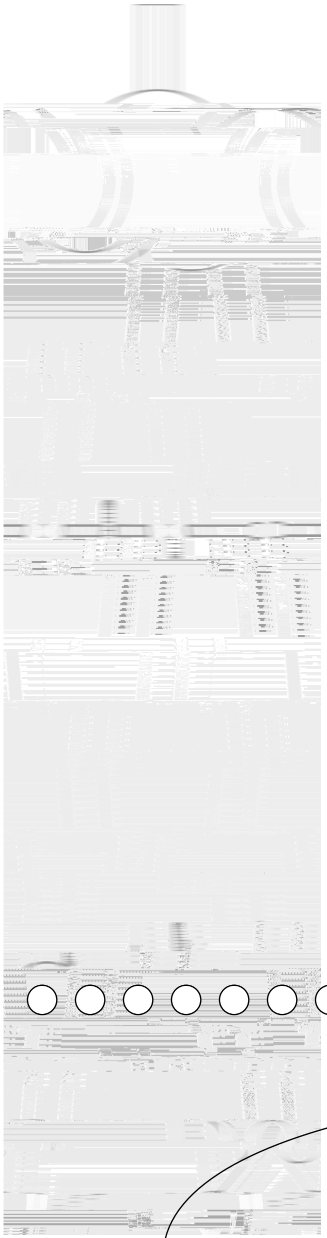
1



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 free-response items (i.e., those with more than two score points).

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

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

$$P_{i1}(x_i = 1|\theta_k, a_i, b_i, c_i) = c_i + \frac{(1 - c_i)}{1.0 + \exp(-1.7a_i(\theta_k - b_i))} \tag{1}$$

where

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

θ_k is the proficiency of a person on a scale k;

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

b_i is its location parameter, characterizing its difficulty;

c_i is its lower asymptote parameter, reflecting the chances of respondents of very low proficiency selecting the correct answer.

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

$$P_{i0} \equiv P(x_i = 0|\theta_k, a_i, b_i, c_i) = 1 - P_{i1}(\theta_k) \quad . \tag{2}$$

The two-parameter logistic (2PL) model was used for the short free-response items that were scored as 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.

In scaling the Benchmarking data, the three- and two-parameter models were used in preference to the one-parameter Rasch model, primarily because they can more accurately account for the differences among items in their ability to discriminate between students of high and low ability. With the Rasch model, all items are assumed to have the same discriminating power, while the 2PL and 3PL models provide an extra item parameter to account for differences among items, and the 3PL model has a parameter that can be used to model guessing behavior among low-ability students.

Modeling item response functions as accurately as possible by using 2PL and 3PL models also reduces errors due to model mis-specification. The error is apparent when the model cannot exactly reproduce or predict the data using the estimated parameters. The difference between the observed data and those generated by the model is directly proportional to the degree of model mis-specification. Current psychometric convention does not

13.3 Item Parameter Estimation

Exhibit 13.2 TIMSS 1999 Grade 8 Science Assessment Example Item Response Function—Polytomous Item

