DEVELOPING INDICATORS OF EDUCATIONAL CONTEXTS IN TIMSS

Alka Arora, International Study Center ,Boston College, USA
María José Ramírez, International Study Center, Boston Colege, USA

Abstract

TIMSS collects a vast amount of information aimed to provide a context to better understand education systems around the globe. This paper discusses how this background information can be reported using three real examples from the 1999 assessment. These examples illustrate the main concepts and steps involved in reporting background data, while presenting methods and statistical techniques that can be used for developing indices. Important criteria to evaluate the overall quality of the measures are introduced, paying special attention to their validity and reliability. The advantages of using indices are discussed both from a conceptual and a measurement perspective.

Keywords: Comparative education, Educational indicators, Large-scale assessment.

BACKGROUND

The purpose of the Trends in Mathematics and Science Study (TIMSS) is to provide a base from which policy makers, curriculum specialists, and researchers can better understand the performance of their educational systems. With this aim, TIMSS collects data on hundreds of contextual variables from nationally representative samples of students, their science and mathematics teachers, and their schools. Background data is collected using four types of questionnaires. The School Questionnaire asks about the community in which the school is situated, its staff and student body, curricular and instructional resources, among others. The Teacher Questionnaire mainly gathers information about teachers' background and preparation to teach, and their instructional practices. The Student Questionnaire focuses on students' attitudes and perceptions about science and mathematics, the educational resources they have at home, and their activities during and after classes. The Curriculum Questionnaires collects information about the structure and organization of the curriculum, and the topics intended to be covered up to the eighth grade.

Once the data are collected, one of the major challenges for TIMSS is reporting this vast array of information in a thorough and meaningful way. There is a need to focus on relevant educational contexts, inputs, and processes, while avoiding overburdening the audiences with unmanageable amounts of information. Finally, time and resource constrains need to be considered. Because of the nature of the TIMSS project, the comparability of the data across different contexts must be ensured. This has to be done by developing technically fair measures of educational indicators. At the same time, these measures must be easy to understand and to communicate to a broad audience.

The purpose of this paper is to explain how the TIMSS data can be further analyzed to decide what information to report and how to report it. Three examples of educational indicators from the TIMSS 1999 international report are presented. In the first example, data from the Student Questionnaire was directly reported, with no or minimal modifications. In the other two examples, data from different source items were combined to form indices. In the second example, an index with an underlying scale was computed by averaging the score points associated with the response categories of its component items. This scale was further divided in qualitative levels—high, medium, low—using cutoff points. In the third example, an index was created based on combinations of item responses, each combination being directly assigned to an index level.

This paper emphasizes the benefit of deriving indices to report the TIMSS data: they provide more valid and reliable measures of complex constructs and allow for effective data summarization. While

developing the three examples, key test theory, psychometric and statistical concepts are addressed, and important criteria to evaluate the overall quality of the measures are discussed. Construct definition and content validity, the identification of construct indicators, the creation of scales, the empirical validations of these scales, the use of measures of association to confirm expected relationships, are the kind of topics addressed while guiding the reader through the process of deriving simple and complex measures.

Explaining the methodology used in reporting TIMSS data is important to help better understand the meaning of these indicators. It is also important for educational researchers and students interested in developing their own indices according to their local contexts, interests, and research questions. The conclusions describe the possibilities and limitations of international measures of educational contexts.

METHODS FOR REPORTING CONTEXTUAL DATA

Direct Reporting Method

The simplest and straightforward method to report TIMSS background data is by using the same response categories provided in the questionnaires, or by introducing only slight modifications in them. This method is appropriate when measuring something directly observable or a somewhat simple phenomenon (e.g., age, school size). In TIMSS 1999, the direct reporting method was used to report students' educational expectations. Item 8 in the students' questionnaire was used to collect this information (Figure 1).

Figure 1: Students' Expectation for Finishing School

8. How far in school do you expect to go?

*Circle **one** letter, A, B, C, D, E, or F.*

<some secondary school>	A
<finish secondary school>	B
<some vocational/technical education after secondary school>	C
<some university>	D
<finish university>	E
I don't know	F

TIMSS requires the participating countries to adapt the international version of the questionnaires to fit the unique characteristics of their educational systems. In the figure, the text in brackets had to be translated and replaced by a *national option* equivalent to the international version of the questions. TIMSS sets up *quality control* procedures to ensure that national adaptations are internationally comparable. In some exceptional cases, countries are allowed to make *national adaptations* that deviate from the international version. For instance, while Cyprus changed the first option "some secondary school" to "finished lower secondary", Belgium Flemish preferred not to include it (Gonzalez and Miles, 2001). Countries have to document any deviation from the international version

and must indicate how their new response categories compare to the international ones. This information is then used to recode the data before putting it in the international dataset.

In some cases, not enough students select a given response option; or, after a few revisions, it is apparent that it would be better to present together information that was asked separately. In these situations, it is a good idea to collapse some of the original response categories. This was the case for the options *C* "some vocational/technical education after secondary school" and *D* "some university," which were collapsed into one category: "some vocational/technical education or university only."

When data from the questionnaires are directly reported, it is recommended to keep the same option labels that were used in the item. This avoids ambiguities and preserves the meaning of the original question, which is important to secure the comparability of the data across the countries. The order of the options may vary, however, to highlight some kind of information. This was clearly the intention when the percentage of students who finished the university was reported in the first column of Exhibit 4.4 in the 1999 Mathematics international report (Table 1).

