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

5. QUALITY CONTROL STEPS FOR FREE-RESPONSE SCORING

Ina V.S. Mullis
Teresa A. Smith

5.1 OVERVIEW

For the TIMSS main surveys, approximately one-third of the written test time was devoted to free-response items, both short-answer and extended-response types. Across the seven tests administered to the three populations (mathematics and science at Populations 1 and 2, as well as literacy, advanced mathematics, and physics at Population 3) and the performance assessments administered to Populations 1 and 2, TIMSS included approximately 300 free-response questions and tasks.

The free-response items were scored using two-digit codes with rubrics specific to each item. The first digit designates the correctness level of the response. The second digit, combined with the first, represents a diagnostic code used to identify specific types of approaches, strategies, or common errors and misconceptions.

The scope of the free-response scoring effort was very complex. With large within-country samples of students responding to the tests, and those student samples representing many countries, ensuring reliability of scoring was a major concern for TIMSS. It was therefore necessary to develop procedures for applying the coding guides reliably and to document coding reliability.

To meet the goal of ensuring reliable scoring, TIMSS used a three-pronged approach.

1. An ambitious schedule of training sessions was designed to assist representatives of national centers who would then be responsible for training personnel in their respective countries to apply the two-digit codes reliably.
2. To gather and document information about the within-country agreement among coders, TIMSS developed a procedure whereby approximately 10% of the student responses were to be coded independently by two readers.
3. To provide information about the cross-country agreement among coders, TIMSS conducted a special study at Population 2 whereby 39 coders from 21 of the participating countries coded common sets of student responses.

This chapter contains information about these three activities. For more details about the training sessions and the procedures for estimating within-country reliability, please see:

- "Training Sessions for Free-Response Scoring and Administration of the Performance Assessment" (Mullis, Jones, and Garden, 1996)
- *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995).

5.2 TRAINING SESSIONS FOR FREE-RESPONSE SCORING

Training sessions for free-response scoring were conducted in seven regions to provide easier access for participants, and smaller groups for the TIMSS trainers to manage. Accommodations also were required to address the TIMSS schedule, which, for the most part, required countries on the Southern Hemisphere timeline to test Populations 1 and 2 in the fall of 1994 and Population 3 in the fall of 1995. The remaining countries tested all three populations in the spring of 1995. Consistency across sessions was provided by using essentially the same training team and training materials for all the sessions. The members of the training team had considerable knowledge of the TIMSS tests and of the procedures used in training scorers to achieve high reliability in scoring. The team consisted of the following members: Mr. Chancey Jones, United States; Mr. Robert Garden, New Zealand; Dr. Graham Orpwood, Canada; Dr. Jan Lokan, Australia; and Dr. Ina Mullis, United States. Although not all training team members attended all of the training sessions, most attended the larger sessions and at least some attended each of the sessions. Representatives from each of the participating countries attended at least one training session. The only exception was Italy, which to date has not had the resources to code and enter the data it collected. The schedule of training sessions and the countries participating in each session are shown in Table 5.1.

A four-day training schedule was developed to introduce attendees to the TIMSS coding approach and give them practice in scoring example papers. The specifics of the schedule varied from session to session depending on the participants' involvement in the different aspects of TIMSS. However, in the most effective schedule for the sessions, the first three days were devoted to scoring procedures for the main survey at Populations 1, 2, and 3, respectively, and the fourth day to the performance assessment. The sessions began with an orientation covering the importance of the coding of free-response questions and performance tasks. The training team described the significance of the first and second digits in the TIMSS codes, explaining that the first digit is a correctness score, and that the second digit provides diagnostic information about the type of response. Other orientation topics included the importance of maintaining high reliability in conducting the coding process, the desirability of planning and conducting similar training in the participants' own countries, and the necessity of finding exemplar student papers within each country to use in the training process. Information also was provided about procedures for conducting the actual coding and the within-country reliability studies (as described in the *Guide to Checking, Coding, and Entering the TIMSS Data,* TIMSS, 1995).

