exist in the form of state or public examinations, and so provide an alternative estimate of students' achievements. Students in Ireland, for example, take Junior Certificate Examinations at about the age of 15 years when they have completed 3 years of secondary schooling. The examinations, which are administered by the Department of Education and Science, are based on 3-year programs of study or syllabi prescribed by the Department. Programs, which are followed in all schools, are specified for, and examinations are provided in, among other subjects, mathematics and science.

The availability of student performance data on these examinations for students who had participated in TIMSS, albeit a year or two later, was seen as providing an opportunity to examine the differential value of TIMSS scores in predicting performance on an alternative assessment. The comparison seemed feasible since most students took examinations in mathematics and science (the two curriculum areas assessed in TIMSS). In the first of 2 years (1996) in which Grade 8 students who had been assessed in TIMSS sat the state examinations, 98.1% of candidates took an examination in mathematics, and 86.7% took an examination in science. In 1997, when Grade 7 TIMSS participants sat the state examinations, the percentage taking mathematics was 92.7, and the percentage taking science, 84.3. A study of performance on the Junior Certificate Examinations and performance in another international survey (Programme for International Student Assessment [PISA]) using plausible values had already revealed moderately strong relationships for both mathematics and science ($r = 0.73$) (Shiel, Cosgrove, Sofroniou, & Kelly, 2001).

In the study reported in this article, three sets of TIMSS scale scores for overall performance in mathematics and science (international achievement scores from the first plausible value in each subject area as well as estimates using all five plausible values; standardized percent correct; and national Rasch scores) are examined as covariates in models predicting later student achievement in state examinations. The strength of the relationships should provide evidence of the differential utility of individual student performance measures derived from data that were designed to be used in estimating population parameters.

Method

Sample

The target population of TIMSS (Population 2) consisted of all students enrolled in the two adjacent grades that contained the largest proportion of 13-year-old students in the country. In a two-stage cluster sample design, the first stage consisted of a sample of schools that could be stratified as an option by a country. In Ireland, implicit strata were used for the selection of 150 schools: administrative type (secondary,

vocational, community, comprehensive) and gender, from which a sample was selected using a probability proportional to size technique. The second stage involved the selection of one mathematics classroom of students in each school at each target grade (8 and 9). These also were the students who took the science test so that in many countries, including Ireland, more than one science class was represented.

The number of students followed through to their Junior Certificate Examination was 2,898 in the lower grade sample attending 129 schools and 2,969 in the upper grade sample attending 132 schools. Matching examination records were obtained for 92.7% of the lower-grade students who had participated in TIMSS and for 96.5% of upper-grade students. While the response variables of interest are students' subsequent performance in their state examinations, the cluster sampling also induces dependencies in the observations, since students within a cluster are more like one another than the other members of their population, and this needs to be allowed for in the statistical analysis.

Variables

Data from the TIMSS International Database Middle School Years 1995 CD-ROM were matched with the state (Junior Certificate) examinations for mathematics and science results obtained from the Irish Department of Education and Science. The Junior Certificate program for mathematics has three curriculum streams leading to separate examinations, based on academic difficulty—Foundation, Ordinary, and Higher. The Junior Certificate program for science has two curriculum streams and examinations—Ordinary and Higher.[1] Performance on each examination paper is assigned a grade from A through F, or "No Grade." Attempts have been made to combine the grading system across the academic streams onto a single numerical performance scale with integer scores (e.g., Miller, Farrell, & Kellaghan, 1998). An example for mathematics is reproduced in Table 1. The integer scores for each grade are essentially shifted

TABLE 1
Performance Scale for the Junior Certificate Examination (Miller, Farrell, & Kellaghan, 1998)

Examination Level			
Higher	Ordinary	Foundation	Performance Score
A			12
B			11
C			10
D	A		9
E	B		8
F	C		7
	D	A	6
	E	B	5
	F	C	4
		D	3
		E	2
		F	1

upwards by 3 points for the Ordinary examination and 6 points for the Higher, resulting in a final 12-point scale. Since the Junior Certificate Examination for science has only Ordinary and Higher streams, the scale has 9 points ranging from 4 to 12. The very small number of cases scoring No Grade were omitted from the performance scale and subsequent analysis (6 cases for mathematics and 1 for science).

To examine the degree of overlap of the scores assigned to grades for each stream, an initial sensitivity analysis was carried out by regressing the Junior Certificate performance scales on the TIMSS scales and comparing the fits obtained with those from wider and more narrowly overlapping scoring systems. The results, reproduced in the Appendix, confirm that Miller et al.'s choice of overlap was the one most closely associated with the single assessment scale of TIMSS, and so their scale is used in the analyses reported in this article.

Three sets of scores based on students' performance on TIMSS were derived: five international plausible values (scaled to an international mean of 500), a standardized percent correct score (scaled to a national mean of 50), and a national Rasch score (scaled to a national mean of 150). The lower grade of the Irish sample had a weighted mean of 475.8 (and weighted standard deviation of 81.0) on the first plausible value, while the upper grade had a mean of 550.0 (with a standard deviation of 85.3). Differences in the range of the scales affect the magnitude of the parameter estimates; however, the use of standardized parameter estimates allows scale-independent comparisons to be made.