Table 1: Students' Expectations for Finishing School and Mathematics Achievement

	Finish University		*Some Voc./Tech. Edu. or Uni. Only*		*Finish Secondary School Only*		*Some Secondary School Only*		*Do Not Know*	
	% of Students	*Avg. Ach.*	*% of Students*	*Avg. Ach.*	*% of Students*	*Avg. Ach.*	*% of Students*	*Avg. Ach.*	*% of Students*	*Avg. Ach.*
AUS	55	554	14	524	17	479	5	460	9	501
CHL	54	428	18	367	19	347	2	—	7	359
ISR	59	492	16	457	11	419	1	—	13	438
PHL	64	374	10	299	9	293	8	293	8	315
TWN	62	624	24	527	2	—	0	—	11	534
AVG	52	517	17	469	15	442	3	390	14	462

Note. AUS = Australia, CHL = Chile, ISR = Israel, PHL = Philippines, TWN = Chinese Taipei, AVG = International average for 38 countries participating in TIMSS 1999.
International report also displays standard errors for percentage of students and average achievement.
As results are rounded to the nearest whole number, some totals may appear inconsistent.
A dash (—) indicates insufficient data to report achievement

While the direct reporting of data from the questionnaires presents the advantage of simplicity, it would be impossible to report the vast amount of information on a one-by-one basis. Some *data reduction* is required to summarize hundreds of contextual variables. Another weakness of the direct reporting method is its inability to measure more complex constructs, like socio-cultural background of the families, or attitudes toward mathematics.

Developing Indices

An index is a *derived variable* that combines data from several items in the TIMSS questionnaires. An index can be used as an indicator of more complex constructs not directly observable, like school climate, and preparation to teach. Because the items making up an index can target different facets of the same object, indices can provide a more global and thorough picture of the phenomenon being studied. Indices are also preferable to single indicators because they can provide more precise or reliable measures of the underlying construct or trait.

TIMSS sometimes classifies the students in three index levels: high, medium, and low. The high level of an index is set so that it corresponds to conditions or activities generally associated with good educational practice or high academic achievement. The levels are also established so that there is a reasonable distribution of students across the three levels. TIMSS strongly emphasizes the *communication* aspect of the indices, which should be easy to understand and interpret by policy makers and school personnel. Efforts are made to create index levels that make sense in the educational practice, and that can be replicated easily.

An essential step in creating indices consists in providing evidence of *criterion related validity.* If the index is measuring the intended trait, and if the trait is supposedly related to an external criterion (e.g., academic performance), then this relationship should be observed in the data. Hence, it is a prerequisite to formulate hypotheses about the relationship between the index levels and achievement scores in order to test them.

The discriminating power of an index indicates the strength of its association with the achievement outcome. Power varies from country to country, and it also varies when combining the information from all the countries together. In international studies like TIMSS, it is a challenge to develop indices that discriminate well both within and across the countries.

Indices allows for *data reduction*. As described in the following sections, one index can condense information from two, five, or dozens of variables (questions), provided that these variables are all conceptually related to the same construct. By doing so, more meaningful and relevant information is made accessible to the TIMSS users.

In the following sections, these and other important measurement concepts are shown with the help of real examples from the TIMSS 1999 Mathematics International Report. The methodology to develop background indices is presented step by step. Data from five countries are used to illustrate the main points and decisions involved in developing indices. These countries—Australia, Chile, Chinese Taipei, Israel and the Philippines—were selected because they allowed illustrating some important points in the scale building process. Therefore their statistics should not be taken as representative of the full sample of countries.

