ED 414 281 TM 027 669

AUTHOR Smith, Teresa A.

TITLE The Generalizability of Scoring TIMSS Open-Ended Items.

PUB DATE 1997-03-00

NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Comparative Education; English; *Error of Measurement; Foreign Countries; *Generalizability Theory; Intermediate Grades; International Education; *Interrater Reliability; Junior High Schools; Mathematics; Middle Schools; Sciences; *Scoring; *Test Items

IDENTIFIERS Free Response Test Items; *Middle School Students; *Third International Mathematics and Science Study

ABSTRACT

The Third International Mathematics and Science Study (TIMSS) measured mathematics and science achievement of middle school students in more than 40 countries. About one quarter of the tests' nearly 300 items were free response items requiring students to generate their own answers. Scoring these responses used a two-digit diagnostic code rubric with the first digit determining the correctness of the response and the second used to identify certain types of responses showing common approaches or misconceptions. This paper discusses relative contributions of the sources of error variance in student and country-level scores on these open-ended items due to rater effects as a function of item type. Fifty student booklets from 7 English speaking countries were analyzed for 350 student responses for each item. Generalizability studies determined the variability of scores associated with effects due to raters. Generalizability coefficients indicated a high degree of reliability in the relative rankings of a country's average score on TIMSS free-response ratings based on data from the cross-country coding study. The generalizability for an individual's score on a particular item was found to be somewhat less stable for some items, but this was not a concern, since the purpose of the TIMSS was to report country-level averages. (Contains six tables and five figures.) (SLD)

TIMSS

The Generalizability of Scoring TIMSS Open-Ended Items

Teresa A. Smith
TIMSS International Study Center
Boston College

Presented at AERA, 1997

Introduction

The Third International Mathematics and Science Study (TIMSS) measured mathematics and science achievement of middle-school students (grades 7 and 8 in most cases) in more than 40 countries.[1] The test instruments comprised nearly 300 items representing a range of mathematics and science topics, about a quarter of which were free-response items requiring students to generate and write their own answers. While the open-ended responses to these items provide the opportunity to investigate common approaches, methods, and misconceptions of students across countries, they also present a challenge in ensuring internationally reliable scoring of the student responses.

The scoring of the free-response items utilized 2-digit diagnostic-code rubrics specific to each item, where the first digit determines the correctness level of the response, and the second digit is used to identify certain types of responses showing common approaches or misconceptions. The majority of free-response items were short-answer items, which were all worth 1 score point, while extended-response items may be worth a total of 2 or 3 points. The scoring rubrics for the free-response items were developed in English and then translated by each of the participating countries into their own language to be used by coders in their own country, making the establishment and verification of uniform coding procedures an important aspect of quality control for TIMSS. In addition to developing detailed coding manuals with student examples and conducting international training sessions to assist each country in their free-response coding effort, the TIMSS quality assurance program provided for both within-country and across-country studies of the interrater reliability of free-response-item coding.[2] The data from these studies were used to provide basic documentation about the level of agreement in the coding of students' responses.[3]

[1] TIMSS was conducted by the International Association for the Evaluation of Educational Achievement (IEA) with international management of the study under the direction of Albert E. Beaton at the Center for the Study of Testing, Evaluation and Educational Policy, Boston College. Funding for the international coordination of TIMSS is provided by the U.S. National Center for Education Statistics, the U.S. National Science Foundation, the IEA, and the Canadian government. Each participating country provides funding for the national implementation of TIMSS.

[2] Funding for the across-country coding reliability study was obtained as part of a special grant from the U.S. National Center for Education Statistics for quality assurance activities.

[3] Results reported in Mullis, I.V.S., and Smith, T.A., (1996), "Quality Control Steps for Free-Response Scoring" in M.O. Martin and I.V.S. Mullis (eds.), *Third International Mathematics and Science Study: Quality Assurance in Data Collection*. Chestnut Hill, MA: Boston College.

A generalized analysis of variance (GENOVA)[4] study of the international cross-country coding data was conducted to determine the relative contributions to the total variance in student scores from main and interaction effects due to students, countries of origin, and raters. This study utilized over 10,000 original student responses from 7 different English-test countries, coded by a group of 39 coders from 21 of the TIMSS countries. The results of this analysis were used to estimate generalizability coefficients for both the country-level average scores and the individual student-level scores on 31 different math and science free-response items. These generalizability coefficients are used as a measure of the reliability of the free-response-item scores, as they reflect the proportion of observed variance in the international student response sample due to "true-score" or "universe score" variance, with the object of measurement defined as either the student-level score or the country-level average score. This paper will discuss the relative contributions of the sources of error variance in student and country-level scores due to rater effects as a function of item type. The generalizability of both student-level and country-level scores are presented, and the effect of the total number of raters and student sample size used within each country on the reliability of country-level average item scores is discussed.