Table 5.1
TIMSS Free-Response Item Coding Training Sessions

Date	Location and Participating Countries
October 10-12, 1994	Wellington, New Zealand (Populations 1 and 2) — Australia, Korea, New Zealand, Singapore
January 18-21, 1995	Hong Kong — Hong Kong, Japan, The Philippines, Thailand
January 25-28, 1995	Boston, United States — United States, Canada, Mexico, Norway, Kuwait
March 7-10, 1995	Enschede, The Netherlands — Belgium (Flemish), Denmark, England, France, Germany, Greece, Indonesia, Iran, Ireland, The Netherlands, Portugal, Scotland, Spain, Sweden, Switzerland
March 13-16, 1995	Budapest, Hungary — Austria, Bulgaria, Canada, Cyprus, Czech Republic, Hungary, Iceland, Israel, Latvia, Lithuania, Norway, Romania, Russian Federation, Slovak Republic, Slovenia, the Ukraine
July 17-18, 1995	Miami, United States — Colombia, Argentina
July 18-19, 1995	Pretoria, South Africa — South Africa
September 6, 1995	Wellington, New Zealand (Population 3) — New Zealand
September 28-29, 1995	Melbourne, Australia (Population 3) — Australia

Each participant in the training sessions needed a considerable amount of material, including the relevant coding guides, manuals, and packets of example papers for practice. TIMSS developed an extensive coding guide for each population, containing the individual rubrics developed for each of the TIMSS free-response items given to that population. Each rubric defined the scoring categories to be used for the item together with example student responses for each category.

Training packets were prepared for a subset of the items considered the most complicated to score. Across the populations, training packets were prepared for 14 mathematics and 23 science items. Each packet began with the rubric for the item followed by coded student responses illustrating each of the categories in the rubric. The packet also contained about 15 to 20 precoded student responses, with the codes known to the training team but not to the participants in the training session. These were used to give the participants an idea of what it is like to actually score student responses.

The purpose was not to conduct the actual training for the coders, but to present a model for use in each country and present an opportunity to practice with the most difficult items. The trainers emphasized the need for participants to prepare training materials for each of the items rather than only a sample of items, and the fact that for more difficult items more example responses might be needed to help coders reach a high degree of reliability.

The trainer for the item would begin by familiarizing the group with the rubric for the item and answering questions about the reasons underlying the categories. Then the trainer would invite the participants to code five or six of the example student responses. After the group had completed the coding for these responses, the trainer would read the scores for the responses and answer any questions from the group. This procedure was iterated until all the precoded responses were scored by the participants. Although there was insufficient time at the training sessions to achieve a consistently high level of agreement on each of the items, the procedures provided some practice for participants and an example for how training might be conducted in each country.

Spending only one day on each of the three populations with a fourth day for countries participating in the performance assessment made for a demanding and intense session for most participants. In the future, it would be beneficial to devote more time to training in free-response scoring. All in all, however, the model of developing detailed coding guides and “training the trainers” appears to have worked successfully.

5.3 WITHIN-COUNTRY RELIABILITY STUDIES

In addition to using well-defined coding rubrics and careful training procedures, TIMSS also implemented procedures to monitor inter-rater reliability within each participating country. The procedures were designed to document the degree to which the same codes were given to the same responses regardless of the coder.

The TIMSS International Study Center recommended that each country use a method whereby 10% of the booklets would be coded independently by two coders. Explained in detail in the *Guide to Checking, Coding, and Entering the TIMSS Data* (TIMSS, 1995), the procedure called for every 10th booklet to be coded by two different individuals, with neither knowing the identity of the other or the codes assigned.

Because it is important that the booklets selected for the reliability study represent the coding process in general, the procedures for the reliability sample needed to be as routine as possible to blend in with the normal coding procedure. The object is for the reliability sample to provide an estimate of the overall quality of the free-response coding.

The general idea was to divide coders into two equivalent groups (Group A and Group B, balanced in terms of numbers, training, and experience) and to divide the booklets into two equivalent sets (Set A and Set B, according to odd versus even school identification numbers). The coders in Group A were to code all the booklets in Set A and the 10% reliability sample of the booklets in Set B, while the coders in Group B coded all of the booklets in Set B and the 10% reliability sample of the booklets in Set A. Each group, therefore, handled both sets of booklets.

Because the coders could not know each other's codes for the reliability sample, ensuring a "blind" coding for the reliability sample necessitated the preparation of separate coding sheets for the 10% reliability sample. Coders were to handle the reliability set of booklets first, recording their results on a separate answer sheet. For the other set, the group coded all the booklets, and the codes were written directly into the booklets.