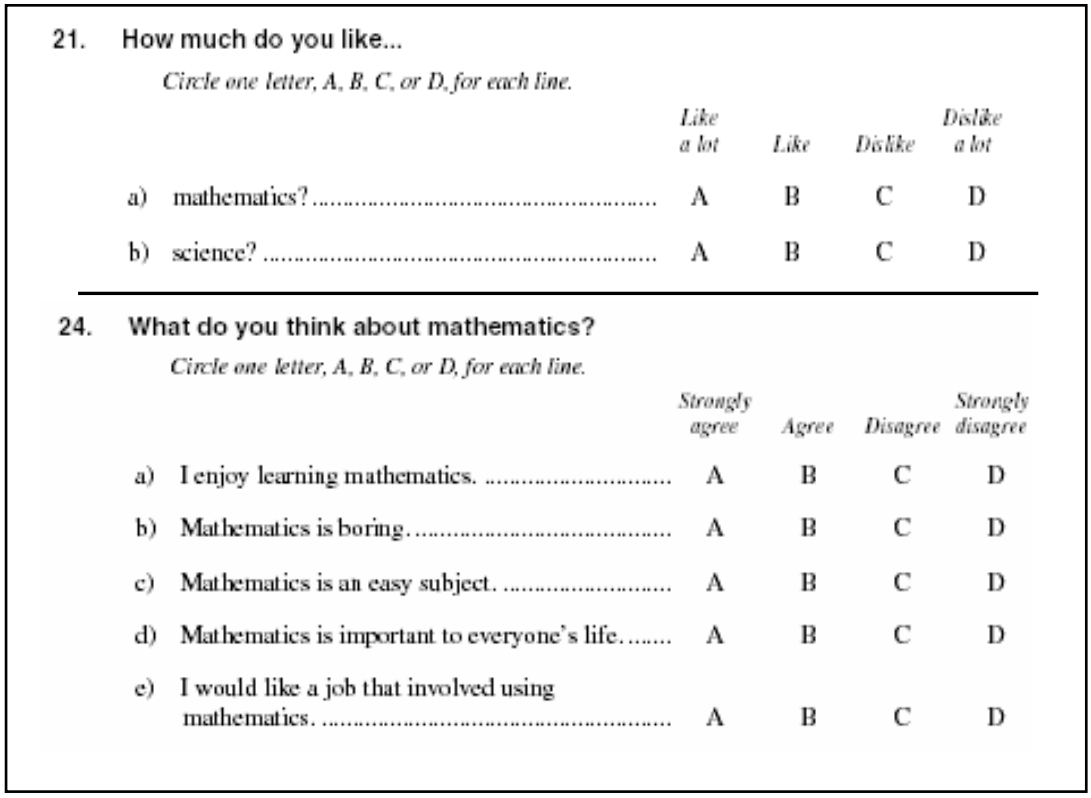
Scale Development Method

The scale development method can be used when several items are conceptually related to the same *construct*, and that construct has an underlying quantitative continuum. This was the case in measuring students' positive attitudes toward mathematics (PATM). Creating a positive attitude in students toward mathematics is an important goal of the curriculum in many countries. Accordingly, it was important to know if the students find the subject enjoyable, place value on the subject, and think it is important for future career aspirations (Mullis et al., 2001). Six main steps were required for developing the PATM index.

Step 1: Identify component questions. The starting point was to identify which items in the TIMSS Student Questionnaire could be used to measure students' attitudes toward mathematics. From a conceptual standpoint, five items were considered appropriate to *operationalize* this construct: items 21a, 24a, 24b, 24d, and 24e (Figure 2). Item 24c was discarded because it focused on the perceived difficulty of mathematics rather than on how the students felt about this subject. Students from countries with very demanding mathematics curricula might perceive mathematics as very difficult, but still like the subject very much.

A simpler way to inform about attitudes toward mathematics could have been to report students' responses to Item 21a only: "How much do you like mathematics?" However the attitude construct is a complex psychological state with many facets: feelings, pressure, behavior, etc. Hence, using *multiple indicators* was more appropriate to capture the complex nature of the phenomenon under study. By so doing, the *content validity* of the scale was improved. Because random errors are cancelled out, using multiple indicators also presented the advantage of increasing the precision or *reliability* of the scale.

Figure 2: *Source Items for Positive Attitudes Toward Mathematics Index*



21. How much do you like...

Circle one letter, A, B, C, or D, for each line.

	Like a lot	*Like*	*Dislike*	*Dislike a lot*
a) mathematics?	A	B	C	D
b) science?	A	B	C	D

24. What do you think about mathematics?

Circle one letter, A, B, C, or D, for each line.

	Strongly agree	*Agree*	*Disagree*	*Strongly disagree*
a) I enjoy learning mathematics.	A	B	C	D
b) Mathematics is boring.	A	B	C	D
c) Mathematics is an easy subject.	A	B	C	D
d) Mathematics is important to everyone's life.	A	B	C	D
e) I would like a job that involved using mathematics.	A	B	C	D

Step 2: Reverse score items. All these questions had a 4-point *Likert scale format*. For item 21a: *like a lot* = 4 points, *like* = 3 points, *dislike* = 2 points, and *dislike a lot* = 1 point; and for items 24a, b, d, e: *strongly agree* = 1 point, *agree* = 2 points, *disagree* = 3 points, and *strongly disagree* = 4 points. A simple way to combine this set of items is by averaging the points associated with each response option. For this to be a meaningful exercise though, all the items must be scaled in the same direction. This means that the 4-point options should always indicate the higher levels of the attribute being measured; that is, they should indicate more positive attitudes toward mathematics. In fact, this was the case for item 24b: a student with highly positive attitudes was expected to *disagree a lot* (4 points) with "mathematics is boring." However, a student with highly positive attitudes was expected to *strongly agree* (1 point) with "I enjoy learning mathematics" (24a). These two items were not scaled in the same direction.

When the items are not scaled in the same direction, it is necessary to *reverse score* for some of them. To build the PATM index, items 24a, 24d, and 24e were reverse scored so that *strongly agree* = 4 points, *agree* = 3 points, *disagree* = 2 points, and *strongly disagree* = 1 point. Now, a student with highly positive attitudes toward mathematics was expected to get 3 or 4 points per response. Accordingly, the student was expected to average closer to 4 points across the five items. On the contrary, a student with negative attitudes was expected to mark responses worth 1 or 2 points, thus getting an average closer to 1.

Step 3: Data screening. Once the component items are selected, the next step is to screen the data to ensure its quality. It is important to check that the data make sense, that they are within the range of possible response values, that there is enough variation in student responses, and that the response rate is adequate. Table 2 presents the number of cases, minimum and maximum scores, mean and standard deviation for each item under analysis (after reverse scoring). In all, 24,876 students (93.8%) from five countries provided a valid response to all five items. Consistent with the 4-point Likert format, the minimum and maximum values for each item were 1 and 4, respectively. In item 21a, a mean of M = 2.90 indicates that, on average, students reported liking mathematics. In item 24d, M = 3.42

indicates that, on average, students were mid-way between agreeing and strongly agreeing with "mathematics is important to everyone's life" (after reverse scoring).

Table 2: Descriptive Statistics for Five Items Related to Attitudes toward Mathematics

	N	Minimum	Maximum	Mean	Std. Deviation
21a_like mathematics	25,830	1	4	2.90	0.846
24a_enjoy [r]	25,675	1	4	2.96	0.845
24b_boring	25,539	1	4	2.67	0.901
24d_important[r]	25,640	1	4	3.42	0.735
24e_job[r]	25,488	1	4	2.65	0.937
Valid N (listwise)	24,876				

[r] Reverse scored variable.

