CHAPTER 11

PIRLS 2016 Achievement Scaling Methodology[1]

The PIRLS approach to scaling the achievement data, based on item response theory (IRT) scaling with marginal estimation, was developed originally by Educational Testing Service for use in the U.S. National Assessment of Educational Progress (NAEP). 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.[2]

Three distinct IRT models, depending on item type and scoring procedure, were used in the analysis of the PIRLS 2016 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.

Two– and Three–Parameter IRT Models for Dichotomous Items

The fundamental equation of the three-parameter logistic (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:

$$P\left(x_i=1\middle|\,\theta_k,a_i,b_i,c_i\right)=c_i+\frac{1-c_i}{1+\exp\left(-1.7\cdot a_i\cdot(\theta_k-b_i)\right)}\equiv P_{i,1}\left(\theta_k\right) \qquad (11.1)$$

1 This description of the PIRLS achievement scaling methodology has been adapted with permission from the TIMSS 1999 Technical Report (Yamamoto and Kulick, 2000).

2 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). For a recent overview, see von Davier and Sinharay (2014) and von Davier (2014). The procedures used in PIRLS have been used in several other large-scale surveys, including 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).

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:

$$P_{i,0} = P\left(x_i = 0 \,\middle|\, \theta_k, a_i, b_i, c_i\right) = 1 - P_{i,1}\left(\theta_k\right) \tag{11.2}$$

The two-parameter logistic (2PL) model was used for the constructed response items that were scored as either correct or incorrect. The form of the 2PL model is the same as Equations (11.1) and (11.2) with the c_i parameter fixed at zero.

IRT Model for Polytomous Items

In PIRLS, 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:

$$P\left(x_i = l \,\middle|\, \theta_k, a_i, b_i, d_{i,1}, \cdots, d_{i,m_i-1}\right) = \frac{\exp\left(\sum_{v=0}^{l} 1.7 \cdot a_i \cdot \left(\theta_k - b_i + d_{i,v}\right)\right)}{\sum_{g=0}^{m_i-1} \exp\left(\sum_{v=0}^{g} 1.7 \cdot a_i \cdot \left(\theta_k - b_i + d_{i,v}\right)\right)} = P_{i,l}\left(\theta_k\right) \tag{11.3}$$

where

m_i is the number of response categories for item i;

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 originally for PIRLS 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:

$$P\left(x \middle| \theta_k, \textit{item parameters}\right) = \prod_{i=1}^{n} \prod_{l=0}^{m_i - 1} P_{i,l}\left(\theta_k\right)^{u_{i,l}} \tag{11.4}$$

where $P_{i,l}(\theta_k)$ 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:

$$u_{i,l} = \begin{cases} 1 \textit{ if response is } x_i \textit{ is in category } l; \\ 0 \textit{ otherwise} \end{cases} \tag{11.5}$$

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. Once items are calibrated in this manner, a likelihood function for the proficiency θ_k is 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.

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 (Martin, Mullis, & Foy, 2015). 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. Indeed, 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. Instead of first computing estimates of individual θ's and then aggregating these to estimate population parameters, the plausible values approach uses all available data, students' responses to the items they were administered together with all background data, to estimate directly the characteristics of student populations and subpopulations. Although these directly estimated population characteristics could be used for reporting purposes, instead the usual plausible values approach is to generate multiple imputed scores, called plausible values, from the estimated ability distributions and to use these in analyses and reporting, making use of standard statistical software. By including all available background data in the model, a process known as "conditioning," relationships between these background variables and the estimated proficiencies will be appropriately accounted for in the plausible values. Because of this, analyses conducted using plausible values will provide an accurate representation of these underlying relationships. A detailed review of the plausible values methodology is given in Mislevy (1991).[3]

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:

$$\begin{aligned} t^*(x,y) &= E\left[t(\underline{\theta},\underline{y}) \,|\, \underline{x},\underline{y} \right] \\ &= \int t(\underline{\theta},\underline{y})\, p(\underline{\theta} \,|\, \underline{x},\underline{y})\, d\underline{\theta} \end{aligned} \tag{11.6}$$

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

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 IALLS. 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 a jackknife variance estimation procedure.