Analysis

To allow for the dependencies among achievement scores introduced by the clustering of students into classes and schools induced by the sampling design, a random effects statistical modelling framework was used. A random effect in the form of a random intercept added to each model was considered sufficient for the purpose of evaluating the predictive utility of the TIMSS scales. More complex models implementing random effects in the form of random slopes could be used to evaluate variation in the coefficients across the clusters, which also might incorporate additional explanatory variables, such as gender, school, and home background. However, this goes beyond the needs of the present analysis, where the random effect is simply a nuisance variable, fitted to allow for the dependencies among student observations induced by the cluster sampling design.

Since the two grades in the TIMSS sample took different state examinations in consecutive years, they are analyzed separately. As the sampling design was based on a single mathematics class examined at each grade within a given school, the data do not allow the separation of the class and school variance components. Therefore, both the class and school levels are treated as a single school-class level with student scores nested within them (i.e., two variance components are estimated rather than the three that would be available if more than one class had been examined at each grade in a given school). As the sample was drawn from intact mathematics classes, for mathematics scores the Level 2 variance component will incorporate that of schools plus classes, and Level 1 will correspond to variation among students. For science scores, each mathematics class in a given school can correspond to several science classes,

with the result that the Level 1 variance component for science includes additional variation at the class level.

The statistical models fitted an unobserved common random effect z_j for every lower level unit (pupil) in jth upper level unit (school/class). The z_j are assumed to initially follow a Gaussian (normal) distribution, and the model is a generalized linear model with the linear predictor $\eta_{ij} = \beta' x_{ij} + \sigma z_j$, where the i subscripts represent lower level units, the j subscripts upper level units, and the σz_j are cluster-dependent deviations from the intercept. The term σ is the standard deviation of the random intercept that varies across clusters with z_j approximated by a normal distribution $N(0, 1)$. Generalized linear models are presented at an introductory level by Hutcheson and Sofroniou (1999) and Lindsey (1997). They are given extensive coverage by McCullagh and Nelder (1989), and extensions for mixed models involving random effects are presented by Fahrmeir and Tutz (2001), McCulloch and Searle (2001), and Lindsey (1999).

Since the analyses used an identity link with a normal lower level random component, with the addition of a random effect to the linear predictor, the models are instances of hierarchical linear models. The normal distribution fitted as the lower level random component has a function relating the mean to the variance, which equals a constant (i.e., the lower level variance is assumed to remain constant as the fitted mean changes). In the present case, the models for normal random effects can be written as

$$\mathrm{PS}_{ij} = \alpha + \beta_{10}\mathrm{TIMSS}_{ij} + \sigma z_j, \tag{1}$$

where PS is the performance score on the Junior Certificate Examination, TIMSS represents one of the TIMSS achievement scales, α is the intercept, and σz_j are school/class deviations from the intercept. The model notation used in this article follows that of Aitkin and Longford (1986) and Longford (1993), incorporating expanded terms that distinguish between normal and nonparametric maximum likelihood (NPML) random effects as given in Aitkin (1999). We feel that this notation is accessible for readers who are likely to be familiar with the similar notation of ordinary least-squares regression. The residual term is implicit in the models (for example, written as $+\epsilon_{ij}$ in classical linear models to denote the "error" of the fitted model) since in generalized linear models the linear predictor given by the systematic component of the model is specified separately from the random component, with a link function connecting the two. Only in the case of a normal random component with identity link does the classical linear model with residual term combine straightforwardly into a single equation. Snijders and Bosker (1999, chapter 14) give parallel forms for the residual, together with the corresponding linear predictor incorporating an added random effect, for several examples of generalized linear mixed models.

The Hierarchical Linear Modeling (HLM) 5 software package (Raudenbush, Bryk, Cheong, & Congdon, 2000) was used in its Windows mode to fit the normal random effects models using the full maximum likelihood estimation option from the Basic Specifications menu option of the HLM2 routine. An alternative random effect was fitted using a NPML approach. The algorithm is implemented in the form of a macro by Aitkin and Francis (in press) for the Generalized Linear Interactive Modelling (GLIM) 4 statistical software package (Francis, Green, & Payne, 1993), which, rather than using the closed form expression available for the special case of both levels

being modeled as normally distributed, employs Gaussian quadrature to replace the integral over the normal z_j with a finite sum of K Gaussian quadrature mass-points z_k with masses π_k. In the case of a normal mixing distribution, the mass-points and masses are known independently of the dataset. This method of estimation has the advantage of allowing the broad family of generalized linear models to be extended to a multilevel framework, though it requires sufficient quadrature points to be chosen, and is consuming of computer processing time. A further advantage is that it becomes possible to estimate the mixing distribution itself, as well as the model parameters, producing a NPML random effects model—this being a discrete distribution on a finite number of mass-points. Thus, both the mass-points of the mixing distribution and their masses (i.e., probabilities) are estimated as additional parameters in the model. Aitkin (1999) gives the statistical theory behind the method, and Aitkin and Francis (in press) present a detailed explanation of their implementation in GLIM along with several worked examples together with the required GLIM macros.

