9 Reporting Achievement in Mathematics and Science Content Areas

Albert E. Beaton
Eugenio J. Gonzalez
Boston College

9.1 ADAPTING AVERAGE PROPORTION-CORRECT TECHNOLOGY FOR TIMSS

Although item response theory (IRT) methods were used to scale the student achievement data for purposes of international reporting, TIMSS also made use of an approach whereby the proportion of items answered correctly by the students in a country was averaged over the set of items in a subject matter content area. This "average-proportion-correct technology" was used for reporting performance in each of the 11 content areas of mathematics and science that were assessed at the seventh and eighth grades, and each of the 10 content areas that were assessed at the third and fourth grades. The content scales assessed in each subject, at each grade level, are presented in Table 9.1. Average proportion-correct technology was also used for the Test Curriculum Matching Analyses (TCMA) described in Chapter 10. This approach allows the averaging across items, even though the items are located in different assessment booklets and individual students do not respond to all of the items being averaged. Using this technology, it is also possible to obtain standard errors for the proportion correct with a slight modification of the jackknife repeated replicate (JRR) variance estimation procedures outlined in Chapter 5.

Table 9.1 Mathematics and Science Content Areas

Third and Fourth Grades (Population 1)

Mathematics	Science
• Whole Numbers	• Earth Science
• Fractions and Proportionality	• Life Science
• Measurement, Estimation, and Number Sense	• Physical Science
• Data Representation, Analysis, and Probability	• Environmental Issues and the Nature of Science
• Geometry	
• Patterns, Relations, and Functions	

Seventh and Eighth Grades (Population 2)

Mathematics	Science
• Fractions and Number Sense	• Earth Science
• Geometry	• Life Science
• Algebra	• Physics
• Data Representation, Analysis, and Probability	• Chemistry
• Measurement	• Environmental Issues and the Nature of Science
• Proportionality	

Unlike the TIMSS IRT scaling, the average proportion-correct approach does not provide scores or plausible values for individual students, and is also sensitive to ceiling effects on sets of items, in particular when a subpopulation of interest responds correctly to most or all of the items in a set. However, the average proportion-correct approach was used in TIMSS for reporting student performance in subject matter content areas and for the TCMA analyses in preference to IRT scaling because of cost considerations, and because of the extra time the more complex scaling approach would have required.

Adapting the average proportion-correct technology for TIMSS posed two particular problems. The first was that some of the TIMSS items had graded responses, that is, the students were assigned a score ranging from 0 to 3 points depending on the item and the degree of correctness of their responses to the item. When an item response can have only two values, 0 for incorrect and 1 for correct, the average score on the item for a sample of students is also the proportion correct. However, this does not hold for an item where responses can score more than 1. For such items, it was necessary to find a way to use the proportion correct to represent the responses.

The second problem was that occasionally an item was found to be unusable for some countries. The item review process (see Chapter 6) revealed that from time to time an item for a country was misprinted, mistranslated, or missing by mistake from the booklet, or had other problems that prevented them from being comparable with the items administered in other countries. Such items were deleted for the country concerned; however, they could affect the overall proportion correct for a specific country if, for example, a country happened to have mistranslated the most difficult item in a content area. While such missing items are handled readily by IRT methods, they cause difficulties for the average proportion-correct approach. The items deleted at each population are documented in Chapter 6.

9.1.1 Treating Graded Response Items

A simple way to handle graded responses would be to compute the average score on each of the items in a particular area and then add up these averages to obtain the average score on the scale. The statistic computed for each country would then be the sum of its averages for the items involved in the area. The average for a binary (right/wrong) item in this situation would be its proportion correct, and for a graded response the average score on the item. However, an average computed this way would have an upper bound equal to the total number of score points possible divided by the total number of items. If any of the items were graded-response items with maximum scores greater than 1, the upper bound for the average would be greater than 1, and the average would not be interpretable as a proportion-correct.