Step 4: Looking for the underlying factor. The next phase was to test empirically if these five items were in fact measuring the same latent variable or factor (attitudes toward mathematics). For instance, a student with moderate positive attitudes towards mathematics would be expected to *agree* with "I enjoy learning mathematics," and *disagree* with "mathematics is boring."

Principal component analysis (PCA) looks for underlying components that account for the common variance shared by the items. In this example, it was hypothesized that only one component (attitudes toward mathematics) would explain most of the covariation among the items.

The *correlation matrix* is the starting place for a PCA. As shown in Table 3, for the five items under analysis, all the Pearson's *r* were positive and significant at $p < .0005$. This was an expected outcome since all the items were scaled in the same direction and the large *N* made the standard errors to shrink. One correlation was high ($r = .72$), four moderate ($.45 < r < .65$), and five low ($r < .35$). Items 21a "how much do you like mathematics" and 24a "I enjoy learning mathematics" had the highest correlation ($r = .72$). This made sense considering how close the meaning of these two items was. Item 24d "mathematics is important to everyone's life" had low correlation with all the other items. However, since conceptually it made sense to use this item as an indicator of attitudes toward mathematics, it was kept for further analyses.

Table 3: *Inter-Item Correlation for Attitude Related Items*

	21a	24[a]	24b	24d	24e
21a_like	-	.720**	.490**	.260**	.520**
24a_enjoy [r]		-	.483**	.309**	.546**
24b_boring			-	.183**	.348**
24d_important[r]				-	.304**
24e_job[r]					-

** $p < .0005$
[r] Reverse scored variable.

Building upon the variance shared among all the items, PCA creates new variables (components) that account for most of the variance in the original items. As shown in Table 4, PCA identified one underlying component, with loadings (component-item correlations) ranging from $r = .87$ to $r = .49$. "I like mathematics" (21a) and "I enjoy learning mathematics" (24a) shared the most variance with the factor. Since these two items can be considered prototype indicators of students' attitudes toward mathematics, these results served as evidence that the correct construct was being targeted.

Table 4: Component Matrix

Item	Component 1
21a_like	.851
24a_enjoy[r]	.868
24b_boring	.751
24d_important[r]	.486
24e_job[r]	.685

Extraction Method: Principal Component Analysis.
[r] Reverse scored variable.

Step 5: Assessing the reliability of the scale. At this stage it was important to know how reliable a scale consisting of the selected items would be. In Table 5, the item-total correlation (or point-biserial correlation) indicates the relationship between each item and the rest of the items in the scale combined. If all the items are targeting the same underlying construct, high item-total correlations are expected. With the exception of item 24d ($r = .33$), correlations were of high and moderate sizes ($r > .5$), thus supporting the hypothesis that the items were measuring a common underlying construct.

Cronbach's alpha (α) was used to measure the *internal consistency* of this 5-item scale. An $\alpha = 0.79$ indicated that 79% of the total scale variance could be attributed to systematic variance in the latent variable (attitudes toward mathematics). This is a fairly respectable reliability considering the small number of items included in the analysis.

Table 5: Item-Total Statistics

	Corrected Item-Total Correlation	Alpha if Item Deleted
21a_like	.6982	.7008
24a_enjoy[r]	.7275	.6908
24b_boring	.4999	.7682
24d_important[r]	.3324	.8100
24e_job[r]	.5789	.7419

[r] Reverse scored variable.

The last column in the table shows what alpha would be if an item were deleted from the scale. In this example, deleting item 24d would lead to a 0.02 increase in the reliability from $\alpha = 0.79$ to $\alpha = 0.81$. Deleting any other item would lower the reliability. However, considering that the reliability increased, though not that much, and that item 24d is conceptually related to the construct of interest, a decision was made to keep it for the following analyses.

Step 6: Computing index scores. An index score is a number derived from several source items. This score was computed by averaging the score points associated with each student response. For example, a student with highly positive attitudes toward mathematics could have answered: 21a = *like at lot* (4 points), 24b = *strongly disagree* (4 points); and 24a, 24d, 24e = *strongly agree* (4 points each after reverse scoring). This student would have a mean index score of (4 x 5) / 5 = 4. Another student with somewhat negative attitudes may have responded: 21a = *dislike* (2 points); 24a = *disagree* (2 points); 24b = *strongly agree* (1 point); 24d = *agree* (3 points); and 24e = *strongly disagree* (1 point). The index score for this student would equal (2 + 2 + 1 +3 + 1) / 5 = 1.8.

This procedure preserves the *original metric* of the item format, thus allowing for a straightforward interpretation of the index scores. For instance, a score of 4 means that the student consistently chose the options indicating the most positive attitude toward mathematics; it is equivalent to strongly agreeing to all the positive statements about mathematics. A score of 1.8, on the contrary, can be interpreted as on average disagreeing to positive statements about mathematics.

Index scores also present the advantage of *maximizing information* when dealing with *missing data*. If attitudes toward mathematics were solely based on item 21a "how much do you like mathematics?," and if a student did not answer it, then she would be counted as missing in the final report. However, if item 21a is only one of the component items for an index, it is still possible to compute a derived score for that student by averaging her valid responses to the other four questions only. In computing index scores, TIMSS requires that at least two-thirds of the component items had valid responses. In a scale based on five items, this rule allows for one missing response only.

