Chapter 11

Scaling the PIRLS 2006 Reading Assessment Data

Pierre Foy, Joseph Galia, and Isaac Li

11.1 Overview

PIRLS 2006 had ambitious goals for broad coverage of the reading purposes and processes as described in its assessment framework[1] and for measuring trends across assessment cycles. To achieve these goals, the PIRLS 2006 assessment consisted of 10 reading passages and items arranged into 40-minute assessment blocks, four of which were retained from the 2001 assessment in order to serve as the foundation for measuring trends. PIRLS used a matrix-sampling design[2] to assign assessment blocks to student booklets—two blocks per student booklet—so that a comprehensive picture of the reading achievement of fourth-grade students in participating countries could be assembled from the booklets completed by individual students. PIRLS relied on Item Response Theory (IRT) scaling to combine the student responses and provide accurate estimates of reading achievement in the student population of each participating country, as well as measure trends in reading achievement among countries that also participated in the 2001 assessment. The PIRLS scaling methodology also uses multiple imputation—or "plausible values"—methodology to obtain proficiency scores in reading for all students, even though each student responded to only a part of the assessment item pool.

This chapter first reviews the psychometric models and the conditioning and plausible values methodology used in scaling the PIRLS 2006 data, and then

[1] The PIRLS 2006 assessment framework is described in Mullis, Kennedy, Martin, & Sainsbury (2006).

[2] The PIRLS 2006 achievement test design is described in Chapter 2.

describes how this approach was applied to the PIRLS 2006 data and to the data from the previous PIRLS 2001 study in order to measure trends in achievement. The PIRLS scaling was carried out at the TIMSS & PIRLS International Study Center at Boston College, using software from Educational Testing Service.[3]

11.2 PIRLS 2006 Scaling Methodology[4]

The IRT scaling approach used by PIRLS was developed originally by Educational Testing Service for use in the U.S. National Assessment of Educational Progress. It is based on psychometric models that were first used in the field of educational measurement in the 1950s and have become popular since the 1970s for use in large-scale surveys, test construction, and computer adaptive testing.[5] This approach also has been used to scale IEA's TIMSS data to measure trends in mathematics and science.

Three distinct IRT models, depending on item type and scoring procedure, were used in the analysis of the PIRLS 2006 assessment data. Each is a "latent variable" model that describes the probability that a student will respond in a specific way to an item in terms of the student's proficiency, which is an unobserved—or "latent"—trait, and various characteristics (or "parameters") of the item. A three-parameter model was used with multiple-choice items, which were scored as correct or incorrect, and a two-parameter model for constructed-response items with just two response options, which also were scored as correct or incorrect. Since each of these item types has just two response categories, they are known as dichotomous items. A partial credit model was used with polytomous constructed-response items, i.e., those with more than two response options.

11.2.1 Two- and Three-Parameter IRT Models for Dichotomous Items

The fundamental equation of the three-parameter (3PL) model gives the probability that a student whose proficiency on a scale k is characterized by the unobservable variable θ_k will respond correctly to item i as:

[3] PIRLS is indebted to Matthias von Davier, Ed Kulick, and John Barone of Educational Testing Service for their advice and support.

[4] This section describing the PIRLS scaling methodology has been adapted with permission from the TIMSS 1999 Technical Report (Yamamoto and Kulick, 2000).

[5] For a description of IRT scaling see Birnbaum (1968); Lord and Novick (1968); Lord (1980); Van Der Linden and Hambleton (1996). The theoretical underpinning of the multiple imputation methodology was developed by Rubin (1987), applied to large-scale assessment by Mislevy (1991), and studied further by Mislevy, Johnson and Muraki (1992), and Beaton and Johnson (1992). The procedures used in PIRLS have been used in several other large-scale surveys, including Trends in International Mathematics and Science Study (TIMSS), the U.S. National Assessment of Educational Progress (NAEP), the U.S. National Adult Literacy Survey (NALS), the International Adult Literacy Survey (IALS), and the International Adult Literacy and Life Skills Survey (IALLS).

(1) $$P\left(x_i = 1 \,\middle|\, \theta_k, a_i, b_i, c_i\right) = c_i + \frac{1 - c_i}{1 + \exp\left(-1.7 \cdot a_i(\theta_k - b_i)\right)} \equiv P_{i,1}\left(\theta_k\right)$$

where

x_i is the response to item *i*, 1 if correct and 0 if incorrect;

θ_k is the proficiency of a student on a scale *k* (note that a student with higher proficiency has a greater probability of responding correctly);

a_i is the slope parameter of item *i*, characterizing its discriminating power;

b_i is the location parameter of item *i*, characterizing its difficulty;

c_i is the lower asymptote parameter of item *i*, reflecting the chances of students with very low proficiency selecting the correct answer.

The probability of an incorrect response to the item is defined as:

(2) $$P_{i,0} = P\left(x_i = 0 \,\middle|\, \theta_k, a_i, b_i, c_i\right) = 1 - P_{i,1}\left(\theta_k\right)$$

The two-parameter (2PL) model was used for the short constructed-response items that were scored as either correct or incorrect. The form of the 2PL model is the same as Equations (1) and (2) with the c_i parameter fixed at zero.

11.2.2 IRT Model for Polytomous Items

In PIRLS 2006, as in PIRLS 2001, constructed-response items requiring an extended response were scored for partial credit, with 0, 1, 2 and 3 as the possible score levels. These polytomous items were scaled using a generalized partial credit model (Muraki, 1992). The fundamental equation of this model gives the probability that a student with proficiency θ_k on scale *k* will have, for the i^{th} item, a response x_i that is scored in the l^{th} of m_i ordered score categories as:

(3) $$P\left(x_i = l \,\middle|\, \theta_k, a_i, b_i, d_{i,1}, \cdots, d_{i,m_i-1}\right) = \frac{\exp\left(\sum_{v=0}^{l} 1.7 \cdot a_i\left(\theta_k - b_i + d_{i,v}\right)\right)}{\sum_{g=0}^{m_i-1} \exp\left(\sum_{v=0}^{g} 1.7 \cdot a_i\left(\theta_k - b_i + d_{i,v}\right)\right)} \equiv P_{i,l}\left(\theta_k\right)$$

where

m_i is the number of response categories for item *i*, either 3 or 4;

x_i is the response to item *i*, ranging between 0 and $m_i - 1$;

θ_k is the proficiency of a student on a scale *k*;

a_i is the slope parameter of item *i*;

b_i is its location parameter, characterizing its difficulty;

$d_{i,l}$ is the category *l* threshold parameter.