By transforming the graded responses into a series of binary items, TIMSS was able to use the proportion-correct technology without losing information, and, in fact ,some additional information was gained. Consider that an item may be assigned a score of 0, 1, 2, or 3. We can code a student's response as if it were three variables ($v_{j,1}, v_{j,2,} v_{j,3}$) as follows:

$v_{j,1}$ equals 1 if the student receives a 1, 2, or 3, and 0 otherwise;

$v_{j,2}$ equals 1 if the student receives a 2 or 3, and 0 otherwise; and

$v_{j,3}$ equals 1 if the student receives a 3, and 0 otherwise.

We can then call $p_{vj,1}$, $p_{vj,2}$, *and* $p_{vj,3}$ the proportions of students who received a 1 on $v_{j,1}$, $v_{j,2}$, and $v_{j,3}$ respectively. Note in particular that $p_{vj,1} \geq p_{vj'2} \geq p_{vj,3}$. The average value of item v_j can then be computed from the proportions of students who receive a score of 1 on $v_{j,1}$, $v_{j,2,}$ $v_{j,3}$, that is,

$$\bar{v}_j = \sum_i p_{vj,i}$$

We can also compute the average proportion correct on these three items (p_j) as

$$p_j = \frac{1}{I} \sum_i p_{vj,i}$$

where I is the maximum score points on the item.

As a numerical example, let us assume the frequency distribution shown in Table 9.2 for a graded-response item administered to 1,000 students within a country.

Table 9.2 Sample Frequency of Responses to Item v_j

Score	Frequency
0	200
1	300
2	400
3	100
Total	1000

The score on this item is 1.4, computed as follows:

$$\bar{v}_j = \frac{200 * 0 + 300 * 1 + 400 * 2 + 100 * 3}{1000} = \frac{1400}{1000} = 1.4$$

The three proportions for this item would then be

$p_{vj,1} = 0.80$

$p_{vj,2} = 0.50$

$p_{vj,3} = 0.10$

from which the same average can be computed by

$$\bar{v}_j = \sum_i p_{vj,\,i} = 0.80 + 0.50 + 0.10 = 1.4$$

Using this method of coding we can treat the graded-response items as binary items and still compute the average score for an item allowing the full range of values. If all graded-response items are coded in this way, then the average proportion correct over any set of items will be proportionally the same as if the averages of graded items were mixed with the percentages correct of binary items. Yet we have gained the advantage of working only with proportions.

We note that the three proportions in our example ($p_{vj,1}$, $p_{vj,2}$, and $p_{vj,3}$) contain some information that the average graded response does not. From these proportions, we can see what proportion of students in a country responded at each score level for that item. When this procedure is used, the number of items is then effectively increased from *j* to *j′*, where *j′* is equal to the total number of possible scores points on the set of items.

9.1.2 Missing Proportions Correct

A second problem with the reporting of average proportion correct was what to do on those rare occasions when an item has to be deleted for a country. It is important that the deletion should neither penalize nor benefit the country. Where an item was found to be unusable for a country, that item could be omitted from the analysis for all countries without any threat to fairness, but since different items exhibited problems in different countries, this would reduce the total item pool unacceptably, and would necessitate discarding perfectly good data for the unaffected countries. On the other hand, if the item is deleted only for the affected country, there is the possibility of unduly influencing the country's overall score. To minimize the effect of deleted items on overall average proportion correct, TIMSS derived a method of estimating the proportion of students in the country that would have performed successfully on the items if they had been included. To achieve this, TIMSS used the information on how the country performed on the remaining items, and how the other countries performed on the item in question. Transforming all items into binary variables as described in the preceding section greatly facilitated the implementation of this procedure.

Note that this approach was used when average proportions correct were used for cross-national comparisons. The IRT scaling did not require this procedure since one of the advantages of IRT scaling is its capacity to deal with missing items.

9.1.3 Computational Method

The TIMSS approach was as follows: Let us assume that we want to estimate the average proportion correct over a set of items for a set of countries but that one country, country *k*, has mistranslated item *j′* and therefore the proportion correct for country *k* on item *j′* cannot be known from the available data.[1] Different countries may have dif-

[1] We will use the notation *j′* for an item to signify the dichotomized version of the item, as described in the section on *Treating Graded Response Items.*

ferent unusable items and thus different missing proportions. The TIMSS procedure may be used as long as there is at least one known proportion for each country and for each item, although of course it works best when there are just a few missing items.