Description of the Cross-Country Interrater Reliability Study

The Item Sets: A total of 31 items, the 14 mathematics and 17 science items contained in three of the eight TIMSS test booklets, were selected for use in the cross-country coding study. These items reflect slightly more than half of the total TIMSS mathematics and science free-response items. With 23 short-answer and 8 extended-response items, the total item set is also a good representative sample of the distribution of short-answer and extended-response items found in the TIMSS tests as a whole.

Language Considerations: The TIMSS study involved 30 different languages, and a cross-country study that included many of these TIMSS test languages would have been ideal. However, in order to maximize the number of across-country coding comparisons that could be made within reasonable budgetary constraints, it was determined that the most feasible study involved the use of student responses in English. To provide information about the coding in the TIMSS countries, the study needed to involve the actual coders from the participating countries. A number of these coders were fluent in English as well as their own language, but very few were bilingual or multilingual in the other languages of interest. Therefore, an English-based coding

[4] All generalizability analyses were conducted using the GENOVA data analysis package described in Crick, J.E., and Brennan, R.L., (1983), *ACT Technical Bulletin Number 43: Manual for GENOVA*, American College Testing Program, Iowa City, IA.

study permitted the inclusion of a large number of countries and the use of original student responses.

The Participating Countries: A total of 39 coders from the 21 TIMSS countries listed in Table 1 participated in the international reliability study. Participation was voluntary, and all countries were invited to participate who could provide coders who were fluent in reading and scoring student responses in English. Countries could send as many as two coders, and all of the countries participating in the study did so except Canada, France, and Germany (who each sent one coder). The participation of two coders per country enabled the study to be conducted in one week and also enabled countries that had divided responsibility for the coding task by subject area to send one coder who specialized in science and another who specialized in mathematics.

Table 1
Countries Participating in the Cross-Country Coding Reliability Study

Australia	Ireland	Romania
Bulgaria	Latvia	Russian Federation
Canada	Lithuania	Singapore
England	New Zealand	Slovak Republic
France	Norway	Sweden
Germany	Philippines	Switzerland
Hong Kong	Portugal	United States

The Student Response Set: Fifty student booklets for each of the three booklet types chosen for the cross-country coding study were provided by each of the 7 English-test countries in Table 1 (Australia, Canada, England, Ireland, New Zealand, Singapore, and the United States). The 50 booklets were selected by essentially choosing every other booklet from the 10% within-country reliability sample[5], excluding those booklets with mostly-blank responses. This provided a total sample of 350 student responses for each item, or a total of 10,850 student responses to serve as the basis for the study.

[5] Each country was to select every 10th booklet to be coded independently by two different coders for the within-country reliability sample.

The Coding Session Design: In order to accomplish all of the coding involved in the study during one week, the 31 items were divided into two sets of 15 and 16 items, respectively. The division was essentially according to mathematics and science items, but because the science items take more time to code there also was an attempt to balance the workload between the two groups. Item Set 1 contained 12 mathematics items and 4 science items; Item Set 2 contained 13 science items and 2 mathematics items. The coders also were divided into two groups, with one coder from each country in each of the groups. Information about the division of items was sent to the countries and coders in advance so that coders could receive refresher training in the items they were to score. Coders were to bring their own coding guides so that they could follow as closely as possible the procedures used in the within-country scoring. For Canada, France, and Germany (the three countries with only one coder), the coders elected to score Item Set 1. Thus, 21 coders worked on scoring Item Set 1 and 18 on scoring Item Set 2. Because time permitted, 4 mathematics and 8 science items were scored by both groups of coders, which was not part of the original study design.

As shown in Table 2, the 350 student responses for each item were divided into seven equivalent stacks of 50 responses. These stacks included responses from each of the seven countries supplying student responses, with each stack containing seven or eight responses from each of the countries. The responses for each item were organized to be distributed across coders according to a balanced rotated design. The seven stacks were placed into groups of three, such that every stack appeared with every other stack. Each unique combination of stacks is referred to as a rotation (A,B,C,D,E,F,G). Coders within each item set were randomly assigned to a rotation number so that each coder would score three stacks of responses for each item, or a total of 150 student responses (comprising 21 or 22 responses for each of the 7 English-test countries). This design also ensured that every coder shared a stack of at least 50 student responses with every other coder scoring the same set of items.