Early research on NPML estimation includes the work of Laird (1978) and Lindsay (1983) addressing theoretical issues. Heckman and Singer (1984) provided applied examples of how parameter estimates can be very sensitive to the nature of the mixing distribution specified for the random effects, and Wood and Hinde (1987) showed how mass-point locations and their masses can be estimated for variance components models using maximum likelihood. NPML random effects models provide robustness and an invariance to model assumptions concerning the unobserved random effects with some loss of efficiency relative to a correct model assumption. They get around the lack of information in the data concerning the mixing distribution, which is only available from the marginal distribution of the data. In the NPML method the mixing distribution is treated as a nuisance parameter or function, avoiding potentially misleading inferences from an unverifiable model assumption concerning the distribution of the random effect that cannot be directly addressed (Aitkin, 1999). A drawback of the NPML method is the lack of an estimate of the precision of the distribution estimates (Mallet, Metré, Steimer, & Lokiec, 1988).

A generalized linear model with a NPML random effect can be written in its general form with the linear predictor $\eta_{ijk} = \beta' x_{ij} + \alpha_k$, where α_k is a parameter for the intercept of the kth component of the discrete mixing distribution. In the analyses presented here, the models involving NPML random effects (having identity link functions) can be written as

$$PS_{ijk} = \beta_{10} TIMSS_{ij} + \alpha_k. \quad (2)$$

The notation for the random effect of the normal random intercept model, given as a separate mean intercept α and a standard deviation σ multiplied by the standard normal distribution $N(0,1)$ of z_j, makes it explicit that α and σ are parameters estimated by the model, while in the NPML random effects model it is the location and weights of the k mass points of the discrete probability distribution for α_k that are estimated. In the NPML case, one can subsequently summarize the properties of the mixing distribution such as its mean and standard deviation from the parameters of the probability mass function allowing comparison of these quantities with those of the normal random effects model.

A decision had to be made regarding the appropriate number of mass-points. In the present study, this was done by incrementing the number of mass-points until the deviance showed little change (a χ^2 statistic, calculated from the change in deviance, is suitable for this purpose). This approach has received theoretical support from Murphy and van der Vaart (2000), along with the simulation studies of Davies (mentioned in Aitkin, 1999), providing justification for the standard asymptotic null χ^2 for comparing nested models.

As the expectation maximization (EM) algorithm used in the GLIM 4 macro does not directly provide standard errors for parameter estimates, the method of Dietz and Böhning (1995) was used. This involved dropping each explanatory variable in turn and calculating the standard error as

$$SE = |\text{pe}|/\sqrt{\text{deviance change}}, \tag{3}$$

where $|\text{pe}|$ is the absolute value of the parameter estimate. This procedure provides a good approximation to the standard error in large samples. Aitkin (1999) argues that, even in the case of small samples with skewed log likelihood functions, this is a more appropriate measure of significance than a standard Wald test based on the inverse information matrix, since Dietz and Böhning's method gives a squared t statistic that is equal to the likelihood ratio statistic.

For the NPML random effects, this method of obtaining a t statistic was combined in the analyses described in this article with that for computing a t statistic from imputed data sets given by Little and Schenker (1995), which applies to plausible values. The parameter estimates and standard errors for each plausible value set are combined as follows. The combined parameter estimate is given by averaging across the value of pe for the K sets of plausible values.

$$\overline{\text{pe}} = \frac{1}{K}\sum_{k=1}^{K} \text{pe}_k. \tag{4}$$

Then the average of the completed-data (i.e., within-imputation) variances is calculated from each standard error, denoted by se

$$\overline{U} = \frac{1}{K}\sum_{k=1}^{K} \text{se}_k^2 \tag{5}$$

and the between-imputation variance is given by

$$B = \frac{1}{(K-1)}\sum_{k=1}^{K} (\text{pe} - \overline{\text{pe}})^2. \tag{6}$$

These are then used to give the combined standard error as the square root of the total variance T where

$$T = \overline{U} + \left(1 + \frac{1}{K}\right)B. \tag{7}$$

The t statistic is then straightforwardly calculated as $\overline{\text{pe}}/\sqrt{T}$. Finally, the associated degrees of freedom are given by first calculating the variance ratio

$$vr = \left(1 + \frac{1}{K}\right) B/\overline{U} , \tag{8}$$

and then

$$df = (K - 1)\left(1 + \frac{1}{\text{vr}}\right)^2 . \tag{9}$$

The first plausible value for each academic area (termed the "International Mathematics Score" and "International Science Score" in the TIMSS international reports that used them) was also analyzed separately.

For analysis purposes, an ordinal logistic regression method such as a proportional odds or continuation ratio model was considered, but this was felt to involve a rather large number of categories with associated parameters (9 and 12 points form the performance scales). Therefore, a hierarchical model with Normal errors and identity link was chosen. This has the advantage of allowing a direct interpretation of parameters as the expected change in mean performance score per unit of a given explanatory variable (one unit is equal to one examination grade). This is especially true for the percent correct explanatory variable, which uses a familiar scale. A comparison of the estimates obtained from the models incorporating normal random effects with those using NPML estimates of a discrete distribution of the random effects allows an examination of the sensitivity of the conclusions to the specified mixing distribution. Throughout this article the models use an identity link between the response variable and the linear predictor together with a normally distributed Level 1 random component, while the Level 2 random component added to the linear predictor is either normally distributed or a discrete distribution estimated by the NPML method.