The TIMSS approach begins by filling in the missing values using the model

$$p_{kj'} = p_{k0} + p_{0j'} - p_{00}$$

where $p_{kj'}$ is the estimated proportion correct of country k on its unusable item j', p_{k0} is the average proportion correct of country k on all of its usable items, $p_{0j'}$ is the average proportion correct of all other countries on item j', and p_{00} is the average proportion correct for all available items over all countries. Imputation under this model implies that there is no interaction between the proportion correct on the imputed item and the countries.

The above model was improved in two ways. First, filling in an estimated value of $p_{kj'}$ affects the values of p_{k0}, $p_{0j'}$, and p_{00}, so the method should be iterative, making successive estimates until all values stabilize. Second, proportion correct is not a good statistic for an additive model such as is specified above; in fact, unless the proportions are transformed to an additive metric, estimated proportions of greater that 1 or less than 0 are possible. The use of the logit transformation of the proportions avoids this problem.

The logit transformation used to transform the percents correct into an additive scale is

$$z_{kj'} = Logit(p_{kj'}) = \ln\left(\frac{p_{kj'}}{1-p_{kj'}}\right)$$

Using this equation transforms a proportion correct for an item ($p_{kj'}$) to a logit value ($z_{kj'}$) that may range from minus to plus infinity. The logit for p = .50 is zero. The logit for 0 is minus infinity and for 1 is plus infinity, and so values of 0 and 1 are not usable. In the unusual case when there is a value of 1.0 or 0.0 for a proportion correct for an item, 0.9999 is substituted for 1.0, and 0.0001 is substituted for 0.0. This logit transformation permits simple and appropriate arithmetic calculations on proportions.

If we now define a matrix of proportions $P_{kj'}$ where k is the number of countries and j' is the number of items, and some of the elements of $P_{kj'}$ are missing, the method used to estimate the missing proportion correct works as described below.

Step 1: The matrix with logit scores $Z_{kj'}$ is produced from the usable elements of the matrix $P_{kj'}$ by the transformation of the elements in $P_{kj'}$ into logit scores as defined above. The elements $z_{kj'}$ when item j' is deemed unusable in country k are left blank in this $Z_{kj'}$ matrix. The matrix $Z_{kj'}$ also has a "zeroth" row and column. The elements in z_{k0} contain the average of the elements on the kth row of the $Z_{kj'}$ matrix. These are the country averages across the usable items. The elements in $z_{0j'}$ contain the average of the elements of the j'th column of the $Z_{kj'}$

matrix. These are the item averages across all countries. The element z_{00} contains the overall average for the elements in vector $z_{0j'}$ and z_{k0}. In the initial matrix $Z_{kj'}$, the averages are defined over the usable $z_{kj'}$ elements and the missing values are not used.

Step 2: The first estimation for the logits of the missing $z_{kj'}$ values is then given by the formula

$$z'_{kj'} = z_{k0} + z_{0j'} + z_{00}$$

Step 3: At this point a new matrix $Z'_{kj'}$ is created where each of the $Z'_{kj'}$ elements are the same as those in $Z_{kj'}$, but where the missing $z_{kj'}$ elements are replaced with the newly estimated $z'_{kj'}$.

Step 4: New averages are computed for the vectors z'_{k0}, $z'_{0j'}$, and z'_{00} with the elements of the newly created $Z'_{kj'}$, matrix. These averages can now be computed over all available values in $Z'_{kj'}$ which is now a complete matrix with no missing elements.

Step 5: New estimates for the missing elements in the $Z_{kj'}$ matrix are then computed as

$$z'_{kj'} = z'_{k0} + z'_{0j'} - z'_{00}$$

where $z'_{kj'}$, z'_{k0}, $z'_{0j'}$, and z'_{00} are the values obtained from the $Z'_{kj'}$ matrix on the succeeding iterations.