This procedure ensured that the coding of the reliability sample was conducted without the coders knowing the codes for the main survey and vice versa. It also ensured that different coders worked on the reliability sample than on the main coding, so that the same coder did not provide the codes for both reliability sample and main survey. As an additional step, countries were encouraged to try as much as possible to balance the reliability sample coding for each of the Group A coders across the different Group B coders, and similarly to balance the reliability sample coding for each of the Group B coders across different Group A coders. Countries also were encouraged to do the reliability scoring throughout the main survey coding. That is, for an hour or so each day, the Group B coders were to code every tenth booklet in the Set A batches, while the Group A coders were coding every tenth booklet in the Set B batches.

Many suggestions were given to countries about how to implement the 10% reliability scoring. Above all, however, the TIMSS International Study Center emphasized the importance of implementing a systematic plan to document the reliability of the coding schemes and stressed the need to enter the information about coding reliability into the database.

The within-country scoring reliability results for Population 2 presented in Tables 5.2 and 5.3 show the average and range of agreement across all mathematics and science free-response items, for each country. These results, showing the percentage of exact agreement for both the correctness score and the full two-digit diagnostic code, reveal a high degree of agreement for the countries that documented the reliability of their coding. Exact agreement between the first and second independent coders was particularly high for the first digit of the code, indicating the correctness score given the response. Since achievement on the TIMSS tests was estimated using only this first digit, it seems reasonable to conclude that scorer agreement within countries was robust. It appears that the use of open-ended items did not lower the reliability of the TIMSS tests, at least from a within-country perspective for countries providing within-country reliability data. Unfortunately, lack of resources precluded several countries from providing this information.

Naturally, the goal was to have 100% or perfect agreement between coders. In actuality, agreement above 85% is considered quite good, and above 70% acceptable. For the mathematics items, a very high percentage of exact agreement was observed for all countries, with averages across items for the correctness score ranging from 97% to 100% and an overall average of 99% across all 26 countries. In addition, all countries had at least 77% agreement on all items. While the percentage of exact agreement for science items was somewhat lower than for the mathematics items, it is still quite good, with averages across items for the correctness score ranging from 88% to 100% and an overall average across the 26 countries of 95%. Also, nearly all countries had greater than 70% agreement on all items. Percentages of agreement below 70% may be a cause for concern. In fact, as part of the database review prior to scaling the TIMSS achievement data, countries were alerted about items where scoring agreement was below 70%. In several instances, this information uncovered a misunderstanding in the coding approach and the student responses were recoded before the achievement data were scaled.

Although the results in Tables 5.2 and 5.3 indicate a high degree of within-country coder agreement in assigning the overall score to students’ responses, the data indicate less agreement concerning the second coding digit, designed to provide a more detailed view of the type of response. Even for the second digit, however, agreement was quite respectable, with averages across items ranging from 89% to 99% for mathematics items and from 73% to 98% for science items. Nevertheless, depending on the items and countries involved, some care should be taken in making comparisons across countries at this finer level.

Table 5.2
TIMSS Within-Country Free-Response Coding Reliability Data for Mathematics Items*

Country	Correctness Score Agreement			Diagnostic Code Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Australia	98%	90%	100%	90%	61%	98%
Belgium (Fl)	100%	98%	100%	99%	92%	100%
Bulgaria	98%	93%	100%	94%	59%	100%
Canada	98%	85%	100%	92%	70%	99%
Colombia	99%	97%	100%	96%	91%	100%
Czech Republic	98%	77%	100%	95%	68%	100%
England	100%	96%	100%	97%	89%	100%
France	100%	96%	100%	98%	93%	100%
Germany	98%	89%	100%	94%	75%	100%
Hong Kong	99%	94%	100%	96%	84%	100%
Iceland	98%	84%	100%	91%	73%	100%
Iran, Islamic Rep.	98%	94%	100%	93%	70%	100%
Ireland	99%	95%	100%	97%	83%	100%
Japan	100%	96%	100%	99%	90%	100%
Netherlands	98%	87%	100%	91%	68%	100%
New Zealand	99%	95%	100%	95%	81%	100%
Norway	99%	90%	100%	95%	79%	100%
Portugal	98%	88%	100%	93%	82%	99%
Russian Federation	99%	94%	100%	96%	84%	100%
Scotland	97%	81%	100%	89%	63%	99%
Singapore	99%	95%	100%	98%	87%	100%
Slovak Republic	97%	84%	100%	91%	70%	98%
Spain	98%	88%	100%	94%	75%	100%
Sweden	99%	90%	100%	94%	75%	100%
Switzerland	100%	95%	100%	98%	83%	100%
United States	99%	95%	100%	96%	85%	99%
AVERAGE	99%	91%	100%	95%	78%	100%