Table 2

The Design for Assigning Student Responses to Coders

Coder	Stacks	Each Stack
Coder A	1, 7, 5	
Coder B	2, 1, 6	50 Student Responses
Coder C	3, 2, 7	Responses from all 7 countries
Coder D	4, 3, 1	- 8 responses from one country
Coder E	5, 4, 2	- 7 responses from the other 6 countries
Coder F	6, 5, 3	
Coder G	7, 6, 4	

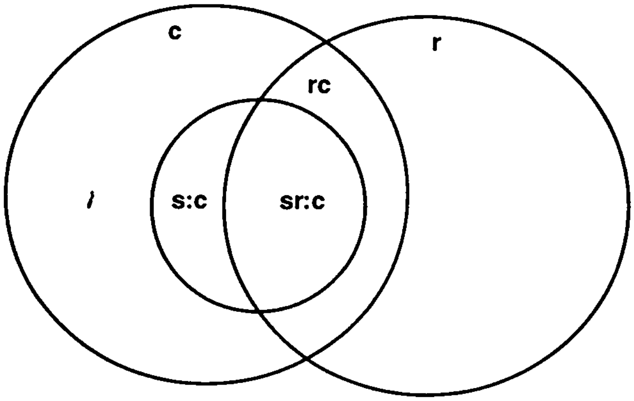
Given that the design for assigning student responses to coders yielded seven combinations of the three stacks of student responses, and that the study involved 21 coders scoring Item Set 1 (primarily mathematics), there were three full rotations of coders for Item Set 1. Thus, for the Item Set 1 Coders, each student response was coded by coders from 9 different countries. For Item Set 2, where 18 coders participated, there were not quite enough coders for three full rotations, and not all responses were scored by nine coders, some receiving seven or eight codes depending on the rotation. For the 12 items scored by both groups of coders, student responses received 16 to 18 codes.

Generalizability Study Design and Analysis

The data collected during the cross-country interrater reliability study were used to conduct generalizability study (G-study) analyses to determine the variability in the scores on TIMSS items associated with effects due to raters.. The measurement model used in these analyses was based on a partially-nested student-within-country by rater design (s:c x r), where sets of student responses from each of the seven English-test countries were each rated by different raters. The Venn diagram shown in Figure 1 depicts the decomposition of the total variance in observed scores into the constituent variance components defined in this G-study measurement model.

Figure 1

Venn Diagram Showing Sources of Variance for each Fixed-Item Generalizability Study



c = country
s:c = student-within-country
r = rater
rc = rater-by-country-by-interaction
sr:c = student-by-rater interaction
(within-country)

Variance Components: $\sigma^2_{Tot} = \sigma^2_c + \sigma^2_{s:c} + \sigma^2_r + \sigma^2_{rc} + \sigma^2_{sr:c,e}$

(The last variance term is confounded with other unknown sources of error not specified in the measurement model.)

These analyses treated the items as fixed, and a separate generalizability analysis was done for each item. This type of analysis seemed the most appropriate for the cross-country coding study data for two primary reasons: (1) Each free-response item has its own unique diagnostic-code scoring rubric, so it was valuable to obtain generalizability information about each item. (2) Due to the TIMSS test design[6], in which items and students are nested within booklets, and the assignment of coders in the cross-country study to different item sets that were divided along math/science lines, only a small number of common items existed for measuring the interaction effects involving students, items and raters. Since these effects vary substantially across items, the variance estimates obtained with a small subset of items would not be very generalizable to the larger universe of the TIMSS free-response items.

Also, in the cross-country coding study design, the seven subsets (rotations) of 150 student responses were each coded by a different set of raters (from 2 to 6, depending on the item and the rotation). Therefore, separate generalizability analyses were conducted for each of the seven rotation sets for each item. The variance estimates obtained for these seven analyses were averaged

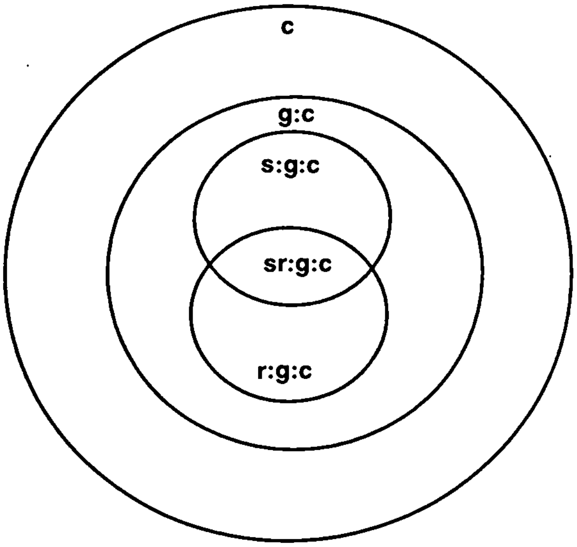
[6] TIMSS test design described in Adam, R.J., and Gonzalez, E.J., (1996), "The TIMSS Test Design" in M.O. Martin and D.L. Kelly (eds.), Third International Mathematics and Science Study: Technical Report Volume 1: Design and Development, Chestnut Hill, MA: Boston College.

to obtain a set of overall variance estimates for each item based on the entire set of coders and student responses utilized in the study.