Steps 3 through 5 above are repeated until a stable solution has been reached. The criterion for convergence is that none of the elements in the $z'_{kj'}$ vectors changes more than .001 from one iteration to the next.

Once a stabilized $Z'_{kj'}$ matrix is obtained, the estimates for the missing elements in $P_{kj'}$ are obtained by creating the matrix $P'_{kj'}$ using the inverse logit transformation

$$p'_{kj'} = \frac{\exp(z'_{kj'})}{1 + \exp(z'_{kj'})}$$

and applying it to each of the elements of the $Z'_{kj'}$ matrix.

The average percent correct on a scale for each country is then obtained by averaging the rows of the $P'_{kj'}$ matrix.

In doing this, notice that the average proportions correct for countries that have all usable data, or for items that were usable for all countries, remain unchanged. In TIMSS the missing proportion-correct values for the unusable items were imputed using only the information for the content area to which the item was assigned. These imputed percent-correct values were then used in the computation of the average percent correct at the content area level and overall for each subject.

9.1.4 Computing Standard Errors

Once the estimates for the missing elements of the $P_{kj'}$ matrix are obtained, the average percent correct for the items of a scale in a country can be computed. These average percents correct are the elements of the vector P_{k0} from the matrix $P_{kj'}$. Each of the $p_{kj'}$ values was computed using the overall sampling weight.

In order to obtain variance estimates for the average percent corrects, it is possible to make use of the replicate weights approach used by the jackknife algorithm to estimate the sampling variability of the data used to fill in the blanks in the $P_{kj'}$ matrix. It is important to keep in mind that the estimated elements of the matrix $P_{kj'}$ are computed using the elements in the vectors $P_{0j'}$ *and* P_{k0} and therefore are subject to variability as repeated replicate samples are drawn from each country. To implement the jackknife repeated replication (JRR) procedure in this case, the sampling zones across the countries are randomly sorted, and information from different zones by country is used to obtain each of the 75 estimates from which the sampling errors are computed. When the sampling zones within a country are sorted they are renumbered and treated as an international zone or international replicate.

The JRR procedure was implemented as follows. TIMSS assigned the schools within each country in pairs to one of up to H jackknife zones, where H is equal to 75. The 75 sampling zones were used to create 75 "pseudo-replicates" of the original sample. Each of the pseudo-replicates consists of a copy of the original data, except that in one of the sampling zones (a different one each time) one school of the pair of schools, chosen at random, is omitted, and the weights for the other member of the pair are doubled. In computing a jackknife estimate of the sampling variability of a statistic such as a mean or a proportion, the statistic is computed once for the data in the original sample, and once again for each of the pseudo-replicate samples. The variation between the original sample estimate and the estimates from each of the replicate samples is the jackknife estimate of the sampling error of the statistic.

Doubling or omitting the weights of the selected school within each sampling zone is accomplished effectively in computational terms by the creation of replicate weights. The replicate-weight approach requires the temporary creation of a new set of weights for each replicate sample. To create the replicate weights for the first replicate sample, one of the pair of schools in the first sampling zone is chosen at random to have its weights doubled, while the other member of the pair has its weights set to zero to compensate. The weights of the schools in all other sampling zones are left unchanged. The replicate weights for the second replicate sample are created in a similar manner. Again, the weights for the schools in all other zones are unchanged from the original weights. This procedure is repeated for all 75 sampling zones, resulting in 75 sets of replicate weights (W_h) for each country.

Using these 75 replicate weights we then compute for each country k a matrix $T^{k}_{h'j'}$ where the row elements across the hth row are the proportion correct of each of the j' items in the scale computed using the hth replicate weight, and the elements down each j'th column are the proportion correct for the j'th item computed using each of the hth replicate weights. The row vectors of this matrix for the kth country are then randomly sorted so that the order of the replicate weights used to compute each row now varies by country. The rows of each of these matrices are now renumbered using the indexing variable h', and the newly sorted matrix is called $T^{k}_{h'j'}$.