*Based on 26 mathematics items, including 6 multiple-part items.
Note: Percent Agreement was computed separately for each part, and each part was treated as a separate item in computing averages and range

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

Table 5.3
TIMSS Within-Country Free-Response Coding Reliability Data for Science Items*

Country	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Australia	91%	69%	99%	78%	48%	97%
Belgium (Fl)	100%	95%	100%	98%	82%	100%
Bulgaria	91%	63%	100%	81%	50%	100%
Canada	92%	76%	100%	80%	59%	99%
Colombia	97%	83%	100%	91%	73%	100%
Czech Republic	96%	87%	100%	90%	61%	100%
England	97%	90%	100%	91%	65%	100%
France	99%	95%	100%	97%	89%	100%
Germany	94%	81%	100%	84%	66%	100%
Hong Kong	94%	72%	100%	87%	56%	100%
Iceland	95%	74%	100%	83%	22%	98%
Iran, Islamic Rep.	88%	67%	100%	73%	33%	99%
Ireland	95%	87%	100%	89%	69%	100%
Japan	100%	96%	100%	98%	87%	100%
Netherlands	92%	75%	100%	79%	17%	100%
New Zealand	97%	90%	100%	90%	63%	100%
Norway	95%	87%	100%	91%	71%	100%
Portugal	96%	88%	100%	91%	75%	100%
Russian Federation	96%	87%	100%	91%	73%	100%
Scotland	89%	73%	99%	74%	52%	96%
Singapore	98%	92%	100%	95%	86%	100%
Slovak Republic	92%	62%	100%	81%	43%	100%
Spain	95%	85%	100%	88%	73%	98%
Sweden	94%	80%	100%	83%	54%	99%
Switzerland	98%	93%	100%	93%	85%	99%
United States	97%	90%	100%	89%	74%	100%
AVERAGE	95%	82%	100%	87%	63%	99%

*Based on 33 science items, including 4 multiple-part items.

Note: Percent Agreement was computed separately for each part, and each part was treated as a separate item in computing averages and ranges.

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

5.4 IMPLEMENTING THE CROSS-COUNTRY RELIABILITY STUDY

At the Salzburg National Research Coordinators' meeting in late 1994, the NRCs suggested that TIMSS also should obtain information about coding reliability across countries. Fortunately, the additional funds for the quality assurance program included some resources for this purpose, and staff set out to design and implement such a study. The TIMSS International Coding Reliability Study was conducted in December 1995 in Boston.

Considering the schedule for TIMSS and its resources, it was clear from the outset that the cross-country coding study could be ambitious, but would be far from all-inclusive. Thus, the purpose would need to be focused, and choices would need to be made about which TIMSS populations to include in the study, numbers of items, and so forth. The next sections discuss the issues involved, the decisions made, and the procedures used in conducting the study.

5.4.1 OVERALL PURPOSE

The goal of the study was to document the level of agreement across countries in coding student responses to the mathematics and science free-response items. More complex aims were discussed, such as studying the sources of potential bias among countries and obtaining a sense of how differences in free-response coding might have affected the overall scores for countries. The TIMSS Technical Advisory Committee, in particular, supported these more complex goals. However, these aims remained largely beyond operational feasibility given the TIMSS schedule and budget.

5.4.2 POPULATION

Free-response items played a substantial role in all of the TIMSS tests, including mathematics and science at Populations 1 and 2. At Population 3, there were three possibilities–one test for the general population, a second for the subpopulation having studied advanced mathematics, and a third for the subpopulation having studied physics. Valuable information could have been obtained from studying the scoring reliability of many of the items included in these various tests. However, since Population 2 was the only mandatory population for participation in TIMSS, it was the population selected for the cross-country study of coding reliability.