The average variance estimates obtained from the initial G-studies were then used to compute generalizability coefficients for country-level and student-level scores based on a set of decision-study (D-study) designs. The D-study designs for the country-level scores were modeled upon typical within-country sampling and coding designs used by the TIMSS countries during the response-coding phase of the study and are used to investigate the effects of sample size and number of raters on the generalizability of the TIMSS item scores. The D-study model used in these analyses and the variance components in observed scores are shown in Figure 2.

Figure 2

Venn Diagram Showing Sources of Variability for the Decision-Study Designs Based On Within-Country Coding Schemes



c = country
g:c = group-within-country
s:g:c = student-within-group-within-country
r:g:c = rater-within-group-within-country
sr:g:c = student-by-rater interaction
(within-group-within-country)

Variance Components: $\sigma^2_{Tot} = \sigma^2_{c} + \sigma^2_{g:c} + \sigma^2_{s:g:c} + \sigma^2_{r:g:c} + \sigma^2_{sr:g:c,e}$

(The last variance term is confounded with other unknown sources of error not specified in the measurement model.)

In this partially-nested (sxr):g:c design , student responses (s) within each country (c) are divided into groups (g), with each student response in the group being coded by a rater assigned to this group. The number of groups in each country is determined by the number of raters available to code the total set of student responses for each item. In order to estimate the variance components for this modified D-study design, a number of assumptions are made relating the G-study results to the expected variance components from the D-study design. First, the division of students into groups within each country is assumed to be random. Therefore, the variance component associated with groups is assumed to be negligible, and the variance associated with students within the groups is assumed to be the same as that within the country as a whole. The effects due to raters (main and interaction effects) are assumed to be the same as that found in the G-study design. In other words, the uncertainty in scores due to a rater is expected to be the same whether that rater is selected from within the same country as the student responses or from the larger universe of all countries. Since in the D-study design, raters are nested within countries, the effect due to the country-by-rater interaction is confounded with the main effect due to rater. Based on these assumptions the variance components for the country-level D-study design are approximated based on estimates obtained for the G-study analyses:

D-Study Term		Estimated Values from G-Study
$\sigma^2_{s:g:c}$	$\cong$	$\sigma^2_{s:c}$
$\sigma^2_{r:g:c}$	$\cong$	$\sigma^2_{r} + \sigma^2_{rc}$
$\sigma^2_{sr:g:c}$	$\cong$	$\sigma^2_{sr:c}$
$\sigma^2_{g:c}$	$\cong$	0

The generalizability of TIMSS item scores were evaluated at two different levels: the country-level average score and the student-level score. To compute generalizability coefficients, the object of measurement must be defined to determine the "universe score", and the error variance associated with the object of measurement must be computed. The generalizability coefficient is calculated from the ratio of the universe score variance to the total variance (universe score variance plus error variance). The universe score variance, relative error variance and generalizability coefficient expressions for the two different levels of TIMSS item scores are shown in Table 3.

Table 3

Equations for Generalizability Coefficients

Object of Measurement	Design	Universe-Score Variance	Generalizability Coefficient
Country-Level Average	(sxr):g:c	σ^2_c	$\frac{\sigma^2_c}{\sigma^2_c + \sigma^2_{rel(c)}}$ ⓐ
Student Score	(s:c)xr *	$\sigma^2_{s:c} + \sigma^2_c$	$\frac{\sigma^2_{s:c} + \sigma^2_c}{\sigma^2_{s:c} + \sigma^2_c + \sigma^2_{rel(s)}}$ ⓑ

*Note: the D-Study model for the student-level scores is based on the original G-Study design that assumes that both the student and rater could be selected from the entire universe of countries and raters.

ⓐ Relative error variance for country-level averages:

$$\sigma^2_{rel(c)} = \overbrace{\frac{\sigma^2_{s:g:c}}{n_{s:g:c}n_{g:c}} + \frac{\sigma^2_{r:g:c}}{n_{r:g:c}n_{g:c}} + \frac{\sigma^2_{sr.g:c}}{n_{s:g:c}n_{r:g:c}n_{g:c}}}^{\text{From D-Study Model}} \cong \overbrace{\frac{\sigma^2_{s:c}}{n_s} + \frac{\sigma^2_r + \sigma^2_{rc}}{N_r} + \frac{\sigma^2_{sr.c}}{n_s}}^{\text{G-Study Estimates}}$$

$n_{r:g:c} = 1$ (1 rater per group of responses), $n_{g:c} = N_r$ (total number of raters per country), $n_{s.g:c}n_{g:c} = n_s$ (total student sample per country)

ⓑ Relative error variance for student-level scores: $\sigma^2_{rel(s)} = \frac{\sigma^2_{rc}}{n_r} + \frac{\sigma^2_{sr:c}}{n_r}$

