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:

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

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

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

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

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

–

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*