5.4.3 NUMBER OF ITEMS

At Population 2, TIMSS included 26 free-response items in mathematics and 33 in science. Clearly, TIMSS would have preferred a reliability study involving all of these items, but coders from the participating countries simply could not commit to a study of such extensive proportions. After careful consideration, three of the eight books at Population 2 were considered appropriate as a basis for the study. Together, these three books included

15 mathematics items and 18 science items. One mathematics and one science item were eliminated to better balance the workload for the coders. As shown in Table 5.4, 31 items were involved in the reliability study–14 mathematics items and 17 science items. The study included about half of all of the free-response items at Population 2 – 54% of the mathematics items and 52% of the science items.

Table 5.4
Number of Mathematics and Science Items in Cross-Country Coding Reliability Study

Booklet	Mathematics			Science			Mathematics and Science Combined		
	SA	ER	Total	SA	ER	Total	SA	ER	Total
#3	2	4	6	2	0	2	4	4	8
#7	3	1	4	4	2	6	7	3	10
#8	4	0	4	8	1	9	12	1	13
Total	9	5	14	14	3	17	23	8	31

Notes: SA=Short Answer; ER=Extended Response

5.4.4 NUMBER OF STUDENT RESPONSES PER ITEM

Three considerations emerged in making decisions about how many student responses to each item should be included in the study:

- The ability to link to the within-country reliability studies
- The language of testing
- The overall coding burden.

To strengthen the study, it was based on the same student responses as the within-country reliability studies. Again, the preference was to involve the full reliability sample across the selected items for all of the participating TIMSS countries. However, the overall coding burden led to a final decision to use half-samples.

The many different languages of testing were one of the most difficult obstacles in conducting the study. The original notion involved student responses from all participating countries. After collecting information about the availability of bilingual coders, however, it became clear that trying to incorporate student responses in a number of languages into the study (e.g., German, French, Spanish, and the Scandinavian languages as well as English) simply was not going to be feasible. 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. On the other hand, there was consensus that translating the students' responses into English introduces unknowns as well as being expensive and time-consuming. Under the circumstances, the best approach appeared to be using responses provided in English for the study. This enabled the use of original student responses and permitted participation by all countries wishing to work on the study.

Once the decision was made to base the study on student responses in English, 7 countries that administered the test in English were asked to provide student responses from their TIMSS testing at Population 2. Each country was asked to provide 50 student responses for each of the 31 items, essentially drawn from every other booklet in their within-country reliability samples. Since there were 7 English-test countries each supplying 50 responses for each item, the coding study involved 350 responses to each of the 31 items. This procedure resulted in a corpus of 10,850 student responses to serve as the basis of the study. The 7 countries devoting time and energy to supplying student responses included Australia, Canada, England, Ireland, New Zealand, Singapore, and the United States.

5.4.5 Number of Participating Countries and Coders

A total of 39 coders from 21 countries participated in the international reliability study. Participation was voluntary, and all countries were invited to participate. Table 5.5 lists the countries that participated, and the names of the coders are listed in Appendix G. Countries could send as many as two coders, and all of the countries participating in the study did so except Canada, France, and Germany (they each sent one coder). Two coders per country enabled the study to be conducted in one week. It 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 5.5
Countries Participating in Cross-Country 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

5.4.6 Two Groups of Items and Coders

In order to accomplish all of the coding involved in the study during one week, the 31 items were divided into sets of 15 and 16 items. 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. Although this was not part of the original plan, it provides an important link between the two groups of coders. During debriefing at the end of the study, it was ascertained that for the countries participating the study, coder responsibilities at Population 2 were more likely to have been assigned by booklet than by subject area. Of the study participants, only the Russian Federation and Hong Kong specialized by subject area. Even though most coders had backgrounds predominantly in either mathematics or science, during the actual coding in their countries they had scored both mathematics and science questions.