The indeterminacy of model parameters in the polytomous model is resolved by setting $d_{i,0} = 0$ and $\sum_{j=1}^{m_i-1} d_{i,j} = 0$.

For all of the IRT models there is a linear indeterminacy between the values of item parameters and proficiency parameters, i.e., mathematically equivalent but different values of item parameters can be estimated on an arbitrarily linearly transformed proficiency scale. This linear indeterminacy can be resolved by setting the origin and unit size of the proficiency scale to arbitrary constants, such as a mean of 500 and a standard deviation of 100, as was done for PIRLS in 2001. The indeterminacy is most apparent when the scale is set for the first time.

IRT modeling relies on a number of assumptions, the most important being conditional independence. Under this assumption, item response probabilities depend only on θ_k (a measure of a student's proficiency) and the specified parameters of the item, and are unaffected by the demographic characteristics or unique experiences of the students, the data collection conditions, or the other items presented in the test. Under this assumption, the joint probability of a particular response pattern *x* across a set of *n* items is given by:

$$\textbf{(4)} \qquad P\left(x \mid \theta_k, \textit{item parameters}\right) = \prod_{i=1}^{n} \prod_{l=0}^{m_i-1} P_{i,l}\left(\theta_k\right)^{u_{i,l}}$$

where $P_{il}\left(\theta_k\right)$ is of the form appropriate to the type of item (dichotomous or polytomous), m_i is equal to 2 for dichotomously scored items, and $u_{i,l}$ is an indicator variable defined as:

(5)
$$u_{i,l} = \begin{cases} 1 & \textit{if response } x_i \textit{ is in category } l; \\ 0 & \textit{otherwise}. \end{cases}$$

Replacing the hypothetical response pattern with the real scored data, the above function can be viewed as a likelihood function to be maximized by a given set of item parameters. In PIRLS 2006, the item parameters for each scale were estimated independently of the parameters of other scales. Once items were calibrated in this manner, a likelihood function for the proficiency θ_k was induced from student responses to the calibrated items. This likelihood function for the proficiency θ_k is called the posterior distribution of the θ's for each student.

11.2.3 Proficiency Estimation Using Plausible Values

Most cognitive skills testing is concerned with accurately assessing the performance of individual students for the purposes of diagnosis, selection, or placement. Regardless of the measurement model used, whether classical test theory or item response theory, the accuracy of these measurements can be improved—that is, the amount of measurement error can be reduced—by increasing the number of items given to the individual. Thus, it is common to see achievement tests designed to provide information on individual students that contain more than 70 items. Since the uncertainty associated with each θ in such tests is negligible, the distribution of θ, or the joint distribution of θ with other variables, can be approximated using each individual's estimated θ.

For the distribution of proficiencies in large populations, however, more efficient estimates can be obtained from a matrix-sampling design like that used in PIRLS. This design solicits relatively few responses from each sampled student while maintaining a wide range of content representation when responses are aggregated across all students. With this approach, however, the advantage of estimating population characteristics more efficiently is offset by the inability to make precise statements about individuals. The uncertainty associated with individual θ estimates becomes too large to be ignored. In this situation, aggregations of individual student scores can lead to seriously biased estimates of population characteristics (Wingersky, Kaplan, & Beaton, 1987).

Plausible values methodology was developed as a way to address this issue by using all available data to estimate directly the characteristics of student populations and subpopulations, and then generating multiple imputed scores, called plausible values, from these distributions that can be used in analyses with standard statistical software. A detailed review of the plausible values methodology is given in Mislevy (1991).[6]

The following is a brief overview of the plausible values approach. Let y represent the responses of all sampled students to background questions or background data of sampled students collected from other sources, and let θ represent the proficiency of interest. If θ were known for all sampled students, it would be possible to compute a statistic $t(\theta, y)$, such as a sample mean or sample percentile point, to estimate a corresponding population quantity T.

Because of the latent nature of the proficiency, however, θ values are not known even for sampled students. The solution to this problem is to follow Rubin (1987) by considering θ as "missing data" and approximate $t(\theta, y)$ by its expectation given (x, y), the data that actually were observed, as follows:

$$\begin{aligned} t^*(x, y) &= E\left| \, t(\underline{\theta}, \underline{y}) \,|\, \underline{x}, \underline{y} \, \right| \\ &= \int t(\underline{\theta}, \underline{y}) \, p\left(\underline{\theta} \,\middle|\, \underline{x}, \underline{y}\right) \, d\underline{\theta} \end{aligned} \tag{6}$$

It is possible to approximate t^* using random draws from the conditional distribution of the scale proficiencies given the student's item responses x_j, the student's background variables y_j, and model parameters for the items. These values are referred to as imputations in the sampling literature, and as plausible values in large-scale surveys such as PIRLS, TIMSS, NAEP, NALS, and IALS. The value of θ for any student that would enter into the computation of t is thus replaced by a randomly selected value from his or her conditional distribution. Rubin (1987) proposed repeating this process several times so that the uncertainly associated with imputation can be quantified. For example, the average of multiple estimates of t, each computed from a different set of plausible values, is a numerical approximation of t^* of the above equation; the variance among them reflects the uncertainty due to not observing $\underline{\theta}$. It should be noted that this variance does not include the variability of sampling from the population. That variability is estimated separately by jackknife variance estimation procedures, which are discussed in Chapter 12.

[6] Along with theoretical justifications, Mislevy presents comparisons with standard procedures; discusses biases that arise in some secondary analyses; and offers numerical examples.