Results

The parameter estimates, standard errors, and standardized parameter estimates are presented in Tables 2 through 5, as well as t ratios, and R^2 statistics for Levels 1 and 2. The explained variances at both levels, R_1^2 and R_2^2, are calculated using the method of Snijders and Bosker (1999) with a representative group size based on the total number of eligible students divided by the number of schools, averaged across each grade (i.e., 26 students). The Level 1 R^2 is calculated as

$$R_1^2 = 1 - \frac{\sigma_1^2 current + \sigma_2^2 current}{\sigma_1^2 empty + \sigma_2^2 empty}, \tag{10}$$

where *current* refers to the model of interest, *empty* to the random intercept only model, and σ_2^2 and σ_1^2 are the residual variances for Level 1 and Level 2, respectively.

The Level 2 R^2 is given by

$$R_2^2 = 1 - \frac{\sigma_1^2 current/c + \sigma_2^2 current}{\sigma_1^2 empty/c + \sigma_2^2 empty}, \tag{11}$$

where c is the typical size of each cluster, termed the representative group size by Snijders and Bosker. Note that no Level 2 explanatory variables were fitted in the analyses reported here.

In the case of the NPML random effects models, the Level 2 variance was calculated from the masses and probabilities of the discrete mixing distribution using the general formulae for the mean and variance of a discrete probability distribution (e.g., as given in Agresti & Finlay, 1997). The mean of the discrete mixing distribution is first calculated as

$$\mu = \sum_{k=1}^{K} \alpha_k \hat{\pi}_k, \tag{12}$$

where α_k is the estimate of the intercept at mass-point k, and $\hat{\pi}_k$ is the estimated probability (i.e., mass) at that point. Then the variance is given by

$$\sigma^2 = \sum_{k=1}^{K} (\alpha_k - \mu)^2 \hat{\pi}_k. \tag{13}$$

For the mixture distributions used to estimate the NPML random effects, five mass-points were required in the case of the lower grade mathematics sample, and six mass-points were required for the remaining mathematics grade sample and the two science grade samples. The parameter estimates can be interpreted as the improvement in the Junior Certificate Examination grade associated with a unit increment in each of the TIMSS scales. To facilitate interpretation, standardized parameter estimates are provided, which give the number of standard deviations of the response variable predicted by a one standard deviation change in the TIMSS scales. These are the multilevel analogs of the Pearson correlation to ordinary least-squares regression estimates. Adopting a modeling approach offers a better scale-independent measure than simply calculating correlation coefficients since it allows evaluation of the model fit by diagnostic methods to determine distortions due to such factors as extreme observations, nonlinearity, and non-normality. The focus of the analyses is on the size of the associations between the performance on the state examination in each curriculum area and TIMSS regressors, along with their test statistics and measures of explained variance.

The raw parameter estimates and standard errors are much larger for the percent correct and national Rasch scores than for the international plausible values, reflecting the differences in dimensions of the scales used in TIMSS. A clearer picture emerges when one examines the standardized parameter estimates. In Tables 2 through 5, it can be seen that the standardized estimates are consistently larger (by a small margin) for the percent correct scale than for the national Rasch scores and larger (by a bigger margin) than for the international plausible values. The order of percent correct > Rasch score > plausible values is maintained throughout the parameter estimates. The size of the t ratio and the explained variances at each level also show this pattern, reflect-

TABLE 2
Estimates for the TIMSS Scales Fitted to the Performance Scale for the Junior Certificate Mathematics Examination Results, Grade 7

Model	Parameter Estimate	Standard Error	Standardized Estimate	t Ratio	df	p Value	R_1^2	R_2^2
Normal random effects models								
Percent correct	0.1555	0.00295	0.68	52.8	2896	< .001	0.58	0.74
Rasch score	0.1588	0.00302	0.67	52.5	2896	< .001	0.57	0.73
First PV	0.0153	0.00034	0.60	44.9	2896	< .001	0.51	0.67
Combined PVs	0.0154	0.00049	0.60	31.8	16	< .001	0.50	0.66
NPML random effects models								
Percent correct	0.1556	0.00355	0.68	43.9	2892	< .001	0.58	0.73
Rasch score	0.1590	0.00363	0.67	43.8	2892	< .001	0.57	0.72
First PV	0.0154	0.00040	0.61	38.9	2892	< .001	0.50	0.66
Combined PVs	0.0155	0.00051	0.61	30.6	29	< .001	0.50	0.65

Note. The First PV (plausible value) is the same as the International Mathematics Score.

ing a better-fitting model for the percent correct explanatory variable. The estimates from the analysis of the first plausible value are very close to those obtained when all five plausible values are combined across analyses. However, the t ratio statistics and associated degrees of freedom are smaller for the combined plausible values, reflecting their incorporation of the between-imputation variance, which is ignored when a single plausible value set is examined in isolation.