$n_r = 1$ (number of raters per student response)

As seen in the expressions in Table 3, the error variance is inversely related to the levels of the facets contributing to sampling error for each type of measurement. For the country-level scores, the relative error has contributions from both the variance due to students within countries and the variance due to rater effects (both main and interaction effects). The error variance decreases (and generalizability increases) with both the number of students (TIMSS within-country sample size) and the total number of raters within each country. For the student-level score, however, only the number of raters may be increased to improve the generalizability. The generalizability coefficient estimates were used to compute the generalizability of student-level scores with just a single rating and to examine the effects of sample size and number of raters on the generalizability of country-level average scores on TIMSS free-response items.

Results

Variance Component Estimates from the G-Study: Estimates of the variance components due to country, students-within-country, and rater (both main and interaction effects) were obtained from the G-study results averaged across all rotations for each item. These results are summarized in Table 4, which shows the average and range of the percent of total variance due to each source of variance across all of the mathematics and science items.

Table 4: Summary of Generalizability Study Estimates of Variance Components from the Cross-Country Coding Study for Mathematics and Science Items

	Percent of Total Variance					
	Math Items			Science Items		
Source of Variability	Average	Range		Average	Range	
		Min	Max		Min	Max
Country (c)	10%	1%	28%	4%	1%	11%
Student:Country (s:c)	84%	65%	96%	68%	38%	97%
Rater	0%	0%	3%	2%	0%	9%
Rater x Country (rc)	0%	0%	1%	0%	0%	1%
(Student x Rater):Country (sr:c)*	5%	1%	30%	26%	1%	54%

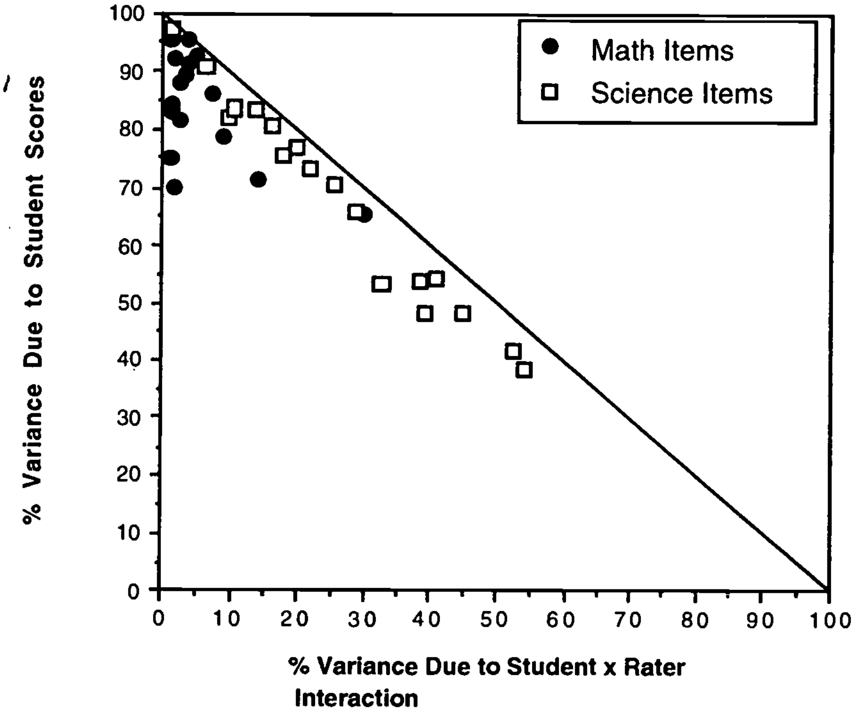
* Confounded with other unknown sources of error not specified in the measurement model.

Source: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Since the rating scales are different for different items, it is difficult to compare the absolute variance estimates; therefore, the results across items are compared on the basis of the percentage of total variance due to each facet. The largest source of variance, on average, is due to the student:country variance for both math and science items (84% for math and 68% for science). The largest difference between the math and science items is observed in the percent of total variance due to the student x rater interaction effect. For the science items, there is a substantial variance associated with the student x rater iteration for many items, with a maximum of 54% and an average value of 26%. In comparison, the largest student x rater interaction found for a math item was 30%, and the average value was just 5%. The student x rater interaction reflects the fact that the relative ranking of students was different for different raters. The coding rubrics for the science items are, in general, more complex and require more judgment on the part of the coder in determining the appropriate diagnostic code, which is consistent with the larger student x rater effect for science items. It should also be noted that the student x rater variance component is confounded with other unknown sources of error not specified by the measurement model. The relatively small main effect due to rater across nearly all items indicates that there is no overall difference in the severity of the raters when averaged across all of the student responses. The country x rater interaction reflects essentially zero percent of the total variance across all of the items in the study. This is a good sign for TIMSS based on the cross-country study results, as it indicates that the relative country-level score would not be affected much by the country of the rater and that the moderate percent of variance due to country reflects main effects due to true differences in country-level average scores.