Note that plausible values are not test scores for individuals in the usual sense, but rather are imputed values that may be used to estimate population characteristics correctly. When the underlying model is correctly specified, plausible values will provide consistent estimates of population characteristics, even though they are not generally unbiased estimates of the proficiencies of the individuals with whom they are associated.[7]

Plausible values for each student j are drawn from the conditional distribution $P\left(\theta_j \middle| x_j, y_j, \Gamma, \Sigma\right)$, where Γ is a matrix of regression coefficients for the background variables, and Σ is a common variance matrix of residuals. Using standard rules of probability, the conditional probability of proficiency can be represented as:.

(7) $$P\left(\theta_j \middle| x_j, y_j, \Gamma, \Sigma\right) \propto P\left(x_j \middle| \theta_j, y_j, \Gamma, \Sigma\right) P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right) = P\left(x_j \middle| \theta_j\right) P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$$

where θ_j is a vector of scale values, $P\left(x_j \middle| \theta_j\right)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$ is the multivariate joint density of proficiencies for the scales, conditional on the observed values y_j of background responses and parameters Γ and Σ. Item parameter estimates are fixed and regarded as population values in the computations described in this section.

11.2.4 Conditioning

A multivariate normal distribution was assumed for $P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right)$, with a common variance Σ, and with a mean given by a linear model with regression parameters Γ. Since in large-scale studies like PIRLS there are many hundreds of background variables, it is customary to conduct a principal components analysis to reduce the number of variables to be used in Γ. Typically, components accounting for 90 percent of the variance in the data are selected. These principal components are referred to as the conditioning variables and denoted as y^c. The following model is then fit to the data:

(8) $$\theta = \Gamma' y^c + \varepsilon$$

[7] For further discussion, see Mislevy, Beaton, Kaplan, and Sheehan (1992).

where ε is normally distributed with mean zero and variance Σ. As in a regression analysis, Γ is a matrix each of whose columns is the effects for each scale and Σ is the matrix of residual variance between scales.

Note that in order to be strictly correct for all functions Γ of θ, it is necessary that $P\left(\theta \middle| y\right)$ be correctly specified for all background variables in the survey. Estimates of functions Γ involving background variables not conditioned on in this manner are subject to estimation error due to misspecification. The nature of these errors is discussed in detail in Mislevy (1991). In PIRLS 2006, however, principal component scores based on nearly all background variables were used. Those selected variables were chosen to reflect high relevance to policy and to education practices. The computation of marginal means and percentile points of θ for these variables is nearly optimal.

The basic method for estimating Γ and Σ with the Expectation and Maximization (EM) procedure is described in Mislevy (1985) for a single scale case. The EM algorithm requires the computation of the mean θ, and variance Σ, of the posterior distribution in equation (7).

11.2.5 Generating Proficiency Scores

After completing the EM algorithm, plausible values for all sampled students are drawn from the joint distribution of the values of Γ in a three-step process.

First, a value of Γ is drawn from a normal approximation to $P\left(\Gamma, \Sigma \middle| x_j, y_j\right)$ that fixes Σ at the value $\hat{\Sigma}$ (Thomas, 1993). Second, conditional on the generated value of Γ (and the fixed value of $\Sigma = \hat{\Sigma}$), the mean θ_j and variance Σ_j^p of the posterior distribution in equation (7), where *p* is the number of scales, are computed using the methods applied in the EM algorithm. In the third step, the proficiency values are drawn independently from a multivariate normal distribution with mean θ_j and variance Σ_j^p. These three steps are repeated five times, producing five imputations of θ_j for each sampled student.

For students with an insufficient number of responses, the Γ's and Σ's described in the previous paragraph are fixed. Hence, all students—regardless of the number of items attempted—are assigned a set of plausible values.

The plausible values can then be employed to evaluate equation (6) for an arbitrary function *T* as follows:

- Using the first vector of plausible values for each student, evaluate *T* as if the plausible values were the true values of θ . Denote the result as T_1.
- Evaluate the sampling variance of T_1, or Var_1, with respect to students' first vector of plausible values.
- Carry out steps 1 and 2 for the second through fifth vectors of plausible values, thus obtaining T_u and Var_u for $u = 2, \ldots, 5$.
- The best estimate of *T* obtainable from the plausible values is the average of the five values obtained from the different sets of plausible values:

$$\hat{T} = \frac{\sum_u T_u}{5}$$

- An estimate of the variance of $\hat{T}$ is the sum of two components: an estimate of Var_u obtained by averaging as in the previous step, and the variance among the T_u's.

Let $\bar{U} = \frac{\sum_u Var_u}{M}$, and let $B_M = \frac{\sum_u \left(T_u - \hat{T}\right)^2}{M-1}$ be the variance among the *M* plausible values. Then the estimate of the total variance of $\hat{T}$ is:

$$Var\left(\hat{T}\right) = \bar{U} + \left(1 + M^{-1}\right) B_M \quad \textbf{(9)}$$

The first component in $Var\left(\hat{T}\right)$ reflects the uncertainty due to sampling students from the population; the second reflects the uncertainty due to the fact that sampled students' θ 's are not known precisely, but only indirectly through *x* and *y*.

11.2.6 Working with Plausible Values

The plausible values methodology was used in PIRLS 2006 to ensure the accuracy of estimates of the proficiency distributions for the PIRLS population as a whole and particularly for comparisons between subpopulations. A further advantage of this method is that the variation between the five plausible values generated for each student reflects the uncertainty associated with proficiency estimates for individual students. However, retaining this component of uncertainty requires that additional analytical procedures be used to estimate students' proficiencies.

If the θ values were observed for all sampled students, the statistic $\left(t-T\right)/U^{1/2}$ would follow a *t*-distribution with *d* degrees of freedom. Then the incomplete-data statistic $\left(T-\hat{T}\right) / \left[Var\left(\hat{T}\right)\right]^{1/2}$ is approximately *t*-distributed, with degrees of freedom (Johnson & Rust, 1993) given by:

(10)

$$\nu = \frac{1}{\frac{{f_M}^2}{M-1}+\frac{\left(1-f_M\right)^2}{d}}$$

where *d* is the degrees of freedom for the complete-data statistic, and f_M is the proportion of total variance due to not observing the values:

(11)

$$f_M = \frac{\left(1+M^{-1}\right)B_M}{Var\left(\hat{T}\right)}$$

When B_M is small relative to $\bar{U}$, the reference distribution for the incomplete-data statistic differs little from the reference distribution for the corresponding complete-data statistic. If, in addition, d is large, the normal approximation can be used instead of the *t*-distribution.