A comparison of the estimates from the mathematics tests with those for the science tests reveals that the former are substantially larger. This may be due to the fact that

TABLE 3
Estimates for the TIMSS Scales Fitted to the Performance Scale for the Junior Certificate Mathematics Examination Results, Grade 8

Model	Parameter Estimate	Standard Error	Standardized Estimate	t Ratio	df	p Value	R_1^2	R_2^2
Normal random effects models								
Percent correct	0.1168	0.00285	0.54	40.9	2967	< .001	0.56	0.69
Rasch score	0.1067	0.00277	0.51	38.5	2967	< .001	0.52	0.64
First PV	0.0104	0.00032	0.44	33.1	2967	< .001	0.45	0.57
Combined PVs	0.0104	0.00033	0.44	31.4	285	< .001	0.45	0.56
NPML random effects models								
Percent correct	0.1163	0.00321	0.54	36.2	2962	< .001	0.56	0.68
Rasch score	0.1104	0.00321	0.52	34.4	2962	< .001	0.53	0.66
First PV	0.0104	0.00034	0.44	30.7	2962	< .001	0.45	0.56
Combined PVs	0.0105	0.00035	0.44	30.0	970	< .001	0.45	0.56

Note. The First PV (plausible value) is the same as the International Mathematics Score.

TABLE 4
Estimates for the TIMSS Scales Fitted to the Performance Scale for the Junior Certificate Science Examination Results, Grade 7

Model	Parameter Estimate	Standard Error	Standardized Estimate	*t* Ratio	*df*	*p* Value	R_1^2	R_2^2
Normal random effects models								
Percent correct	0.0930	0.00295	0.47	31.5	2635	< .001	0.35	0.48
Rasch score	0.0950	0.00309	0.46	30.8	2635	< .001	0.34	0.47
First PV	0.0077	0.00031	0.38	24.6	2635	< .001	0.25	0.36
Combined PVs	0.0078	0.00035	0.38	22.3	106	< .001	0.26	0.36
NPML random effects models								
Percent correct	0.0927	0.00323	0.47	28.7	2630	< .001	0.35	0.48
Rasch score	0.0950	0.00337	0.46	28.2	2630	< .001	0.34	0.47
First PV	0.0077	0.00033	0.38	23.2	2630	< .001	0.26	0.36
Combined PVs	0.0079	0.00037	0.39	21.4	165	< .001	0.26	0.37

Note. The First PV (plausible value) is the same as the International Science Score.

the smaller number of academic streams in the science examination, with the associated narrower performance scale, served to constrain variation in the response variable relative to that found in mathematics. Grade differences are observable in the size of the parameter estimates for both domains. The estimates in mathematics are noticeably smaller for the upper grade than for the lower grade, with similar standard errors. The same pattern is found in the case of science though the grade difference is smaller.

TABLE 5
Estimates for the TIMSS Scales Fitted to the Performance Scale for the Junior Certificate Science Examination Results, Grade 8

Model	Parameter Estimate	Standard Error	Standardized Estimate	*t* Ratio	*df*	*p* Value	R_1^2	R_2^2
Normal random effects models								
Percent correct	0.0831	0.00269	0.46	30.9	2643	< .001	0.41	0.56
Rasch score	0.0764	0.00253	0.44	30.3	2643	< .001	0.39	0.52
First PV	0.0071	0.00028	0.38	25.7	2643	< .001	0.32	0.43
Combined PVs	0.0067	0.00038	0.36	17.7	16	< .001	0.30	0.42
NPML random effects models								
Percent correct	0.0818	0.00290	0.45	28.2	2638	< .001	0.41	0.55
Rasch score	0.0751	0.00271	0.43	27.7	2638	< .001	0.39	0.51
First PV	0.0069	0.00029	0.37	24.0	2638	< .001	0.31	0.42
Combined PVs	0.0067	0.00039	0.36	17.4	20	< .001	0.30	0.41

Note. The First PV (plausible value) is the same as the International Science Score.

The variance of the intercept for the mathematics models was estimated to be between 0.47 to 0.61 for the lower grade and between 0.95 to 1.37 for the upper grade. The intercept variance for the science models ranges between 0.71 and 0.88 for the lower grade and between 0.68 and 0.92 for the upper grade. In each range, the lower value corresponds to that for the percent correct covariate, and the upper value corresponds to the combined plausible value. The variance estimates of the normal random effects models correspond well to those of the models with NPML random effects.