Step 7: Creating index levels. After computing index scores for each student, TIMSS classifies each case into one of the three levels: high, medium, and low. To form these three groups, two cut-points must be established in the underlying continuous scale. In the PATM scale, this was done by assigning to the high level of the index all those students with an index score greater than 3, and to the low level those with index score less than or equal to 2. The remaining students (with scale score greater than 2 and less than or equal to 3) were assigned to the medium level. Conceptually, this corresponds to grouping in the high level those students who at least *agree* to the positive statements about mathematics and in the low level those students who *disagree* to *strongly disagree* in their responses to the positive statements.

It is important to note that these cut-points have an *absolute* value and similar interpretation across all the countries. No matter from which country a student comes from, if he was classified in the medium level of the PATM index, it is possible to infer what his pattern of response could have been on the component questions. This shows how the countries differ in their distribution of students in the high, medium, and low levels of the PATM index.

In choosing the cut-points for index levels, TIMSS prioritizes the *interpretability* of the categories within and across the countries. The index levels have to make sense conceptually (rather than just having a mathematical justification). Readers can easily compare the percentage of students at each level of the index, both within and across the countries. The relationship between index levels and achievement in mathematics can be easily stated, by presenting together the percentage of student and the average mathematics score obtained by the students in each level.

Step 8: Assessing the index. Once the measurement strategy for an index has been established, it is important to check if the data behave according to expectations. Failure to do so may suggest problems with the validity and/or reliability of the measures. Table 6 presents the percentage of students in the high, medium and low levels of the PATM index for five countries. According to expectations, students in the high level of PATM outperformed their peers in the medium and low levels. It is noteworthy that, across the five countries, the vast majority of the students were classified

in the high and medium level of the index. This suggests that the countries have been relatively successful in developing positive attitudes toward mathematics in their students.

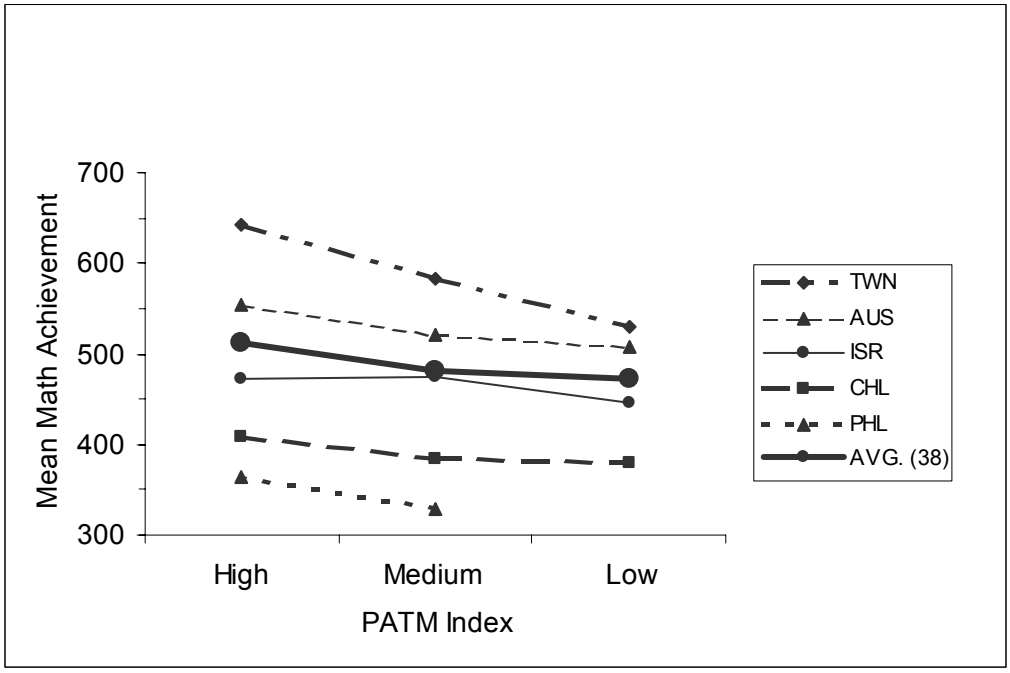
Table 6: Index of Students' Positive Attitudes Toward Mathematics (PATM) and Mathematics Achievement

	High PATM		Medium PATM		Low PATM	
	Percent of Students	Average Achievement	Percent of Students	Average Achievement	Percent of Students	Average Achievement
Philippines	59	365	38	328	2	—
Chile	45	408	47	385	8	379
Israel	44	472	45	474	10	445
Australia	30	544	55	520	15	508
Chinese Taipei	23	643	59	582	18	529

Because results are rounded to the nearest whole number, some totals may appear inconsistent.
A dash (—) indicates insufficient data to report achievement.

Line graphs are useful to checking the hypothesized positive trend between index levels and achievement. Figure 3 shows the relationship between the PATM index and mathematics achievement for five countries separately, plus the international average for all the 38 countries participating in TIMSS 1999. The slope of the lines can be used as an indicator of how well the index discriminates among students with different achievement levels. The steeper the lines, the greater are the differences between the average mathematics scores of one index level and the next. The capacity of the PATM index to differentiate between students varied from country to country: it was higher for Chinese Taipei and lower for Israel. For Chile and Australia, the discrimination between high and medium levels was better than the discrimination between medium and low levels. For the Philippines, only 2% of the total cases lay in the low level of the index; achievement for the low PATM level was not reported due to insufficient data to produce a reliable estimate. The international average line shows that the index could discriminate well among students with different mathematics achievement.

Figure 3: Relationship Between Index of Students' Positive Attitudes Toward Mathematics (PATM) and Mathematics Achievement



Note. AUS = Australia, CHL = Chile, ISR = Israel, PHL = Philippines, TWN = Chinese Taipei, AVG = International average for 38 countries participating in TIMSS 1999.