For a *k*-dimensional function *T*, such as the *k* coefficients in a multiple regression analysis, each U and $\bar{U}$ is a covariance matrix, and B_M is an average of squares and cross-products rather than simply an average of squares. In this case, the quantity $\left(\underline{T}-\hat{\underline{T}}\right) Var^{-1}\left(\hat{\underline{T}}\right)\left(\underline{T}-\hat{\underline{T}}\right)'$ is approximately *F*-distributed with degrees of freedom equal to *k* and ν, with ν defined as above but with a matrix generalization of f_M:

(12)
$$f_M = \left(1+M^{-1}\right) Trace\left[B_M Var^{-1}\left(\hat{T}\right)\right] \Big/ k$$

For the same reason that the normal distribution can approximate the *t*-distribution, a chi-square distribution with *k* degrees of freedom can be used in place of the *F*-distribution for evaluating the significance of the above quantity $\left(\underline{T}-\hat{\underline{T}}\right) Var^{-1}\left(\hat{\underline{T}}\right)\left(\underline{T}-\hat{\underline{T}}\right)'$.

Statistics $\hat{T}$, the estimates of proficiency conditional on responses to cognitive items and background variables, are consistent estimates of the corresponding population values *T*, as long as background variables are included in the conditioning variables. The consequences of violating this restriction are described by Beaton & Johnson (1990), Mislevy (1991), and Mislevy & Sheehan (1987). To avoid such biases, the PIRLS 2006 analyses included nearly all background variables.

11.3 Implementing the Scaling Procedures for the PIRLS 2006 Assessment Data

The application of IRT scaling and plausible value methodology to the PIRLS 2006 assessment data involved four major tasks: calibrating the achievement test items (estimating model parameters for each item), creating principal components from the student and home questionnaire data for use in conditioning; generating IRT scale scores (proficiency scores) for overall reading, the two purposes of reading (reading for literary experience and reading to acquire and use information) and the two processes of reading (processes of retrieving and straightforward inferencing and processes of interpreting, integrating, and evaluating); and placing the proficiency scores on the metric used to report the results from 2001. The PIRLS reporting metric was established by setting the average of the mean scores of the countries that participated in PIRLS 2001 to 500 and the standard deviation to 100. To enable comparisons between 2006 and 2001, the PIRLS 2006 data also were placed on this metric.

11.3.1 Calibrating the PIRLS 2006 Test Items

In striving to measure trends in a changing world, PIRLS releases a number of assessment blocks after each assessment year and replaces them with newly developed blocks that incorporate current thinking of reading literacy and approaches to reading instruction. A number of assessment blocks also are kept secure to be used again in future assessments. The PIRLS 2006 item calibration is based on all items from 2006 and 2001 and all countries that participated in both assessments. This is known as concurrent calibration. The common items are used to ensure that there is sufficient overlap between the current assessment and the previous one, however, the 2001 items that were ultimately released and the items that were developed for 2006 also contribute to setting the PIRLS 2006 scales. Exhibit 11.1 shows the distribution of items included in the PIRLS 2006 calibrations for all five PIRLS scales. The 174 items included in the overall scale were divided between those measuring reading for literary experience (89 items) and for information (85 items) for calibrating the two reading purposes scales, and between those measuring retrieving and straightforward inferencing (96 items) and those measuring interpreting, integrating, and evaluating (78 items) for calibrating the two comprehension processes scales. Exhibit 11.2 lists the countries included in the item calibrations and their sample sizes for both assessment years. A total of 225,542 students from 26 countries contributed to the item calibrations.

Exhibit 11.1 Items Included in the PIRLS 2006 Item Calibrations

PIRLS Scales		Items Unique to PIRLS 2001		Items Unique to PIRLS 2006 [1]		Items in Both Assessment Cycles		Total	
		Number	Points	Number	Points	Number	Points	Number	Points
Overall Reading		49	67	76	99	49	66	174	232
Purposes of Reading	Literary Experience	25	33	38	51	26	33	89	117
	Acquire and Use Information	24	34	38	48	23	33	85	115
Processes of Reading	Retrieving and Straightforward Inferencing	22	24	44	47	30	36	96	107
	Interpreting, Integrating, and Evaluating	27	43	32	52	19	30	78	125

[1] Item R021S08M was removed from all item calibrations because of poor psychometric properties.

Exhibit 11.2 Samples Included in the PIRLS 2006 Item Calibrations

Countries	Sample Sizes	
	PIRLS 2006	PIRLS 2001
Bulgaria	3,863	3,460
England	4,036	3,156
France	4,404	3,538
Germany	7,899	7,633
Hong Kong SAR	4,712	5,050
Hungary	4,068	4,666
Iceland	3,673	3,676
Iran, Islamic Rep. of	5,411	7,430
Israel	3,908	3,973
Italy	3,581	3,502
Latvia	4,162	3,019
Lithuania	4,701	2,567
Macedonia, Rep. of	4,002	3,711
Moldova, Rep. of	4,036	3,533
Morocco	3,249	3,153
Netherlands	4,156	4,112
New Zealand	6,256	2,488
Norway	3,837	3,459
Romania	4,273	3,625
Russian Federation	4,720	4,093
Scotland	3,775	2,717
Singapore	6,390	7,002
Slovak Republic	5,380	3,807
Slovenia	5,337	2,952
Sweden	4,394	6,044
United States	5,190	3,763
Total	119,413	106,129

In line with the PIRLS assessment framework, IRT scales were constructed for reporting student achievement in overall reading, as well as for reporting separately for each of the two purposes of reading and the two processes of reading. The first step in constructing these scales was to estimate the IRT model item parameters for each item on each of the five PIRLS scales. This item calibration was conducted using the commercially-available PARSCALE software (Muraki & Bock, 1991; version 4.1). Item calibration included data from PIRLS 2006 and PIRLS 2001 for countries that participated in both assessment years in order to measure trends from 2001 to 2006. The assessment data were weighted to ensure that the data from each country and each assessment year contributed equally to the item calibration.