In the present article uncentered explanatory variables were used, which gave identical parameter estimates for the TIMSS covariates to grand-mean centering. Some authors have advocated group-mean centering to examine the within-cluster parameter estimate of an explanatory variable (e.g., Raudenbush & Bryk, 2002). However, Snijders and Bosker (1999) suggest grand-mean centered estimates be calculated (especially if one also wishes to examine models with random slopes) unless one is interested in the relative position of an individual in his or her group, in which case group-mean centered estimates may be of interest. For comparison purposes the normal random effects models were refitted with group-mean centered TIMSS covariates. While the estimates are slightly lower, the conclusions to be drawn from the analyses remain unchanged. For mathematics, Grade 7, the group-mean centered parameter estimates are percent correct = 0.1504, national Rasch score = 0.1536, first plausible value = 0.0147, and combined plausible values = 0.0148. For Grade 8, the corresponding values are percent correct = 0.1096, national Rasch score = 0.1005, first plausible value = 0.0098, and combined plausible values = 0.0098. The group-mean centered parameter estimates in the case of science, Grade 7, are percent correct = 0.0890, national Rasch score = 0.0909, first plausible value = 0.0073, and combined plausible values = 0.0075. For Grade 8, the values are percent correct = 0.0779, national Rasch score = 0.0720, first plausible value = 0.0067, and combined plausible values = 0.0064.

Conclusion

Of the three overall assessment scales provided with the TIMSS 1995 International Database, the scale most strongly associated with students' subsequent performance in state examinations was consistently found to be the standardized percent correct scale. This was quite closely followed by national Rasch scores, and, by a wider margin, international plausible values. The range of predictive utility varies from the best result of a 0.68 standard deviation improvement in mathematics performance on the state examination per standard deviation increase in standardized percent correct, down to the worst of a 0.36 standard deviation improvement for science per standard deviation increase in international plausible value. Thus, the predictive utility of the percent correct scale in relation to Junior Certificate Examination performance is greatest, and it forms a good candidate variable for further modelling of these students' performance with additional individual and cluster-level explanatory variables. A similar pattern is visible in the correlations between the TIMSS scale and Junior Certificate Examination performance in Tables A1 and A2.

Once the TIMSS regressor with the best predictive utility has been determined, the use of the raw parameter estimate allows one to make statements in terms of the

change in the state examination scale that is associated with a unit increment in the explanatory variable. This has advantages for interpretation and the dissemination of findings since an increase of one grade in an examination is readily understood by professionals involved in education, as well as by parents. A normal random effects model with the standardized percent correct scale fitted has parameter estimates for the TIMSS mathematics scale of 0.156 for the lower grade and 0.117 for the upper grade. Similarly, the corresponding model for the TIMSS science scale has parameter estimates of 0.093 for the lower grade and 0.083 for the upper grade. Multiplying these estimates by 10 suggests that an increase of 10 percentage points in a student's TIMSS score (the weighted standard deviation) is associated with an average increase in the Junior Certificate mathematics examination of 1.56 examination grades for the lower-grade population and of 1.17 examination grades for the upper-grade population. Likewise in the science examination, an increment of 10 percentage points in TIMSS is associated with an increase of 0.93 of an examination grade for lower-grade students and 0.83 of an examination grade for upper-grade students. The estimates from the NPML random effects model are very similar.

The smaller size of the science estimates compared to the mathematics estimates may be to some extent a function of the narrow performance scale, which may result in a constraint on the degree of variation in performance scores. The finding that estimates in the case of both curriculum domains are smaller for the higher grade is counter to what might be expected on the basis of a hypothesis that assessments that are closer in time result in more closely related scores. However, it should be noted that the state examinations taken by candidates differed for the two grades.

The parameter estimates for a normal mixing distribution are close to those from nonparametric distributions estimated by NPML. This adds support to the accuracy of the fixed effects and, together with the similar variation in the random intercepts in both cases, suggests that the assumption of a normal distribution for the unobserved common random effect in each model is a reasonable one in the present case. However, the use of empirical Bayes shrinkage estimates to calculate posterior means for the school-class clusters may differ since shrinkage is toward the common mean of the random effects in the parametric case and toward discrete mass-points in the NPML case (Aitkin, 1996).

The variances explained by the percent correct and national Rasch scores in mathematics might be considered to be quite high at Level 1 and high at Level 2, suggesting a considerable overlap in the performance of students on TIMSS and the Junior Certificate Examination. The corresponding values for science are more moderate and closer in magnitude for the two levels. A difference between mathematics and science in the relationship of the international assessment scores to the Junior Certificate Examination was not observed for Irish students in the PISA study (Shiel et al., 2001).

Since our study revealed relatively strong relationships between individual student scores, derived from a study designed to estimate population parameters and subsequent performance on local examinations, its findings may be interpreted as providing support for within-country analyses at the individual student level. The strength of the relationships are substantial given the time interval between assessments and the fact that the tests used in international assessments are unlikely to fully represent

the curriculum of an individual participating country on which the local state examinations in our study were based (see Bottani & Tuijnman, 1994; Kellaghan, 1996; Kellaghan & Grisay, 1995).