5.4.7 The Design for Each Group of Coders

As shown in Table 5.6, the 350 student responses for each item were divided into seven stacks of 50 responses. These stacks included responses across all seven countries supplying student responses, with each stack containing seven to 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. Also, in this assembly care was taken that each stack appeared once as the first set of student responses, once as the second set, and once as the third set. Each coder, then, scored three stacks of responses for each item. This meant that for each item, each coder scored a total of 150 student responses (comprising 21 to 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 5.6
The Design for Assigning Student Responses to Coders

Coder	Stacks	Each Stack
Coder A	1, 7, 5	· 50 Student Responses · Responses from all 7 countries - 8 responses from one country - 7 responses from the other 6 countries
Coder B	2, 1, 6	
Coder C	3, 2, 7	
Coder D	4, 3, 1	
Coder E	5, 4, 2	
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. Since for each rotation the combination of stacks already ensured that each stack and each student response in it was coded by three coders, the three rotations resulted in each student response being scored by coders from nine different countries (including one from the country that did the original coding). A similar situation existed for Item Set 2, where 18 coders participated. Here, though, there were not quite enough coders for three full rotations. For Item Set 2, 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.

5.4.8 Operational Aspects of the Study

The TIMSS International Study Center prepared the necessary set of student responses for each coder participating in the study. Thus, within daily guidelines specifying which three to five items were to be scored each day, each coder was able to work at his or her own pace and the International Study Center could rest assured that the coding sequences were being maintained in accordance with the study design. The International Study Center engaged two supervisors from the United States TIMSS free-response coding effort to act as the table leaders for the two groups. They began each coding session by giving their group of coders an overview of the work for the day. Then, each group of coders began with a particular item. Within the group, in accordance with the design shown in Table 5.6, each coder had possession of the 150 responses they were to score for that item. Once coders had finished coding those responses, they moved on the next item until the day's work was completed.

Codes for each stack of 50 responses were recorded on answer sheets devised by the International Study Center that included the booklet and item number, the country and student identification numbers for the responses, the coder's identification, and the codes given to each response. After scoring an item, the coder submitted the answer sheets to a clerk so they could be checked for completeness and to ensure that the codes were within the range valid for the item. This quality control step was conducted throughout the study.

The data from the coding study were entered by a professional data entry agency. The entry process and the database were subjected to a series of quality control checks, including the accuracy of data entry and any appropriate recoding necessitated by the use of special within-country codes by some coders.

5.5 THE RESULTS OF THE INTERNATIONAL CROSS-COUNTRY CODING STUDY

5.5.1 PERCENT AGREEMENT

To provide direct comparisons with the results obtained from the within-country reliability studies, the International Study Center computed the percentage exact agreement for both correctness scores and diagnostic codes. The entire student sample of 350 responses for each free-response item was used to compute these measures. All coder pairs who coded a common subset of at least one stack of 50 student responses contributed to the overall percent agreement measure for each item. In the cross-country study design, each student response was coded by 7 to 18 coders; the measure of percent agreement obtained reflects an overall pairwise percentage of agreement based on all possible coder pairs for each item. As a result of the study design, nearly all of the across-coder comparisons included in the percent agreement calculations are across-country comparisons. For the items coded by both groups of coders, the percent agreement measures include approximately 2% comparisons between coders from the same country.

Table 5.7 summarizes the average percent agreement across the 31 items used in the international reliability study and compares these results with the corresponding within-country measures for the same items. The within-country measures are reported as the average and the range of percent agreement measures across the 26 countries submitting within-country reliability results.

Table 5.7
Average Percent Exact Agreement for All Items in International Free-Response Coding Reliability Study

Subject Area	Number of Items[1]	Average Correctness Score Agreement				Average Diagnostic Code Agreement			
		International Study	Within-Country Study[2]			International Study	Within-Country Study[2]		
			Average	Min	Max		Average	Min	Max
Mathematics	14	97%	98%	92%	100%	89%	93%	83%	99%
Science	17	87%	95%	82%	100%	71%	86%	63%	99%
OVERALL AVERAGE	31	92%	96%	87%	100%	80%	90%	73%	99%

[1] Includes four math and one science 2-part items. Percent Agreement was computed separately for each part, and each part was treated as a separate item in computing averages and ranges.
[2] Average and range of within-country percent exact agreement results from 26 countries reported in Tables 5.2 & 5.3.