Five separate item calibrations were run: one for the overall reading scale; one for each of the two purposes of reading—literary experience and acquire and use information; and one for each of the two processes of reading—retrieving and straightforward inferencing and interpreting, integrating, and evaluating. Exhibits D.1 through D.5 in Appendix D display the item parameters estimated from the five calibration runs. All items and all students involved in the calibration process were included in the calibration of the overall reading scale. Interim reading scores[8] were produced as a by-product of this first calibration for use in generating conditioning variables. For the calibration of the literary experience scale, only items from literary assessment blocks and only those students completing a booklet with a literary block (183,431) were included. Similarly, only items from information assessment blocks and only those students completing a booklet with an information block (183,253) were included in the calibration of the acquire and use information scale. The situation was somewhat different for the two processes of reading since all assessment blocks, regardless of their purpose of reading, had a mix of items classified in the two processes of reading. Thus, only items classified in the retrieving and straightforward inferencing process and nearly all students[9] (225,539) were included in the calibration of the retrieving and straightforward inferencing scale, and only items classified in the interpreting, integrating, and evaluating process and nearly all students (225,435) were included in the calibration of the interpreting, integrating, and evaluating scale.

11.3.2 Omitted and Not-Reached Responses

Apart from missing data on items that by design were not administered to a student, missing data could also occur because a student did not answer an item—whether because the student did not know the answer, omitted it by mistake, or did not have time to attempt the item. An item was considered not reached when (within part 1 or part 2 of the booklet) the item itself and the item immediately preceding were not answered, and there were no other items completed in the remainder of the booklet.

In PIRLS 2006, as in 2001, not-reached items were treated differently in estimating item parameters and in generating student proficiency scores. In estimating the values of the item parameters, items that were considered not to have been reached by students were treated as if they had not been administered.

[8] Because each student responded to only a subset of the assessment item pool, these interim scores, known as expected a priori—or EAP—scores, were not sufficiently reliable for reporting PIRLS results. The plausible value proficiency scores were used for this purpose.

[9] Three students did not respond to any items classified in the "retrieving and straightforward inferencing" process and 107 students did not respond to any items classified in the "interpreting, integrating, and evaluating" process.

This approach was considered optimal for item parameter estimation. However, not-reached items were considered as incorrect responses when student proficiency scores were generated.

11.3.3 Evaluating Fit of IRT Models to the PIRLS 2006 Data

After the calibrations were completed, checks were performed to verify that the item parameters obtained from PARSCALE adequately reproduced the observed distribution of responses across the proficiency continuum. The fit of the IRT models to the PIRLS 2006 data was examined by comparing the theoretical item response function curves generated using the item parameters estimated from the data with the empirical item response function curves calculated from the posterior distributions of the θ's for each student that responded to the item. Graphical plots of the theoretical and empirical item response function curves are called item characteristic curves (ICC).

Exhibit 11.3 shows an ICC plot of the empirical and theoretical item response functions for a dichotomous item. In the plot, the horizontal axis represents the proficiency scale, and the vertical axis represents the probability of a correct response. The theoretical curve based on the estimated item parameters is shown as a solid line. Empirical results are represented by circles. The empirical results were obtained by first dividing the proficiency scale into intervals of equal size and then counting the number of students responding to the item whose EAP scores from PARSCALE fell in each interval. Then the proportion of students in each interval that responded correctly to the item was calculated.[10] In the exhibit, the center of each circle represents this empirical proportion of correct responses. The size of each circle is proportional to the number of students contributing to the estimation of its empirical proportion correct.

[10] These calculations were performed using the SENWGT.

Exhibit 11.3 PIRLS 2006 Reading Assessment Example Item Response Function for a Dichotomous Item