At this point we then proceed to form each of the 75 $P^{h'}_{kj'}$ matrices by taking the h'^{th} row from the $T^{k}_{h'j'}$ matrices. After the estimation of the missing elements of each of the $P^{h'}_{kj'}$ matrices takes place, the resulting 75 $P'^{h'}_{k0}$ vectors will contain the H replicates for each of the k countries in the sample. At this point the standard method for estimating the sampling variance is used by applying the following equation for each country:

$$jse_{P'_{k0}} = \sqrt{\sum_{h'}(p'_{k0} - p'^{h'}_{k0})}$$

9.2 PROFILES OF RELATIVE PERFORMANCE BY CONTENT AREAS

In addition to performance on mathematics and science overall, it was of interest to see how countries performed on the content areas within each subject relative to their performance on the subject overall. If the results for all countries are summarized in a table of average percents correct organized by country and by content area, then differences in relative performance across content areas for a country may be thought of as a country-by-content area interaction. There were six content areas in mathematics at each population, and four science content areas at Population 1 and five at Population 2, that were used in this analysis. The relative performance for the countries on the content areas was examined separately for each subject.

Suppose now that we have computed the vector of average percent corrects (P'_{k0}) for each of the content areas on the test using the procedures described earlier, and that we join each of these column vectors to form a new matrix called R_{ks} where a row contains the average percent correct for country k on scale s for a specific subject. This R_{ks} matrix has also a "zeroth" row and column. The elements in r_{k0} contain the average of the elements on the kth row of the R_{ks} matrix. These are the country averages across the content areas. The elements in r_{0s} contain the average of the elements of the sth column of the R_{ks} matrix. These are the content area averages across all countries. The element r_{00} contains the overall average for the elements in vector r_{0j} or r_{k0}. Based on this information we can then construct the matrix R'_{s} in which the elements are computed as

$$r'_{ks} = r_{ks} + r_{00} - r_{0s} - r_{k0}$$

Each of these elements can be considered as the interaction between the performance of country k on content area s. A value of zero for an element r'_{s} indicates a level of performance for country k on content area s that would be expected given its performance on other content areas and its performance relative to other countries on that content area. A negative value for an element r'_{s} indicates a performance for country k on content area s lower than would be expected on the basis of the country's overall performance. A positive value for an element r'_{s} indicates a performance for country k on content area s better than expected.

Although we can compute the values for the country by content area interaction, this value is of little interest unless we can determine whether it is significantly different from zero. To do this we need to compute the corresponding standard error for each of the r'_s elements and perform a test of significance, taking into account the multiple comparisons by using the Dunn-Bonferroni procedure (see Chapter 8).

To compute the JRR standard error, suppose that we have computed the vector of average percents correct for each of the international replicates $P'^{h'}_{k0}$ for each of the content areas on the test using the procedures described in the previous chapter, and that we join each of these column vectors to form a new set of matrices each called R^h_{ks} where a row contains the average percent correct for country *k* on content area *s* for a specific subject, for the *h*th international set of replicates. Each of these R^h_{ks} matrices has also a "zeroth" row and column. The elements in r^h_{k0} contain the average of the elements on the *k*th row of the R^h_{ks} matrix. These are the country averages across the content areas. The elements in r^h_{0s} contain the average of the elements of the *s*th column of the matrix. These are the content area averages across all countries. The element r^h_{00} contains the overall average for the elements in vector r^h_{0j} or r^h_{k0}. Based on this information we can then construct the set of matrices R'^h_{ks} in which the elements are computed as

$$r'^h_{ks} = r^h_{ks} + r^h_{00} - r^h_{0s} - r^h_{k0}$$

The JRR standard error is given by the formula

$$jse_{r_{ks}} = \sqrt{\sum_h (r'_{ks} - r'^h_{ks})^2}$$

A relative performance was considered significantly different from the expected if the 95 percent confidence interval built around it did not include zero. The confidence interval for each of the r'_{ks} elements was computed by adding and subtracting to the r'_{ks} element its jackknifed standard error multiplied by the critical value for the number of comparisons.