Combination of Responses Method

Indices can be based on actual *combinations of responses* to questions related to the same underlying construct. Here the cases are directly classified in one of the index levels, depending upon the combination of responses to the source questions. An example of this procedure can be found in the Home Educational Resources (HER) index reported in TIMSS 1999. The steps used to develop this index are detailed below.

Step 1: Identify component questions. The HER index was based on students' responses to six items: mother and father education (items 7a, b), number of books in the home (item 10); and having a computer (item 11b), study desk/table for own use (item 11c), and/or dictionary (item 11d) at home (Figure 4). Since the countries differ in the definition of their educational levels, it was important to make sure that the national adaptations had been recoded in order to make them internationally comparable.

The HER index is a *proxy* for the cultural capital and economic resources of the students' families. This is a variable external to the school with a known and pervasive effect on educational achievement. Accordingly, it was important for TIMSS to report on how the students from different countries differed in their level of HER.

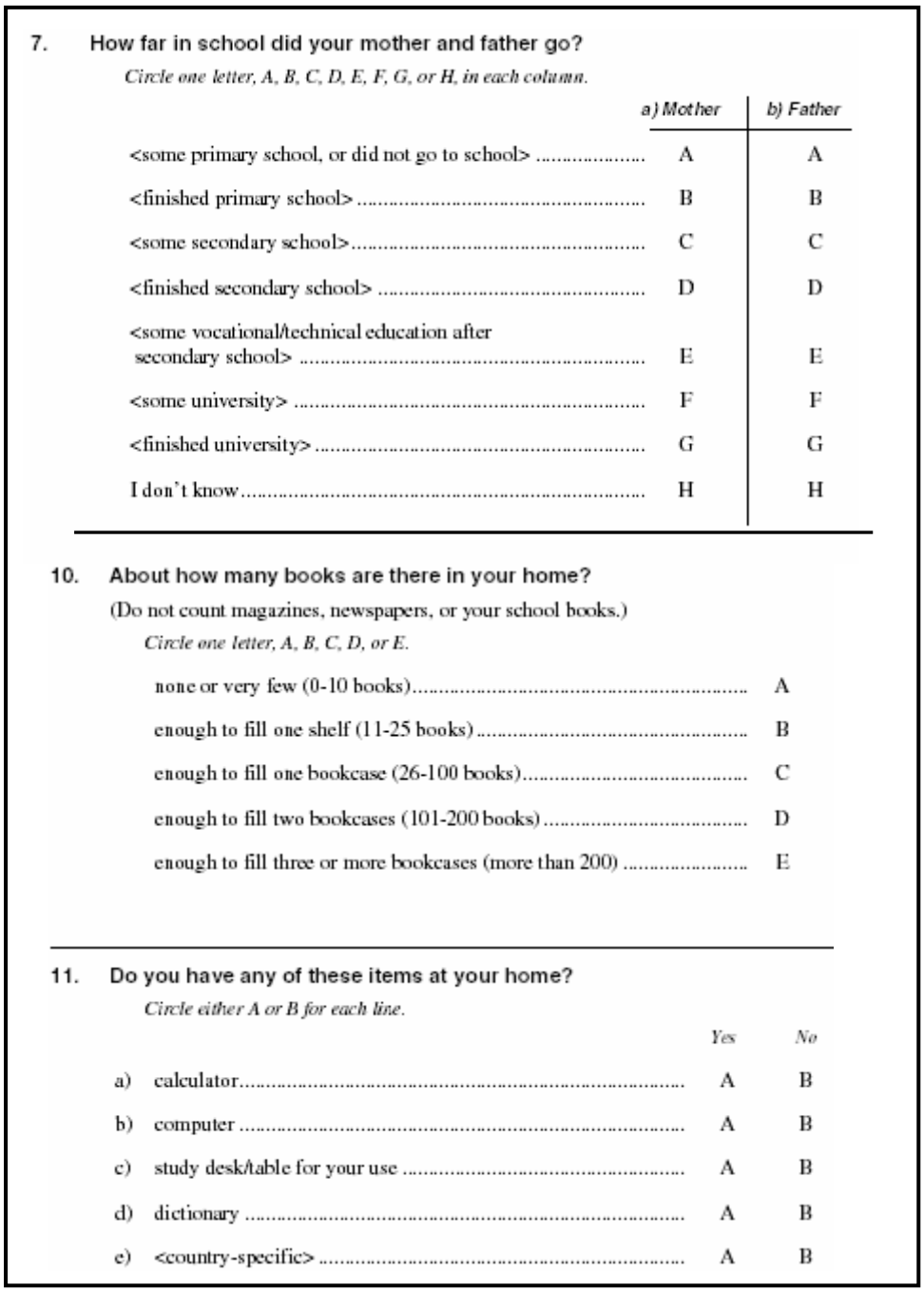
The highest level of education of either parent was a strong predictor of students' achievement in all countries. Although, in some countries mothers education alone predicted the achievement score equally well or better. For most of the TIMSS countries and for all the countries on average, the highest level of either parent discriminated better than mother education alone. Finally, deriving the highest education level allowed for maximizing the number of valid cases when data was missing for one parent. For all these reasons, a decision was made to use information from both parents in the HER index.

The number of books at home, the availability of computer, desk, and dictionary are all indicators of the socio-cultural level of the parents and of the acquisition power of the families. As such, these variables were expected to show a positive association with students' achievement.

Step 2: Data screening. Once the potential component questions have been identified, the next step was to check that the expected relationships between outcome and questions, and among the questions themselves, held true. Additionally, the questions were checked for the students' response rates. For the example countries, analysis of variance (ANOVA) showed that students differed significantly in their mathematics scores depending upon their response options to each question ($p < .0005$). In all cases the relationship was according to expectations, supporting the use of these items as source variables for the HER index.