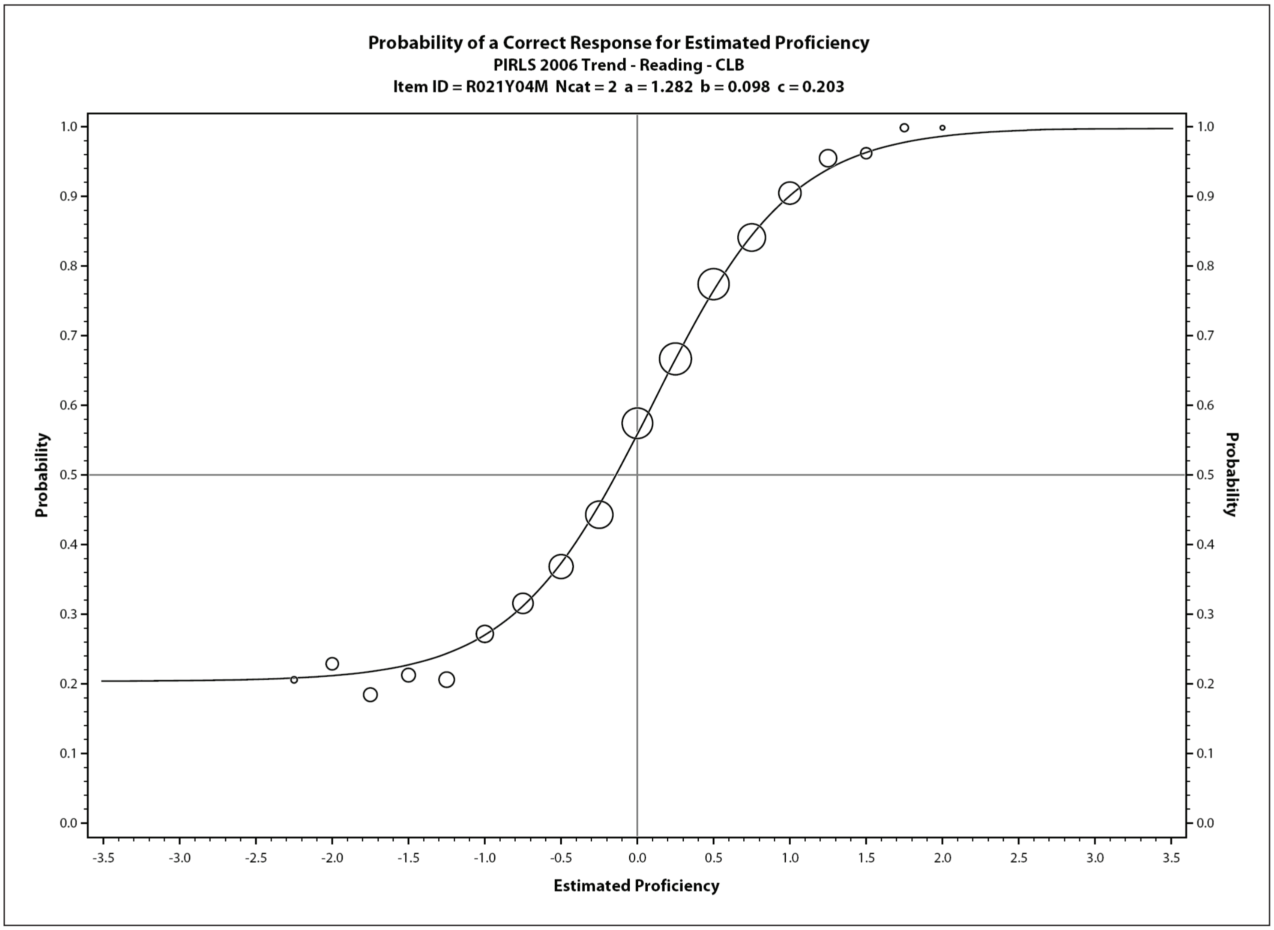
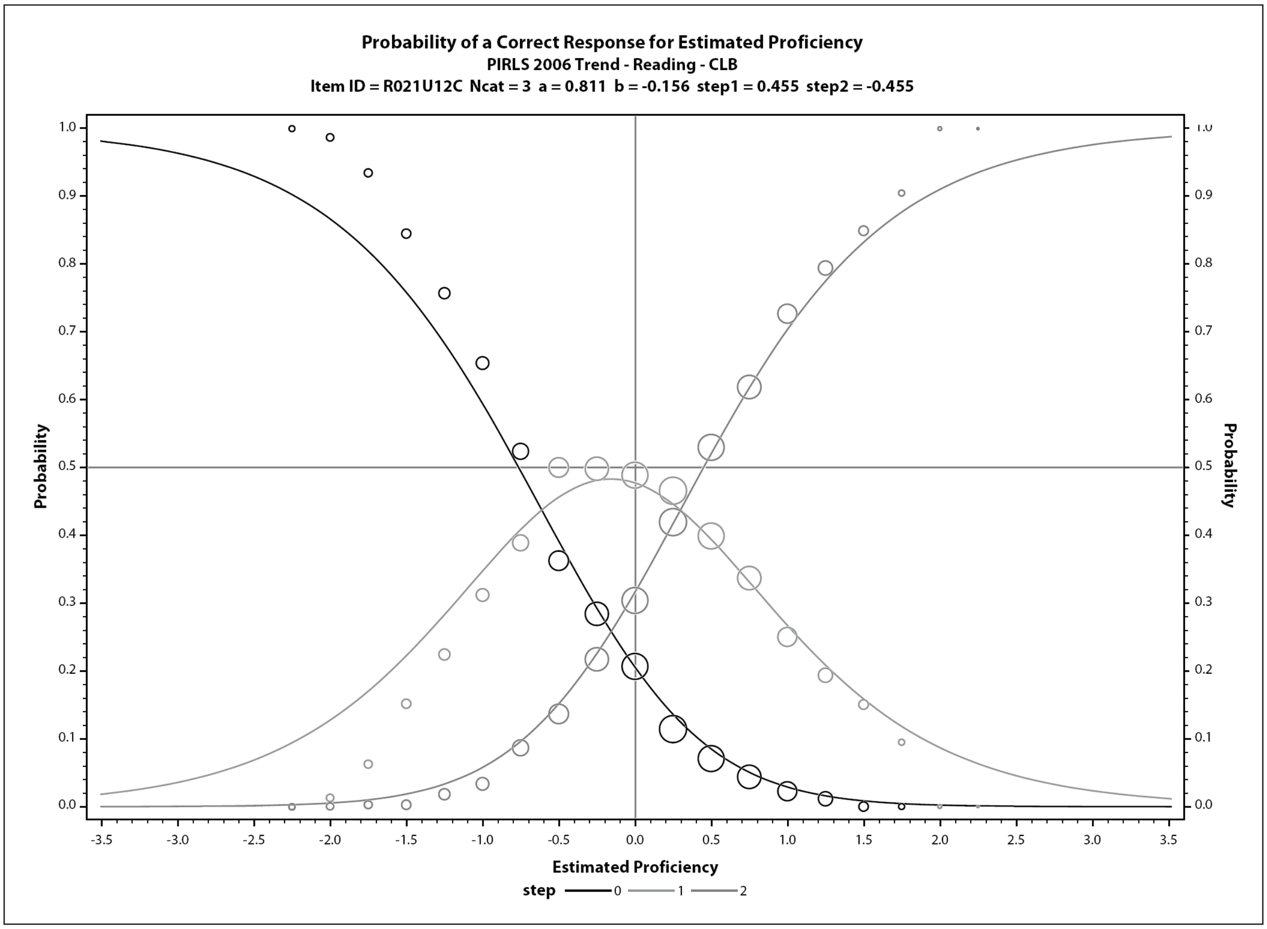


Exhibit 11.4 contains an ICC plot of the empirical and theoretical item response functions for a polytomous item. As for the dichotomous item plot, the horizontal axis represents the proficiency scale, but the vertical axis represents the probability of having a response in a given response category. The theoretical curves based on the estimated item parameters are shown as solid lines. Empirical results are represented by circles. The interpretation of the circles is the same as in Exhibit 11.3. For items where the IRT model fits the data well, the empirical results fall near the theoretical curves.