The study findings also indicate, however, that the equivalence of measures that vary in the procedure they use to estimate student achievement in international studies cannot be assumed. Perhaps it is not surprising that plausible values do not fare very well given that they are designed to estimate population parameters. However, neither might one have expected the simple percent correct on an assessment of one-eighth of tasks to predict later achievement better than scores derived from item response models that took account of the sampling of subjects and the latent nature of variables. It may be that the utility of plausible values would be improved if more than one conditioning variable (grade) had been used in generating them; however, these are not available for TIMSS data at seventh and eighth grades. Mislevy (1991), in an examination of conditioning variables used in the National Assessment of Educational Progress reading survey in years 1984 and 1988, found biases of around −30% and −10%, respectively, for the estimated effects of excluding a range of background variables, including gender, race, geographical variables, and grade. These biases were reduced to within −3% by conditioning on the first 32 principle components of the 64 contrasts available. His findings highlight the potential importance of conditioning on background variables when publicly releasing datasets, such as TIMSS, to decrease biases in secondary analyses. One relatively straightforward method would appear to be to condition on a sufficient number of principle components of the background variables.

While the findings of our study provide prima facie evidence that analyses of individual student performance based on studies designed to provide population parameters may be justified, they cannot, of course, be automatically generalized. For example, differences in the country profiles of students' responses (Zabulionis, 2001), which could reflect international differences in the structure of students' achievements, may mean that international assessment instruments are less appropriate (at both national and individual levels) in some countries than in others, thus limiting the value of data derived from them for within-country analyses. Further studies in a variety of countries would be required to throw light on these issues. An implication of our findings is that such studies should be sensitive to the fact that the three measures released with the TIMSS database may produce findings that differ and that the plausible values used in research to date may not provide the best estimates of individual student proficiency.

Appendix

Tables A1 and A2 present Pearson correlations between the TIMSS scales and performance scales for the Junior Certificate Examination in mathematics and science. These correlations equal the standardized parameter estimates from ordinary least-squares regressions with single explanatory variables, which were used for initial explorations of the properties of the performance scales. For this purpose, only the first of the international plausible values was used. The 12-point mathematics performance scale and the 9-point science performance scale correspond to the final ones chosen for the analyses reported in this article.

TABLE A1

Correlations Between the TIMSS Scales and a Variety of Performance Scales for the Junior Certificate Mathematics Examination

	Number of Scale Points						
TIMSS Scale	10	11_{s1}	11_{s2}	12	13_{s1}	13_{s2}	14
Grade 7							
Percent correct	0.74	0.74	0.75	0.76	0.75	0.75	0.75
Rasch score	0.74	0.74	0.75	0.75	0.75	0.75	0.75
First PV	0.69	0.69	0.70	0.71	0.70	0.71	0.70
Grade 8							
Percent correct	0.77	0.76	0.79	0.79	0.78	0.79	0.78
Rasch score	0.75	0.74	0.77	0.77	0.76	0.77	0.77
First PV	0.71	0.70	0.73	0.73	0.72	0.74	0.73

Note. *s*1 and *s*2 indicate the two possible separations of the Higher and Ordinary streams available for the given scales.

TABLE A2

Correlations Between the TIMSS Scales and a Variety of Performance Scales for the Junior Certificate Science Examination

	Number of Scale Points		
TIMSS Scale	8	9	10
Grade 7			
Percent correct	0.58	0.59	0.58
Rasch score	0.57	0.58	0.57
First PV	0.50	0.50	0.50
Grade 8			
Percent correct	0.65	0.67	0.66
Rasch score	0.64	0.66	0.65
First PV	0.58	0.60	0.60

Notes

The authors would like to thank Professor Murray Aitkin for providing the GLIM macros for fitting NPML random effects models in advance of publication and for advice on their use.

[1]A little over one-seventh of students taking mathematics took the Foundation examination, one-half took the Ordinary, and just over one-third took the Higher. Of students taking science, just under one-third took the Ordinary examination, and a little over two-thirds took the Higher (Ireland, Department of Education and Science, 1997, 1998).

References

Adams, R. J., Wu, M. L., & Macaskill, G. (1997). Scaling methodology and procedures for the mathematics and science scales. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study technical report, Volume II: Implementation and analysis—Primary and middle school years* (pp. 111–115). Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences* (3rd ed.). London: Prentice-Hall.

Aitkin, M. (1996, July). *Empirical Bayes shrinkage using posterior random effects means from nonparametric maximum likelihood estimation in general random effects models.* Paper presented at the 11th International Workshop on Statistical Modelling, Orvieto, Italy.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics, 55,* 117–128.

Aitkin, M., & Francis, B. J. (in press). Fitting generalized linear models by nonparametric maximum likelihood. *GLIM Newsletter.*

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

Beaton, A. E., & Gonzalez, E. J. (1997). Reporting achievement in mathematics and science content areas. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study technical report, Volume II: Implementation and analysis—Primary and middle school years* (pp. 175–185). Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Bottani, N., & Tuijnman, A. C. (1994). The design of indicator systems. In A. C. Tuijnman & T. N. Postlethwaite (Eds.), *Monitoring the standards of education: Papers in honour of John P. Keeves* (pp. 47–77). Oxford, UK: Pergamon.

Dietz, E., & Böhning, D. (1995). Statistical inference based on a general model of unobserved heterogeneity. In G. U. H. Seeber, B. J. Francis, R. Hatzinger, & G. Steckel-Berger (Eds.), *Statistical modelling* (pp. 75–82). New York: Springer.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modelling based on generalized linear models* (2nd ed.). New York: Springer.