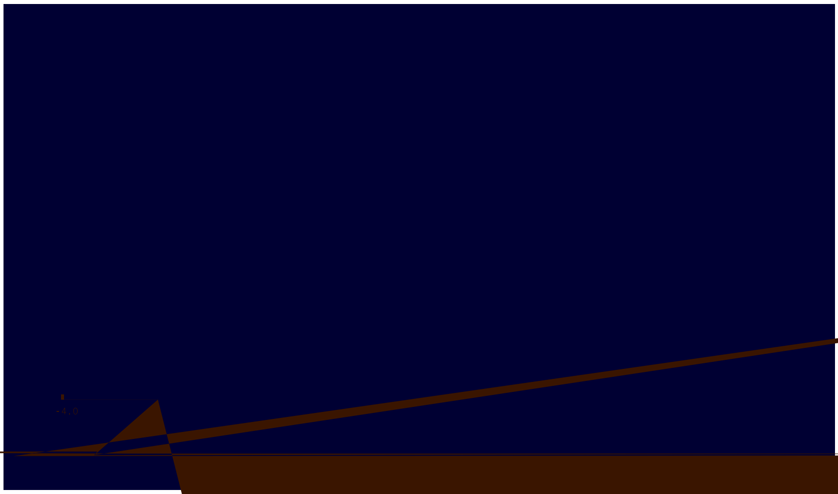


Exhibit 13.1 sh [illegible] response functi [illegible] s represents the p [illegible] nts the probability of [illegible] theoretical curve based [illegible] e centers of the small ci [illegible] tions correct. The size of [illegible] sum of the posteriors at each po [illegible] r all of those who received the item [illegible] ber of respondents contributing to [illegible] pirical proportion correct. Exhibit 13.2 sh [illegible] rical and theoretical item response function [illegible] tem. Again, the horizontal axis represents th [illegible] le, but the vertical axis represents the probabilit [illegible] esponse fall in a given score category. The interp [illegible] the small circles is the same as in Exhibit 13.1. For item [illegible] the model fits the data well, the empirical and theoretic [illegible] ves are close together.

13.4 Scaling Mathematics and Science Domains and Content Areas

In order to estimate student proficiency scores for the subject domains of mathematics and science, all items in each subject domain were calibrated together. This approach was chosen because it produced the best summary of student proficiency across the whole domain for each subject. Treating the entire mathematics or science item pool as a single domain maximizes the number of items per respondent, and the greatest amount of information possible is used to describe the proficiency distribution. This was found to be a more reliable way to compare proficiency across countries than to make a scale for each content area, such as algebra, geometry, etc., and then form a composite measure of mathematics by combining the content area scales.

A disadvantage of this approach is that differences in content scales may be underemphasized as they tend to regress toward the aggregated scale. Therefore, to enable comparisons of student proficiency on content scales, TIMSS provided separate scale scores of each content area in mathematics and science. If each content area is treated separately when estimating item parameters, differential profiles of content area proficiency can be examined, both across countries and across subpopulations within a country.

13.4.1 Omitted and Not-Reached Responses.

Apart from data that by design were not administered to a student, missing data could also occur when 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. In TIMSS 1999, 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 as not having been reached by students were treated as if they had not been administered. This approach was optimal for parameter estimation. However, since the time allotment for the TIMSS 1999 tests was generous, and enough for even marginally able respondents to complete the items, not reached items were considered to have incorrect responses when student proficiency scores were generated.

13.4.2 Proficiency Estimation Using Plausible Values

Most cognitive skills testing is concerned with accurately assessing the performance of individual respondents for the purposes of diagnosis, selection, or placement. Regardless of the measurement model used, 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 individual θs.

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 sampled student *j*. These values are referred to as imputations in the sampling literature, and as plausible values in large-scale surveys such as NAEP, NALS, and IALLS.[3] The value of θ for any respondent that would enter into the computation of *t* is thus replaced b52.

IALLSnot as ouldomi/F8 1 36D(is 7 T3 Tw(bles caledosTn density of proficiencies of the scales, conditional on the observed value $\underline{y}_j$ 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.

13.4.3 Conditioning

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

$$\theta = \Gamma'\underline{y}^c + \varepsilon \tag{9}$$

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

In order to be strictly correct for all functions Γ of $\underline{\theta}$, it is necessary that $p(\underline{\theta}|\underline{y})$ be correctly specified for all background variables in the survey. In Benchmarking, 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 $\underline{\theta}$ for these variables is nearly optimal. Estimates of functions Γ involving background variables not conditioned in this manner are subject to estimation error due to mis-specification. The nature of these errors is discussed in detail in Mislevy (1991).

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, $\underline{\theta}$, and variance Σ, of the posterior distribution in equation (7). For the multiple content area scales of TIMSS 1999, the computer program CGROUP (Thomas, 1993) was used. The program implemented a method to compute the moments using higher-order asymptotic corrections to a normal approximation. Case weights were employed in this step.