Exhibit 11.4 PIRLS 2006 Reading Assessment Example Item Response Function for a Polytomous Item



11.3.4 Variables for Conditioning the PIRLS 2006 Data

PIRLS 2006 used all background variables from the student background questionnaire and the Learning to Read Survey questionnaire. Because there were so many background variables that could be used in conditioning, PIRLS followed the practice established in other large-scale studies of using principal components analysis to reduce the number of variables while explaining most of their common variance. Principal components for the PIRLS 2006 background data were constructed as follows:

- For categorical variables (questions with a small number of fixed response options), a "dummy coded" variable was created for each response option, with a value of one if the option was chosen and zero otherwise. If a student omitted or was not administered a particular question, all dummy coded variables associated with that question were assigned the value zero.

- Background variables with numerous response options (such as year of birth, or number of people who live in the home) were recoded using criterion scaling.[11] This was done by replacing each response option with the mean interim (EAP) score of the students choosing that option.
- Separately for each PIRLS country, all the dummy-coded and criterion-scaled variables were included in a principal components analysis. Those principal components accounting for 90 percent of the variance of the background variables were retained for use as conditioning variables. Because the principal components analysis was performed separately for each country, different numbers of principal components were required to account for 90% of the common variance in each country's background variables.

In addition to the principal components, student gender (dummy coded), the language of the test (dummy coded), an indicator of the classroom in the school to which the student belonged (criterion scaled), and an optional, country-specific variable (dummy coded) were included as conditioning variables. These additional variables are characterized as primary conditioning variables. Exhibit 11.5 shows the total number of variables that were used for conditioning.

[11] The process of generating criterion-scaled variables is described in Beaton (1969).

Exhibit 11.5 Number of Variables Used for Conditioning in PIRLS 2006

Countries	Sample Sizes	Number of Background Variables Available	Conditioning Variables	
			Principal Components	Primary Conditioning Variables
Austria	5,067	526	295	2
Belgium (Flemish)	4,479	520	286	2
Belgium (French)	4,552	514	291	2
Bulgaria	3,863	528	282	2
Canada, Alberta	4,243	495	274	3
Canada, British Columbia	4,150	495	274	3
Canada, Nova Scotia	4,436	495	279	3
Canada, Ontario	3,988	495	272	3
Canada, Quebec	3,748	495	275	3
Chinese Taipei	4,589	519	295	2
Denmark	4,001	528	289	2
England	4,036	528	280	2
France	4,404	516	293	2
Georgia	4,402	518	297	2
Germany	7,899	520	291	2
Hong Kong SAR	4,712	530	299	2
Hungary	4,068	497	278	2
Iceland	3,673	506	284	2
Indonesia	4,774	492	291	2
Iran, Islamic Rep. of	5,411	530	297	2
Israel	3,908	530	296	3
Italy	3,581	530	292	2
Kuwait	3,958	509	299	2
Latvia	4,162	525	292	3
Lithuania	4,701	511	290	2
Luxembourg	5,101	522	292	2
Macedonia, Rep. of	4,002	528	303	3
Moldova, Rep. of	4,036	530	294	3
Morocco	3,249	506	286	2
Netherlands	4,156	520	281	2
New Zealand	6,256	520	287	8
Norway	3,837	522	283	3
Poland	4,854	501	284	2
Qatar	6,680	526	310	2
Romania	4,273	530	289	3
Russian Federation	4,720	500	282	2
Scotland	3,775	528	277	2
Singapore	6,390	526	296	2
Slovak Republic	5,380	524	293	3
Slovenia	5,337	518	290	2
South Africa	14,657	503	312	12
Spain	4,094	528	285	6
Sweden	4,394	528	289	2
Trinidad and Tobago	3,951	491	281	2
United States[1]	5,190	285	166	7

[1] The United States did not administer the "Learning to Read Survey" questionnaire, thus reducing the number of background variables available for conditioning.

11.3.5 Generating IRT Proficiency Scores for the PIRLS 2006 Data

The MGROUP program (Sheehan, 1985; version 3.2)[12] was used to generate the IRT proficiency scores. This program takes as input the students' responses to the items they were given, the item parameters estimated at the calibration stage, and the conditioning variables, and generates as output the plausible values that represent student proficiency. For each of the 45 PIRLS participants listed in Exhibit 11.5, it was necessary to run MGROUP three times to produce the PIRLS 2006 assessment scales: one unidimensional run for the overall reading scale, one multidimensional run for the reading purposes scales, and one multidimensional run for the comprehension processes scales. Thus a total of 135 (45x3) MGROUP runs were required to obtain proficiency scores for PIRLS 2006.

In addition to generating plausible values for the PIRLS 2006 data, the parameters estimated at the calibration stage also were used to generate plausible values on all five PIRLS scales using the 2001 data for the 26 trend countries that participated in both assessment years. These plausible values for the trend countries are called "link scores." Link scores were also produced for the Canadian provinces of Ontario and Quebec for evaluation purposes. Producing the link scores required 84 additional MGROUP runs.

Plausible values generated by the conditioning program are initially on the same scale as the item parameters used to estimate them. This scale metric is generally not useful for reporting purposes since it is somewhat arbitrary, ranges between approximately –3 and +3, and has an expected mean of zero across all countries.

11.3.6 Transforming the Proficiency Scores to Measure Trends between 2001 and 2006

To provide results for PIRLS 2006 comparable to the results from the PIRLS 2001 assessment, the 2006 proficiency scores (plausible values) had to be transformed to the metric used in 2001. To accomplish this, the means and standard deviations of the link scores for all five PIRLS scales were made to match the means and standard deviations of the scores reported in the 2001 assessment by applying the appropriate linear transformations. These linear transformations are given by:

(13)

$$PV_{k,i}^{\ *} = A_{k,i} + B_{k,i} \cdot PV_{k,i}$$

[12] The MGROUP program was provided by ETS under contract to the TIMSS and PIRLS International Study Center at Boston College.

where

$PV_{k,i}$ is the plausible value *i* of scale *k* prior to transformation;

$PV_{k,i}^{*}$ is the plausible value *i* of scale *k* after transformation;

and $A_{k,i}$ and $B_{k,i}$ are the linear transformation constants.