The difference between the math and science items is also depicted in Figure 3, which shows the distribution of items based on the relative percentages of variance due to student:country and student x rater interaction. The math items (solid circles) are all clustered in the upper left-hand region of the plot, indicating a low variance due to the student x rater interaction, except for one item. In contrast, the science items cover a range of variance levels due to the student x rater interaction. The vertical distance between each point and the diagonal line reflects, primarily, the percent of variance due to the main country effect. Examination of the different data point locations indicates that the largest country effects are found for some of the math items.

Figure 3
Generalizability Analysis Results: Components of Variance By Item Type



Source: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

BEST COPY AVAILABLE

D-Study Results: Generalizability coefficients were computed based on the equations in Table 3 using the variance estimates obtained in the G-study analyses for each of the items. The generalizability coefficients for country-level average scores reflect the reliability of the relative ranking of a country's average score on an item based on the total sample of students, given that each student response receives one rating by a rater within that country. In general, there were many raters participating in coding, and the full set of student responses in each country was divided among these raters. Therefore, the generalizability coefficient is a function of the total sample size for each item and the total number of raters involved in rating the entire set of student responses in each country.

In Tables 5 and 6, generalizability coefficients for math and science items, respectively, are presented for two sample sizes (500 and 1000) and three levels of number of raters (5, 15 and 25) to be representative of the ranges of values encountered in most of the countries in the TIMSS study. The generalizability of country-level averages is quite high for most of the mathematics and science items, with generalizability coefficients greater than 0.7 at the lower levels of raters and students for all but three of the science items and all but one of the mathematics items. Increasing the number of raters from 5 to 15 results in an increase in the generalizability to above 0.7 for all of these items. This analysis suggests that the generalizability of country-level averages on free-response items would be an issue only if very small numbers of raters were involved in the coding in each country. Also, since the generalizability analyses reflect only the seven English-test countries represented in the international study, the variance in average scores for this particular set of countries is lower than what would be obtained if all TIMSS countries were represented in the analysis. Provided that the rater and student effects are comparable for the countries not included in the generalizability study sample, it is likely that the generalizability coefficients presented here underestimate the generalizability of country-level averages for the entire TIMSS population.

Table 5
Generalizability of Scores on Mathematics Items

Item	Generalizability Coefficients for Country-Level Averages[1]						Generalizability Coefficient for Student-Level Scores[4]
	Sample Size = 500[2]			Sample Size = 1000[2]			
	Number of Raters[3]			Number of Raters[3]			
	5	15	25	5	15	25	
M8	0.99	0.99	0.99	1.00	1.00	1.00	0.98
M1	0.99	0.99	0.99	1.00	1.00	1.00	0.99
M5	0.99	0.99	0.99	0.99	0.99	0.99	0.99
M9	0.99	0.99	0.99	0.99	0.99	0.99	0.98
M3	0.99	0.99	0.99	0.99	0.99	0.99	0.98
M6	0.99	0.99	0.99	0.99	0.99	0.99	0.99
[5] M11B	0.98	0.99	0.99	0.99	0.99	0.99	0.91
[5] M13B	0.98	0.99	0.99	0.99	0.99	0.99	0.85
[5] M11A	0.98	0.98	0.98	0.99	0.99	0.99	0.98
[5] M13A	0.97	0.97	0.98	0.98	0.99	0.99	0.97
M4	0.97	0.97	0.97	0.98	0.98	0.98	0.98
M12	0.97	0.97	0.97	0.98	0.98	0.98	0.92
M14	0.96	0.96	0.96	0.98	0.98	0.98	0.96
[5] M2A	0.95	0.95	0.95	0.97	0.97	0.97	0.99
M7	0.93	0.93	0.93	0.96	0.96	0.96	0.99
[5] M2B	0.92	0.93	0.93	0.95	0.96	0.96	0.95
[5] M10A	0.86	0.87	0.87	0.92	0.93	0.93	0.97
[5] M10B	0.58	0.74	0.79	0.61	0.79	0.84	0.69
Average	0.94	0.95	0.96	0.96	0.97	0.98	0.95

[1] Generalizability of the average country-level score on an item, based on one rating for each student.
[2] Total number of students within a country responding to each item.
[3] Total number of raters within each country scoring a subset of the student responses for each item.
[4] Generalizability of an individual student's score on an item, based on one rating.
[5] Two-part items; each part analyzed separately.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Table 6
Generalizability of Scores on Science Items