Eta-squared (η^2) was used to characterize the strength of the relationship between mathematics scores and each source variable (Table 7). η^2 represents the proportion of variance in a continuous variable, such as mathematics achievement, that can be attributed to differences between groups, such as those with high, medium, and low values on an index. The highest level of education of either parent had a relatively stronger effect in Chile ($\eta^2 = .20$) and a weaker one in Chinese Taipei ($\eta^2 = .07$). In four countries, parents' education discriminated better than books and possessions at home. In Chinese Taipei, number of books was the best predictor of students' achievement, though.

Figure 4: Source items for Index of Home Educational Resources



7. How far in school did your mother and father go?

Circle one letter, A, B, C, D, E, F, G, or H, in each column.

	a) Mother	b) Father
<some primary school, or did not go to school>	A	A
<finished primary school>	B	B
<some secondary school>	C	C
<finished secondary school>	D	D
<some vocational/technical education after secondary school>	E	E
<some university>	F	F
<finished university>	G	G
I don't know	H	H

10. About how many books are there in your home?

(Do not count magazines, newspapers, or your school books.)

Circle one letter, A, B, C, D, or E.

none or very few (0-10 books)	A
enough to fill one shelf (11-25 books)	B
enough to fill one bookcase (26-100 books)	C
enough to fill two bookcases (101-200 books)	D
enough to fill three or more bookcases (more than 200)	E

11. Do you have any of these items at your home?

Circle either A or B for each line.

		Yes	No
a)	calculator	A	B
b)	computer	A	B
c)	study desk/table for your use	A	B
d)	dictionary	A	B
e)	<country-specific>	A	B

Chi-square (χ^2) and Spearman's correlation for rank ordered variables showed that a significant association existed among the items: the more educated the parents, the more books, and the more possessions ($p < .0005$). This co-variation pattern supported the hypothesis that the three items were measuring the same underlying construct.

Here it is interesting to comment on some cultural differences that showed up while analyzing the HER variables. The difficulties in measuring home background resources across countries were evident. While in some places the norm is that students do their schoolwork at the dining table (which may be the only table at home), in other places not having a desk/table is a sign of poverty. The strength of the association between number of books and achievement presented some regional patterns that reflected important *cultural differences* among the participants. In some countries, someone interested in a book will usually buy it if she can afford to. However, in other countries the

common practice is to borrow the book from a library, whether it can be afforded or not. In this latter situation, the number of books in the home is a less effective indicator of the socio-cultural level of the family. These cultural differences are always a challenge in designing and reporting data from large-scale surveys like TIMSS.

Table 7: Association Between Measures of Home Educational Resources and Mathematics Achievement

	Highest level of education of either parent η^2	Number of books at home η^2	Possessions at home η^2
Australia	.096	.043	.033
Chile	.202	.128	.091
Chinese Taipei	.072	.154	.079
Israel	.122	.075	.102
Philippines	.068	.048	.027

Note. First plausible value of mathematics scores used as outcome.

Step 3: Combining the data. Item options were combined so that students could be classified in the high, medium or low levels of the index. After doing some exploratory data analysis, the index levels were defined as follow: "High level of HER indicates having more than 100 books in the home; having all three educational aids; and either parent has finished university. Low level indicates having 25 or fewer books in the home; having none of the educational aids; and both parents' education is some secondary or less or is not known. Medium level includes all other possible combination of responses" (Mullis et al., 2000, p. 118). Again, since the index levels were based on the same combination of responses in all the countries, meaningful comparisons can be made across systems.

Step 4: Assessing the index. After the students are assigned to the different index levels, it is necessary to check the data for the distribution of students in various index levels. Table 8 presents the distribution of students in the high, medium, and low levels of the HER index, together with their mathematics score, for each example country. In all the cases, the expected positive trend was confirmed: the more home educational resources the students had, the higher their mathematics score. There were astonishing differences in HER from country to country: while only 3% of the students in Australia were in the low level of the HER index, just 3% of the students in the Philippines were in the high level.