The linear transformation constants were obtained by first computing, using the senate weight, the international means and standard deviations of the proficiency scores for all five PIRLS scales using the plausible values generated in 2001 for the 26 trend countries. Next, the same calculations were done using the 2006 link scores of the 26 trend countries. The linear transformation constants are defined as:

(14)

$$\begin{aligned} B_{k,i} &= \sigma_{k,i} / \sigma_{k,i}^{*} \\ A_{k,i} &= \mu_{k,i} - B_{k,i}\ \mu_{k,i}^{*} \end{aligned}$$

where

$\mu_{k,i}$ is the international mean of scale *k* based on plausible value *i* released in 2001;

$\mu_{k,i}^{*}$ is the international mean of scale *k* based on plausible value *i* of the 2006 link scores;

$\sigma_{k,i}$ is the international standard deviation of scale *k* based on plausible value *i* released in 2001;

$\sigma_{k,i}^{*}$ is the international standard deviation of scale *k* based on plausible value *i* of the 2006 link scores.

Exhibit 11.6 shows the linear transformation constants that were computed.

Once the linear transformation constants were established, all of the proficiency scores from the 2006 assessment were transformed by applying the

Exhibit 11.6 Linear Transformation Constants Used for the PIRLS 2006 Data

Scale		Plausible Values	PIRLS 2001 Scores		2006 "Link Scores"		$A_{k,i}$	$B_{k,i}$
			Mean	Standard Deviation	Mean	Standard Deviation		
Overall Reading		PV1	514.9855	91.9947	-0.0435	0.9018	102.0152	519.4237
		PV2	514.8861	92.1770	-0.0399	0.8999	102.4341	518.9761
		PV3	514.8006	92.3735	-0.0400	0.9006	102.5698	518.8983
		PV4	514.8252	92.2470	-0.0390	0.8991	102.5944	518.8265
		PV5	514.7781	92.2987	-0.0414	0.9012	102.4191	519.0163
Purposes of Reading	Literary Experience	PV1	514.6110	92.5091	0.1699	0.9962	92.8640	498.8316
		PV2	514.5735	92.5937	0.1694	0.9947	93.0840	498.8075
		PV3	514.4664	92.4649	0.1722	0.9979	92.6575	498.5120
		PV4	514.6021	92.6655	0.1711	0.9965	92.9868	498.6943
		PV5	514.4937	92.7265	0.1723	0.9988	92.8355	498.5015
	Acquire and Use Information	PV1	514.5481	92.2754	0.0737	0.9664	95.4885	507.5074
		PV2	514.3908	92.2856	0.0735	0.9672	95.4108	507.3758
		PV3	514.6731	92.0767	0.0708	0.9656	95.3604	507.9254
		PV4	514.5654	92.0253	0.0709	0.9671	95.1549	507.8201
		PV5	514.5440	92.1947	0.0681	0.9663	95.4119	508.0446
Processes of Reading	Retrieving and Straightforward Inferencing	PV1	514.3950	93.7040	0.0233	0.9984	93.8557	512.2098
		PV2	514.6367	93.6133	0.0216	0.9995	93.6559	512.6105
		PV3	514.4507	93.7383	0.0206	0.9971	94.0063	512.5182
		PV4	514.3605	93.5307	0.0215	1.0014	93.4021	512.3508
		PV5	514.3732	93.7112	0.0220	1.0000	93.7112	512.3132
	Interpreting, Integrating, and Evaluating	PV1	515.2249	90.9159	-0.1075	0.9767	93.0841	525.2357
		PV2	515.0767	91.1286	-0.1112	0.9806	92.9353	525.4140
		PV3	515.0542	91.1681	-0.1101	0.9814	92.8949	525.2859
		PV4	515.0299	90.9888	-0.1118	0.9845	92.4232	525.3639
		PV5	515.0590	91.1741	-0.1133	0.9826	92.7873	525.5682

same linear transformations for all countries. This provided achievement scores for the PIRLS 2006 assessment that were directly comparable to the scores from the 2001 assessment.

References

Beaton, A.E. (1969). Scaling criterion of questionnaire items. *Socio-Economic Planning Sciences, 2,* 355-362.

Beaton, A.E., & Johnson, E.G. (1990). The average response method of scaling. *Journal of Educational Statistics, 15,* 9-38.

Beaton, A.E., & Johnson, E.G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement*, *26,* 163–175.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores*, (pp.397–479). Reading, MA: Addison-Wesley Publishing.

Johnson, E.G., & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17,* 175–190.

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

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

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80,* 993-97.

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

Mislevy, R.J., Beaton, A., Kaplan, B.A., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161.

Mislevy, R.J., Johnson, E.G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*, 131–154.

Mislevy, R.J., & Sheehan, K. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983–84 technical report* (pp. 293–360). (no. 15-TR-20) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Mullis, I.V.S., Kennedy, A.M., Martin, M.O., & Sainsbury, M. (2006). *PIRLS 2006 assessment framework and specifications* (2nd ed.). Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., & Gonzalez, E.J., (2004), *International achievement in the processes of reading comprehension: Results from PIRLS 2001 in 35 countries*, Chestnut Hill, MA: Boston College.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2)*, 159-176.

References *(continued)*

Muraki, E., & Bock, R.D. (1991). PARSCALE: Parameter scaling of rating data [Computer software and manual], Chicago, IL: Scientific Software, Inc.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Sheehan, K. M. (1985). M-GROUP: Estimation of group effects in multivariate models [Computer program]. Princeton, NJ: Educational Testing Service.

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2,* 309–22.

Van Der Linden, W.J. & Hambleton, R. (1996). *Handbook of modern item response theory.* New York. Springer-Verlag.

Wingersky, M., Kaplan, B.A., & Beaton, A.E. (1987). Joint estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983–84 technical report* (pp.285–92) (No. 15-TR-20). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Yamamoto, K., & Kulick, E. (2000). Scaling methodology and procedures for the TIMSS mathematics and science scales. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS 1999 technical report*. Chestnut Hill, MA: Boston College.