The critical values were determined by adjusting the critical value for a two-tailed test, at the alpha 0.05 level of significance for multiple comparisons according the Dunn-Bonferroni procedure. Since the number of scales varied by subject, and the number of countries varied by grade, eight different critical values were computed. Table 9.3 summarizes the number of comparisons performed by subject at each grade level.

Table 9.3 Number of Comparisons and Critical Values Used for the Test of Significance of the Relative Performance Within Country

Subject	Grade	Countries	Scales	Comparisons	Critical Value
Science	8th	41	5	205	3.6683
Science	7th	39	5	195	3.6554
Mathematics	8th	41	6	246	3.7148
Mathematics	7th	39	6	234	3.7020
Science	3rd	24	4	96	3.4698
Science	4th	26	4	104	3.4913
Mathematics	3rd	24	6	144	3.5774
Mathematics	4th	26	6	156	3.5984

9.3 PERCENT CORRECT FOR INDIVIDUAL ITEMS

To portray student achievement as fully as possible, the TIMSS international reports present many examples of the items used in the TIMSS tests, together with the percentage of students in each country responding correctly to the item. For multiple-choice items this was the weighted percentage of students that answered the item correctly. This percentage was based on the total number of students that were administered the items. Omitted and not-reached items were treated as incorrect. For free-response items with more than one score level the percent correct for these example items was computed as the weighted percentage of students that achieved the highest score possible on the item.

When the percent correct for example items were computed, student responses were classified in the following way. For multiple-choice items, the responses to item *j* were classified as correct *(C_j)* when the correct option for an item was selected, incorrect *(W_j)* when the incorrect option for an item was selected, invalid *(I_j)* when two or more choices were made on the same question, not reached *(R_j)* when it was determined that the student stopped working on the test before reaching the question, and not administered *(A_j)* when the question was not included in the student's booklet or the question was mistranslated or misprinted. For free-response items student responses to item *j* were classified as correct *(C_j)* when the maximum number of points was obtained on the question, incorrect *(W_j)* when the wrong answer or an answer not worth all the points in the question was given, invalid *(N_j)* when, although something was written in the answer sheet, what was written was not legible or interpretable, not reached *(R_j)* when it was determined that the student stopped working on the test before reaching the question, and not administered *(A_j)* when the question was not included in the student's booklet or the question was mistranslated or misprinted. The percent correct for an item *(P_j)* was computed as

$$P_j = \frac{c_j}{c_j + w_j + i_j + r_j + n_j}$$

where c_j, w_j, i_j, r_j and n_j are the weighted counts of the correct, wrong, invalid, not reached, and not interpretable responses to item *j*, respectively.

Note that although the not-reached responses were treated as missing for the purpose of estimating the item parameters in the international IRT scaling, they were considered to be wrong answers for an individual when percents correct for an item were computed.

9.4 REPORTING GENDER DIFFERENCES BY CONTENT AREAS

Differences between the performance of boys and girls in the subject matter content areas were also examined using the average percent-correct approach. The performance difference was determined to be significant if the standardized difference between the average percent correct for boys and girls within a country exceeded the critical value, corrected using the Dunn-Bonferroni procedure for multiple comparisons.

The standardized difference between the average percent corrects (t_k) was computed as

$$t_k = \frac{\bar{p}_{kb} - \bar{p}_{kg}}{\sqrt{pse_{kb}^2 + pse_{kg}^2}}$$

where $\bar{p}_{kp}$ and $\bar{p}_{kg}$ are the average percents correct within the content area for boys and girls, respectively, within country *k*, and pse_{kb} and pse_{kg} are the standard errors of the average percents correct for boys and girls, respectively, within country *k* computed using the jackknife procedure for estimating sampling error. The critical value for the seventh grade was 3.22005, and for the eighth grade was 3.23431. These critical values are corrected using the Dunn-Bonferroni procedure for multiple comparisons. At the seventh grade, the critical value was corrected for 39 comparisons, and at the eighth grade for 41 comparisons. The critical value used for the third and fourth grade tests of significance was 1.960. This critical value was not adjusted for multiple comparisons.