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

A high overall average exact agreement of 92% for correctness scores and 80% for diagnostic codes was obtained for the 31 items in the international study. These are somewhat lower than the corresponding average within-country results of 96% and 90%. For the mathematics items, high average percent agreements of 97% for correctness score and 89% for diagnostic code were obtained, which compare very favorably with the respective average within-country measures of 98% and 93%. The science items, which in general use more complex coding rubrics than the mathematics items, had average percent agreement values that were lower for both the international and within-country studies. In addition, the difference between the two studies was greater for the science items, with average cross-country percent agreement measures of 87% and 71% compared with within-country measures of 95% and 86% for diagnostic code and correctness score, respectively. Although the average international study results are lower than the corresponding within-country measures, they still fall well within the range of within-country results obtained across all 26 countries. Moreover, they exceed by a substantial margin the correctness score agreement threshold of 70% used to identify items exhibiting within-country coding reliability problems.

More detail about the percent agreement measures for each item in the international study is shown in Tables 5.8 and 5.9 for mathematics items and science items, respectively. These tables also show the total number of individual comparisons used in computing the cross-country percent agreement measures, since this number varies for different items due to the rotation design and the division of items and coders into two sets of unequal size. The percent agreement for the twelve items that were coded by both groups of coders was first computed for the two sets separately. A comparison of the two measures revealed a maximum difference of less than 5% for any item. The differences in most cases reflected a slightly higher percent agreement for the group to which the item was originally assigned. Since the differences between coder groups was small, the calculations for these twelve items include comparisons for coders in both groups to obtain the broadest across-country comparisons possible.

The percent of exact agreement for each mathematics item was very high, with only two items having correctness score agreement measures below 90%. Diagnostic code agreements were, in general, lower, ranging from 61% to 98%. Even for the diagnostic agreement, however, 13 out of 18 items had percent agreement greater than 90%. For the majority of items, the difference between the international and within-country average diagnostic code percent agreement measures was 5% or less, and for all but two items, the international measure fell within the range of within-country values. For the correctness score agreement, all items were well within the range of the within-country results. The percent of exact agreement for science items was, in general, lower and exhibited a much broader range, with diagnostic code agreement ranging from 50% to 98%. When only the first-digit correctness score is considered, a percent exact agreement range of 72% to 99% is observed. Even the items with the lowest international diagnostic agreement have correctness score agreement levels that fall within the range of within-country measures and exceed the 70% threshold. Also, only a few of the science items had large diagnostic code agreement differences between the international and within-country measures (20% or greater).

Table 5.8
Percent Exact Agreement for Coding of Mathematics Items

Item Label[1]	Total Valid Comparisons[2]	Correctness Score Agreement				Diagnostic Code Agreement			
		International Study	Within-Country Study[3]			International Study	Within-Country Study[3]		
			Average	Min	Max		Average	Min	Max
M1	9150	100%	99%	96%	100%	97%	97%	84%	100%
[4] M2A	46050	100%	100%	96%	100%	98%	98%	94%	100%
M3	12600	99%	99%	95%	100%	98%	97%	92%	100%
M4	46050	99%	99%	96%	100%	99%	98%	87%	100%
M5	45985	99%	100%	96%	100%	97%	98%	92%	100%
M6	12600	99%	99%	98%	100%	97%	98%	91%	100%
M7	12600	99%	99%	96%	100%	95%	98%	92%	100%
M8	12600	99%	99%	94%	100%	91%	95%	89%	100%
M9	9150	99%	99%	94%	100%	94%	97%	90%	100%
[4] M2B	46050	99%	99%	95%	100%	91%	94%	74%	100%
[4] M10A	45938	98%	100%	98%	100%	95%	97%	90%	100%
[4] M11A	12592	97%	98%	84%	100%	91%	94%	77%	100%
M12	12600	97%	99%	95%	100%	93%	95%	88%	99%
[4] M11B	12600	96%	98%	95%	100%	74%	88%	68%	100%
[4] M13A	12600	95%	97%	90%	100%	85%	92%	75%	99%
M14	12600	91%	96%	81%	100%	77%	89%	72%	98%
[4] M13B	12592	89%	96%	84%	100%	71%	88%	75%	100%
[4] M10B	46050	84%	93%	77%	99%	61%	82%	61%	97%
AVERAGE MATH ITEMS		97%	98%	92%	100%	89%	94%	83%	100%