Francis, B., Green, M., & Payne, C. (Eds.). (1993). *The GLIM system: Generalized linear interactive modelling.* Oxford, UK: Oxford University Press.

Gonzalez, E. J., & Smith, T. A. (Eds.). (1997). *User guide for the TIMSS international database—Primary and middle school years.* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Heckman, J. J., & Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models of duration. *Econometrica, 52,* 271–320.

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models.* London: Sage.

Ireland. Department of Education and Science. (1997). *Statistical report 1995/96.* Dublin: Stationery Office.

Ireland. Department of Education and Science. (1998). *Statistical report 1996/97.* Dublin: Stationery Office.

Kellaghan, T. (1996). IEA studies and educational policy. *Assessment in Education, 3,* 143–160.

Kellaghan, T. (2003). Local, national, and international levels of system evaluation: Introduction. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 873–882). Boston: Kluwer Academic.

Kellaghan, T., & Greaney, V. (2001a). The globalisation of assessment in the 20th century. *Assessment in Education, 8,* 87–102.

Kellaghan, T., & Greaney, V. (2001b). *Using assessment to improve the quality of education.* Paris: International Institute for Educational Planning.

Kellaghan, T., & Grisay, A. (1995). International comparisons of student achievement: Problems and prospects. In *Measuring what students learn. Measurer les résultats scolaires* (pp. 41–61). Paris: Organisation for Economic Co-operation and Development.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association, 73,* 805–811.

Lindsay, B. G. (1983). The geometry of mixture likelihoods, part 1: A general theory. *Annals of Statistics, 11,* 86–94.

Lindsey, J. K. (1997). *Applying generalized linear models.* New York: Springer.

Lindsey, J. K. (1999). *Models for repeated measurements* (2nd ed.). Oxford, UK: Oxford University Press.

Little, R., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39–69). New York: Plenum Press.

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

Mallet, A., Metré, F., Steimer, J. K., & Lokiec, F. (1988). Nonparametric maximum likelihood estimation for population pharmacokinetics, with application to cyclosporine. *Journal of Pharmacokinetics and Biopharmaceutics, 16,* 311–327.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear and mixed models.* New York: Wiley.

Miller, D., Farrell, E., & Kellaghan, T. (1998). *From Junior to Leaving Certificate: A longitudinal study of 1994 Junior Certificate candidates who took the Leaving Certificate examination in 1996.* Dublin: National Council for Curriculum and Assessment.

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

Murphy, S. A., & van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association, 95,* 449–485.

O'Dwyer, L. M. (2002). Extending the application of multilevel modeling to data from TIMSS. In D. F. Robitaille & A. E. Beaton (Eds.), *Secondary analysis of the TIMSS data* (pp. 359–373). Boston: Kluwer Academic.

Pelgrum, W. J., & Plomp, T. (2002). Indicators of ICT in mathematics: Status and covariation with achievement measures. In D. F. Robitaille & A. E. Beaton (Eds.), *Secondary analysis of the TIMSS data* (pp. 317–330). Boston: Kluwer Academic.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2000). *HLM5: Hierarchical linear and nonlinear modelling.* Lincolnwood, IL: Scientific Software International.

Shen, C. (2002). Revisiting the relationship between students' achievements and their self-perceptions: A cross-national analysis based on TIMSS data. *Assessment in Education, 9,* 161–181.

Shen, C., & Pedulla, J. J. (2000). The relationship between students' achievement and their self-perception of competence and rigour of mathematics and science: A cross-national analysis. *Assessment in Education, 7,* 237–253.

Shiel, G., Cosgrove, J., Sofroniou, N., & Kelly, A. (2001). *Ready for life? The literacy achievements of Irish 15-year olds with comparative international data.* Dublin: Educational Research Centre.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Thorndike, R. L. (1974). The relation of school achievement to differences in the backgrounds of children. In A. C. Purves & D. U. Levine (Eds.), *Educational policy and international assessment. Implications of the IEA surveys of achievement* (pp. 93–103). Berkeley, CA: McCutchan.

Wilkins, J. L. M., Zembylas, M., & Travers, K. J. (2002). Investigating correlates of mathematics and science literacy in the final year of secondary school. In D. F. Robitaille & A. E. Beaton (Eds.), *Secondary analysis of the TIMSS data* (pp. 291–316). Boston: Kluwer Academic.

Wood, A., & Hinde, J. (1987). Binomial variance component models with a non-parametric assumption concerning random effects. In R. Crouchley (Ed.), *Longitudinal data analysis* (pp. 110–128). Aldershot: Avebury.

Zabulionis, A. (2001, September 14). Similarity of mathematics and science achievement of various nations. *Education Policy Analysis Archives, 9*(33). Retrieved August 21, 2002, from http://epaa.asu.edu/epaa/v9n33/

Authors

NICK SOFRONIOU is a Research Fellow at the Educational Research Centre, St Patrick's College, Dublin 9, Ireland; nick.sofroniou@erc.ie. His research interests include the applications of multilevel and generalized linear models in educational research.

THOMAS KELLAGHAN is Director of the Educational Research Centre, St Patrick's College, Dublin 9, Ireland. His research interests include assessment and program evaluation.