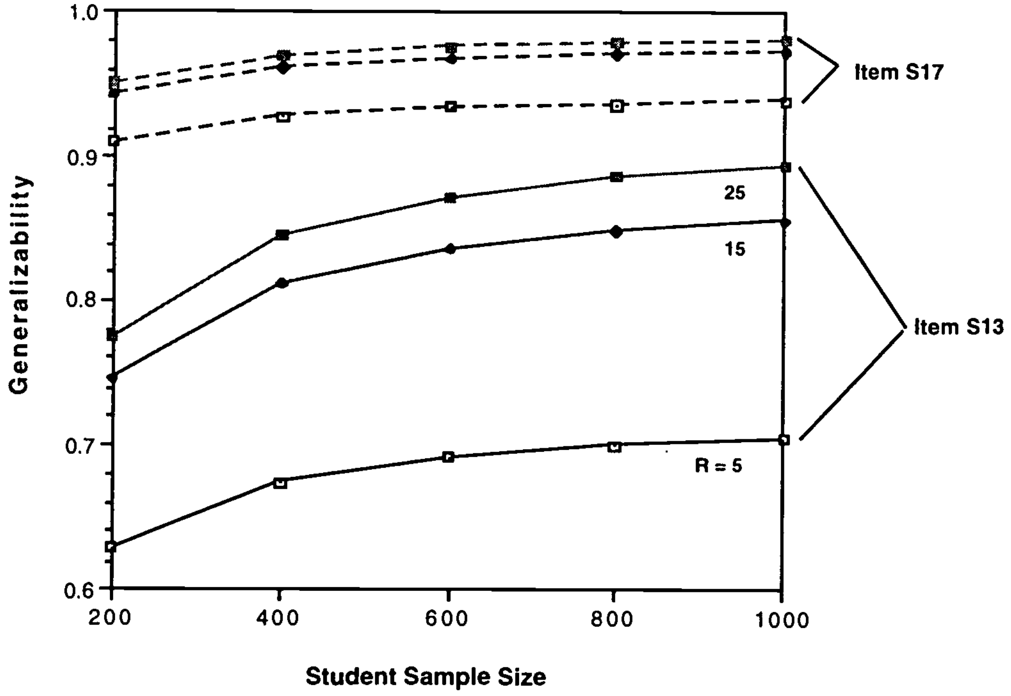
Item	Generalizability Coefficients for Country-Level Averages[1]						Generalizability Coefficient for Student-Level Scores[4]
	Sample Size = 500[2]			Sample Size = 1000[3]			
	Number of Raters[3]			Number of Raters[4]			
	5	15	25	5	15	25	
S9	0.97	0.98	0.98	0.98	0.99	0.99	0.90
S10	0.94	0.95	0.96	0.95	0.97	0.98	0.89
S17	0.93	0.97	0.97	0.94	0.97	0.98	0.66
S3	0.93	0.94	0.94	0.96	0.97	0.97	0.94
S6	0.92	0.95	0.96	0.94	0.97	0.97	0.82
S11	0.92	0.94	0.94	0.94	0.96	0.97	0.70
S2	0.90	0.92	0.93	0.93	0.95	0.96	0.86
S12	0.89	0.92	0.93	0.91	0.95	0.96	0.74
S4	0.88	0.93	0.94	0.90	0.95	0.96	0.54
[5] S7B	0.88	0.92	0.93	0.90	0.95	0.96	0.78
S1	0.87	0.87	0.87	0.93	0.93	0.93	0.99
[5] S7A	0.86	0.91	0.93	0.88	0.94	0.95	0.46
S8	0.84	0.89	0.90	0.87	0.93	0.94	0.80
S15	0.82	0.88	0.90	0.85	0.92	0.93	0.84
S14	0.76	0.87	0.89	0.78	0.89	0.92	0.59
S13	0.68	0.83	0.86	0.70	0.86	0.89	0.42
S16	0.60	0.79	0.84	0.61	0.81	0.87	0.56
S5	0.59	0.75	0.79	0.62	0.80	0.84	0.57
Average	0.84	0.90	0.91	0.87	0.93	0.94	0.72

[1] Generalizability of the average country-level score on an item, based on one rating for each student.
[2] Total number of students within a country responding to each item.
[3] Total number of raters within each country scoring a subset of the student responses for each item.
[4] Generalizability of an individual student's score on an item, based on one rating.
[5] Two-part items; each part analyzed separately.