Plausible values are not intended to be estimates of individual student scores, but rather are imputed scores for like students—students with similar response patterns and background characteristics in the sampled population—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 generally biased estimates of the proficiencies of the individuals with whom they are associated. Taking the average of the plausible values still will not yield suitable estimates of individual student scores.[4]

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:

$$P\left(\theta_j \middle| x_j, y_j, \Gamma, \Sigma\right) \propto P\left(x_j \middle| \theta_j, y_j, \Gamma, \Sigma\right) P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right) = P\left(x_j \middle| \theta_j\right) P\left(\theta_j \middle| y_j, \Gamma, \Sigma\right) \quad (11.7)$$

where θ_j is a vector of scale values, $P(x_j|\theta_j)$ is the product over the scales of the independent likelihoods induced by responses to items within each scale, and $P(\theta_j|y_j, \Gamma, \Sigma)$ 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.

4 For further discussion, see Mislevy, Beaton, Kaplan, and Sheehan (1992).

Conditioning

A multivariate normal distribution was assumed for $P(\theta_j|y_j, \Gamma, \Sigma)$, 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:

$$\theta = \Gamma' y^c + \varepsilon \tag{11.8}$$

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 $P(\theta|y)$ be correctly specified for all background variables in the survey. Estimates of functions Γ involving background variables not conditioned 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, however, the principal components account for almost all of the variance in the student background variables, so that 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 (11.7).

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 $P(\Gamma,\Sigma\,|x_j, y_j)$ 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 (11.7), 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 (11.6) 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:

$$\hat{T} = \frac{\sum_u T_u}{5} \tag{11.9}$$

- 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 $\overline{U} = \frac{\sum_u Var_u}{M}$, and let $B_M = \frac{\sum_u (T_u - \hat{T})^2}{M-1}$ be the variance among the M plausible values

Then the estimate of the total variance of $\hat{T}$ is:

$$Var(\hat{T}) = \overline{U} + (1 + M^{-1}) B_M \tag{11.10}$$

The first component in $Var(\hat{T})$ 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.

Working with Plausible Values

The plausible values methodology is used in PIRLS to ensure the accuracy of estimates of the proficiency distributions for the PIRLS populations 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 $(t-T)/U^{\frac{1}{2}}$ would follow a t-distribution with d degrees of freedom. Then the incomplete-data statistic $(T-\hat{T})/\left[Var(\hat{T})\right]^{\frac{1}{2}}$ is approximately t-distributed, with degrees of freedom (Johnson & Rust, 1992) given by:

$$v = \frac{1}{\frac{f_M^{\ 2}}{M-1} + \frac{(1-f_M)^2}{d}} \tag{11.11}$$

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:

$$f_M = \frac{(1+M^{-1})\,B_M}{Var(\hat{T})} \tag{11.12}$$

When B_M is small relative to $\overline{U}$, 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 $\overline{U}$ 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 $(\underline{T}-\underline{\hat{T}})\ Var^{-1}(\underline{\hat{T}})\ (\underline{T}-\underline{\hat{T}})'$ is approximately F-distributed with degrees of freedom equal to k and v, with v defined as above but with a matrix generalization of f_M:

$$f_M = (1+M^{-1})\ \text{Trace}\left[B_M\ Var^{-1}(\hat{T})\right]/k \tag{11.13}$$

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 $(\underline{T}-\underline{\hat{T}})\ Var^{-1}(\underline{\hat{T}})\ (\underline{T}-\underline{\hat{T}})'$.

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 and Johnson (1992), Mislevy (1991), and Mislevy and Sheehan (1987). To avoid such biases, the PIRLS analyses include nearly all student background variables, in the form of principal components, as well as the class means to preserve between-class differences—the between-classroom and within-classroom variance structure essential for hierarchical modeling.

References

Beaton, A.E., & Johnson, E.G. (1992). Overview of the scaling methodology used in the National Assessment. *Journal of Educational Measurement, 26*(2), 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 (Eds.), *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*(2), 175–190.

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

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

Martin, M.O., Mullis, I.V.S., & Foy, P. (2015). Assessment design for PIRLS, PIRLS Literacy, and ePIRLS in 2016. In I.V.S. Mullis & M.O. Martin (Eds.), *PIRLS 2016 Assessment Framework, 2nd Edition* (pp. 55-69). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*, 993–997.

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

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*(2), 133–161.

Mislevy, R.J., Johnson, E.G. & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*(2), 131–154.

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

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

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

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics*, *2*, 309–322.

Van Der Linden, W.J., & Hambleton, R. (1996). *Handbook of Modern Item Response Theory*. New York. Springer-Verlag.

von Davier, M. (2014). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues and methods of data analysis* (pp. 175–201). Boca Raton: Chapman & Hall/CRC.

von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues and methods of data analysis* (pp. 155–174). Boca Raton: Chapman & Hall/CRC.

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