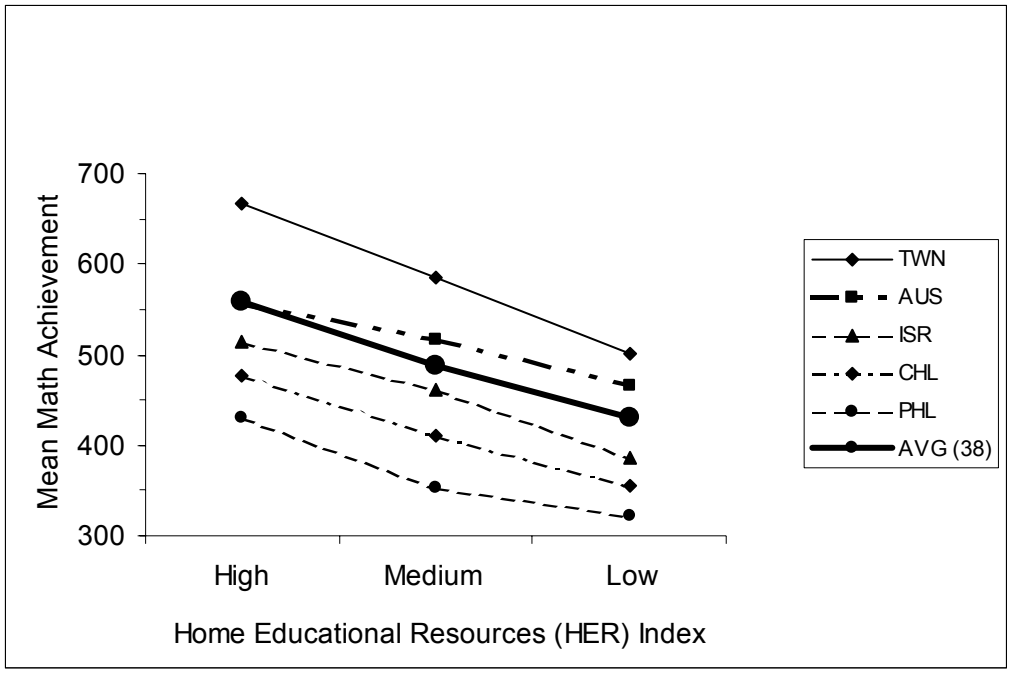
Figure 5 shows the relationship between the HER index levels and the mean achievement scores for the five countries, together with the international average for the 38 countries that participated in TIMSS 1999. According to expectations, higher index levels were associated with higher achievement scores. The international average line shows that the index discriminated well, on average, for all the TIMSS countries. Within the countries, the discrimination power of the index varied. It was higher for Chinese Taipei, lower for Australia. In the Philippines, it is worth noting that the discrimination between the high and medium levels was better than the discrimination between the medium and low levels, as indicated by the slope of the line.

Table 8: Index of Home Educational Resources (HER) and Mathematics Achievement

	High HER		Medium HER		Low HER	
	Percent of Students	Average Achievement	Percent of Students	Average Achievement	Percent of Students	Average Achievement
Australia	24	557	72	517	3	466
Israel	23	514	72	461	5	387
Chinese Taipei	8	666	84	586	8	502
Chile	6	476	56	410	38	355
Philippines	3	431	67	353	30	322

Because results are rounded to the nearest whole number, some totals may appear inconsistent.
A dash (—) indicates insufficient data to report achievement.

Figure 5: Index of Home Educational Resources (HER) and Mathematics Achievement



Note. AUS = Australia, CHL = Chile, ISR = Israel, PHL = Philippines, TWN = Chinese Taipei, AVG = International average for 38 countries participating in TIMSS 1999.

The next step was testing the discrimination power of the HER index. Different measures of association (ANOVA, η^2) showed that the index discriminated well among students with different achievement levels ($p < .0005$). In all the cases, the expected relationship was observed: the higher the HER levels, the higher the student achievement. Table 9 below indicates that the HER index discriminated somewhat better among Chilean students, while its power was somewhat weaker in the Philippines.

Table 9: Association Between the Index of Home Educational Resources (HER) and Mathematics Achievement

	η^2
Australia	.062
Chile	.160
Chinese Taipei	.099
Israel	.088
Philippines	.047

CONCLUSIONS

This paper presents three methods frequently used in TIMSS to report background data. The direct reporting method uses a single question from the questionnaires and report the raw data with minimal or no modifications. The scale development method summarizes data from multiple indicators to inform about a complex construct (e.g., attitudes toward mathematics). Numerical codes assigned to individual questions are averaged so that an underlying continuous scale is formed, and cutoff points are established to create index levels: high, medium, low. The third method creates index levels by specifying a unique combination of responses to the source items that must be satisfied for a student to be classified in its high, medium, or low level.

In deciding what to report and how to report it, TIMSS evaluate the suitability of different indices for all countries together and separately for each country. Suitability is here understood as the conceptual and statistical adequacy of the measures. Indices that combine data from several background questions present the advantage of being more stable (reliable) measures, thus increasing the validity of the data. The indices reported by TIMSS are conceptually related to the traits or phenomenon being measured, and also are related to academic achievement, meaning that they can discriminate adequately both across and within the countries. Representatives from each country review the proposed measures and by approving them, provide a seal of their utility and appropriateness to inform about their local educational contexts.

While there are many approaches for developing index scores (e.g., standardized measures, IRT scale scores) the ones here introduced present the advantage of being valid and reliable measures across different contexts, while being simple enough to be understood by non-technical users. Parsimony in reporting background data allows for a better communication with the readers of the TIMSS reports. It also provides a base for replicating and developing new indicators of educational context. This is of special importance for those interested in carrying on secondary analyses using the TIMSS data.

Reporting background information from large-scale international surveys like TIMSS is a challenging task. The unique characteristics of the educational systems of each country, language and cultural differences, must all be carefully taken into consideration to ensure the comparability of the data. In some countries, students may be more inclined to choose more extreme response options (e.g., *strongly agree*, *strongly disagree*), while in other places these options may be reserved for extreme opinions/feelings only. Standards for judging how prepared teachers are to teach, or how adequate the school facilities are, are also likely to vary from country to country. Survey data built on people's perceptions rather than on objective conditions, and TIMSS data must be interpreted accordingly. Awareness of these cultural differences is a must to develop fair and meaningful indicators of educational contexts.

References

Gonzalez , E. J. and Miles, J. A. (2001). *TIMSS 1999 User Guide for the International Database, Supplement Two: Documentation of National Adaptations of Background Questionnaire Items.* Chestnut Hill, MA: International Study Center Lynch School of Education Boston College.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., Chrostowski, S. J. & O'Connor, K. M. (2001). *TIMSS Assessment and Frameworks and Specifications 2003*. Chestnut Hill, MA: International Study Center Lynch School of Education Boston College.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., Chrostowski, S. J. & Smith, T. A. (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: International Study Center Lynch School of Education Boston College