CHAPTER 12

TIMSS 2015 Achievement Scaling Methodology[1]

The TIMSS 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]

where

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

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

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

is the location parameter of item , characterizing its difficulty;

is the lower asymptote parameter of item , 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:

$$P_{,0} = P(\quad = 0 | \quad , \quad , \quad , \quad) = 1 - P_{,1}(\quad) \qquad (2)$$

The two-parameter (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 (1) and (2) with the parameter fixed at zero.

IRT Model for Polytomous Items

In TIMSS, constructed response items requiring an extended response were scored for partial credit, with 0, 1, and 2 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 on scale will have, for the th

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

For the distribution of proficiencies in large populations, however, more efficient estimates can be obtained from a matrix-sampling design like that used in TIMSS. 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 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 ’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

It is possible to approximate * using random draws from the conditional distribution of the scale proficiencies given the student’s item responses

$-$

Conditioning

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

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

$$ = \frac{1}{\frac{{}_{M}^{2}}{M-1} + \frac{\left(1- {}_{M}\right)^{2}}{}} \tag{11}$$

where is the degrees of freedom for the complete-data statistic, and ${}_M$ is the proportion of total variance due to not observing the values:

$${}_M = \frac{\left(1+M^{-1}\right) B_M}{ (T)} \tag{12}$$

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, is large, the normal approximation can be used instead of the -distribution.

For a -dimensional function T, such as the coefficients in a multiple regression analysis, each U