13.4.4 Generating Proficiency Scores

After completing the EM algorithm, the plausible values are drawn in a three-step process from the joint distribution of the values of Γ for all sampled. First, a value of Γ is drawn from a normal approximation to that fixes Σ at the value (Thomas, 1993). Second, conditional on the generated value of Γ (and the fixed value of Σ=), the mean, , and variance, Σ_j^p, of the posterior distribution in equation (2) 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 and variance Σ_j^p. These three steps are r thlQcw1:/pse imeas,(prducsing/pseiompuuatioes of)-698.4((for)]TJT*

5. An estimate of the variance of *T* is the sum of two components: an estimate of $Var(T_u)$

duces slightly different means and standard deviations than in the original TIMSS 1995 results. Comparison of the original and rescaled 1995 proficiency scores is not appropriate because of this difference in the scale metric.

13.5.2 Scaling the 1999 Data and Linking to the 1995 Data

Since the achievement item pools used in 1995 and 1999 had about one-third of the items in common, the scaling of the 1999 data was designed to place both data sets on a common IRT scale. Although the common items administered in 1995 and 1999 formed the basis of the linkage, all of the items used in each data collection were included in the scaling since this increases the information for proficiency estimation and reduces measurement error.

The linking of the 1995 and 1999 scales was done at the mathematics and science domain levels only, since there were not enough common items to enable reliable linking within each content area.

13.5.3 Creating IRT Scales for Mathematics and Science Content Areas for 1995 and 1999 Data

IRT scales were also developed for each of the content areas in mathematics and science for both 1995 and 1999. Because there were few items common to the two assessments, and because of some differences in their composition, the two scales were not linked, but rather each was established independently.

For TIMSS 1999, the international mean for mathematics was 487 and the international mean for science was 488. The international mean for each content area was set to be equal to the subject area international mean.

13.5.4 Proficiency Scores for Benchmarking Students

Benchmarking plausible values for each student were generated using item statistics obtained from the international study. Consequently, the benchmarking plausible values are directly comparable to those obtained in the international study. For each student, five plausible values were produced for each of the five mathematics content areas (fractions and number sense; measurement; data representation, analysis, and probability; geometry; and algebra), as well as for mathematics overall. Similarly, plausible values were generated for each student in each of the six science content areas (earth science; life science; physics; chemistry; scientific inquiry; and the nature of science) and science overall.

13.6 Summary

IRT was used to model the TIMSS achievement data. TIMSS used two- and three-parameter IRT models, and plausible-value technology to reanalyze the 1995 achievement data and analyze the 1999 achievement data. Plausible-value methodology was used to generate proficiency estimates for each subject and each content area.

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.), *TIMSS technical report volume II: Implementation and analysis* (pp. 111-145). Chestnut Hill, MA: Boston College.

Andersen, E.B. (1980). Comparing latent distributions. *Psychometrika, 45*, 121-134.

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*(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, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley Publishing.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

Engelen, R.J.H. (1987). *Semiparametric estimation in the Rasch model*. (Department of Education Research Report No. 87-1). Twente, The Netherlands: University of Twente.

Gonzalez, E.J. (1997). Reporting student achievement in mathematics and science. In M. O. Martin & D. L. Kelly (Eds.), *TIMSS technical report volume II: Implementation and analysis* (pp. 147-174). Chestnut Hill, MA: Boston College.

Hoijtink, H. (1991). *Estimating the parameters of linear models with a latent dependent variable by nonparametric maximum likelihood*. (Research Bulletin No. HB-91-1040-EX). Groningen, The Netherlands: University of Groningen, Psychological Institute.

Johnson, E.G., & Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics, 17*(2), 175-190.

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

Lindsey, B., Clogg, C.C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*, 96-107.

Little, R.J.A., & Rubin, D.B. (1983). On jointly estimating parameters and missing data. *American Statistician, 37*, 218-220.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical analysis with missing dat*a. New York, NY: John Wiley and Sons.

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

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

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*(2), 133-161.

Mislevy, R.J., & Bock, R.D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* (Computer program). Morresville, IN: Scientific Software.

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. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report* (pp. 293-360). Princeton, NJ: Educational Testing Service.

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

Muraki, E., & Bock, R.D. (1991). *PARSCALE: Parameter scaling of rating data*. Chicago, IL: Scientific Software, Inc.

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

Rubin, D.B. (1991). EM and beyond. *Psychometrika*, *56*, 241-254.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, *82*, 528-550.

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