SOURCE: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Generalizability of the country-level averages for all items increases as both the levels of student sample size and total number of raters increase, but for the typical sample sizes used in the TIMSS study (500 - 1000 students), generalizability is more sensitive to the number of raters than to increases in sample size for many items. The sensitivity of generalizability to numbers of raters and students differs from item to item, depending on the relative contributions to total variance due to country, student, and rater effects. This is demonstrated in Figure 4, which compares the effect of sample size for three levels of raters on the country-level generalizability coefficients for two different science items (S13 and S17). Item S13 was one of the science items with a very large student x rater interaction effect (54% of total variance) and a small main effect due to country (2%). In contrast, Item S17 had a larger main effect due to country (10%), with the student x rater interaction accounting for 32% of the total variance. These two example items and their coding guides are shown in Figure 5. The generalizability of Item S17 is high (>0.9) for all levels of students and raters compared to Item S13. The generalizability of Item S13, on the other hand, is much more sensitive to sample size and number of raters. With only five raters, a generalizability of at least 0.7 cannot be attained for any sample size. Increasing the number of raters to 15 results in a large increase in generalizability at all sample size levels, with generalizability ranging from about 0.75 to 0.84 for sample sizes of 200 to 1000. Further increasing the number of raters to 25 results in only moderate additional increases in generalizability. The effect of sample size is less pronounced and fairly comparable for all levels of raters. These results have implications for how to improve the generalizability of country-level item averages. If a relatively small number of coders are available for coding student responses within a country, it is better to divide the student responses to all items across as many coders as possible rather than to have coders specialize in a just a few items, coding large numbers of student responses to these items. This is particularly true for the science items.

Figure 4: Effect of Sample Size and Number of Raters on the Generalizability of Country Averages



Source: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

BEST COPY AVAILABLE

Figure 5

Example Items S13 and S17 with Coding Guides

Advantage of Two Eyes

Item S13

What is the advantage of having two eyes to see with rather than one?

Coding Guide	
Code	**Response**
Correct Response	
10	Mentions that two eyes allow depth perception or better perception of distance.
11	Mentions that two eyes allow seeing more or a wider field of vision.
12	Mentions that with two eyes one is still working if one eye is damaged.
19	Other correct.
Incorrect Response	
70	Mentions seeing twice as much.
71	Refers to energy or effort.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

New Species in Area

Item S17

What could be the unwanted consequences of introducing a new species to a certain area? Give an example.

Coding Guide	
Code	**Response**
Correct Response	
20	States that the natural (ecological) balance will be upset. A realistic example of a species is given.
21	States that the new species may take over and gives examples.
29	Other correct responses with examples.
Partial Response	
10	Adequate explanation (as in codes 20, 21), but no concrete and realistic example is given.
11	Only the realistic example is given, but no explanation.
12	States the new species cannot live here.
19	Other partially correct.
Incorrect Response	
70	Only an unrealistic example is given.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

Source: IEA Third International Mathematics and Science Study (TIMSS), 1994-95.

Conclusions

Generalizability coefficients computed for country-level averages and student-level scores indicate a high degree of reliability in the relative ranking of a country's average score on TIMSS free-response items based on using the data from the cross-country coding study. The generalizability of country-level averages is quite high for most of the items, particularly for math, with coefficients generally greater than 0.7. As might be expected, the generalizability for an individual student's score on a particular item was found to be somewhat less stable for some items, ranging from 0.42 to 0.99. Since the goal of TIMSS is to report country-level averages and not individual scores, the lower generalizability for individual scores is not a concern for the international TIMSS reporting on the average free-response item scores. These results serve as a caution, however, in performing secondary analyses that involve making generalizations from individual student scores on specific items. The effect of student sample size and number of raters within a country on the estimated country-level generalizabilities indicates that increasing the total number of raters among whom the student responses for each item are divided will result in improved generalizability of country-level averages. This finding is particularly true for some of the science items with higher student-by-rater interaction effects.

TM027669

ERIC

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: The Generalizability of Scoring TIMSS Open-Ended Items	
Author(s): Teresa A. Smith	
Corporate Source: TIMSS International Study Center, Boston College	Publication Date: August 1, 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒ Check here **For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.	The sample sticker shown below will be affixed to all Level 1 documents PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) **Level 1**	The sample sticker shown below will be affixed to all Level 2 documents PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) **Level 2**	□ Check here **For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: Teresa A. Smith	Printed Name/Position/Title: Teresa A. SMith/ Research Associate	
Organization/Address: TIMMS Int'l Study Ctr Boston College	Telephone: 617-552-8972	FAX: 617-552-8419
	E-Mail Address: smithtw@bc.edu	Date: August 11, 1997

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064

800 464-3742 (Go4-ERIC)

April 25, 1997

Dear AERA Presenter,

Hopefully, the convention was a productive and rewarding event. We feel you have a responsibility to make your paper readily available. If you haven't done so already, please submit copies of your papers for consideration for inclusion in the ERIC database. If you have submitted your paper, you can track its progress at http://ericae2.educ.cua.edu.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are soliciting all the AERA Conference papers and will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form on the back of this letter and stet **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can mail your paper to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1997/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/E