[1] See Appendix H for item descriptions and coding guides.
[2] Values for items coded by the same coder group differ slightly due to a small number of missing responses or invalid codes.
[3] Average and range of within-country percent exact agreement results from 26 countries reported in Tables 5.2 & 5.3.
[4] Two-part items; each part is analyzed separately

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

Table 5.9
Percent Exact Agreement for Coding of Science Items

Item Label[1]	Total Valid Comparisons[2]	Correctness Score Agreement				Diagnostic Code Agreement			
		International Study	Within-Country Study[3]			International Study	Within-Country Study[3]		
			Average	Min	Max		Average	Min	Max
S1	9078	99%	99%	95%	100%	98%	97%	80%	100%
S2	46035	94%	97%	77%	100%	74%	86%	64%	100%
S3	9150	93%	96%	81%	100%	85%	91%	54%	100%
S4	12600	93%	95%	83%	100%	67%	80%	52%	99%
S5	46050	92%	97%	88%	100%	78%	88%	58%	100%
S6	46050	91%	96%	90%	100%	79%	91%	79%	99%
[4] S7A	9150	90%	95%	83%	100%	71%	87%	67%	99%
[4] S7B	9150	89%	95%	87%	100%	77%	89%	74%	98%
S8	45930	89%	96%	90%	100%	70%	84%	65%	98%
S9	46050	88%	93%	74%	100%	74%	87%	64%	100%
S10	9150	88%	96%	86%	100%	83%	91%	65%	100%
S11	9122	86%	95%	86%	100%	72%	87%	61%	100%
S12	45930	86%	95%	81%	100%	59%	80%	53%	96%
S13	46034	82%	93%	74%	100%	66%	87%	65%	100%
S14	9150	80%	93%	82%	100%	59%	82%	47%	100%
S15	46050	78%	92%	75%	100%	70%	89%	69%	99%
S16	12600	75%	91%	74%	100%	51%	78%	55%	100%
S17	9129	72%	90%	70%	100%	50%	82%	59%	100%
AVERAGE SCIENCE ITEMS		86%	94%	81%	100%	70%	86%	62%	99%

[1] See Appendix H for item descriptions and coding guides.

[2] Values for items coded by the same coder group differ slightly due to a small number of missing responses or invalid codes.

[3] Average and range of within-country percent exact agreement results from 26 countries reported in Tables 5.2 & 5.3.

[4] Two-part items; each part is analyzed separately

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

5.5.2 Investigating Differences Between the Cross-Country and Within-Country Studies

Although the cross-country percent agreement measures are systematically somewhat lower than those reported from the within-country samples, it should be remembered that these results also reflect differences between the two types of studies. First, for many coders the student responses used in the international study were in a different language and reflected a different culture from those encountered by the coders in their own country. Both of these factors added to the complexity of coding in the cross-country study. Second, while the within-country measures were obtained during the actual coding sessions in each country, when the training and guides were fresh in the coders' minds, for most participants the international study was conducted several months after the main study coding sessions. Although each coder involved in the international study received some refresher training before participation, the potential for coder agreement might well have decreased after the main study coding sessions. Third, the coding environment in the international study was somewhat different from those within the individual countries. For example, several countries indicated that during their country's coding sessions, coding difficulties encountered with some student responses were resolved by consensus, which was not the case in the international study. Also, coders in the international study had to work with photocopies rather than the original booklets.

To investigate differences between the percentage of agreement reported for the two types of reliability studies, the 12 items that were coded by both groups of coders were used to determine the percent diagnostic code agreement between the two coders from each country. The average of these measures for the 18 countries that sent two coders to the international study was used as a measure of the percent agreement obtained under the conditions of the international study, excluding any across-country coder effects. Table 5.10 provides a comparison of this within-country measure from the international study with both the across-country measure and the average percent agreement from the within-country reliability studies conducted in 26 countries.

Table 5.10
Comparison of Diagnostic Code Agreement from the International and Within-Country Studies

	International Study		Within-Country Studies Average[3]
	Average Within-Country Percent Agreement[1]	Overall Across-Country Percent Agreement[2]	
Mathematics Items			
M2A	98%	98%	98%
M2B	91%	91%	94%
M4	99%	99%	98%
M5	98%	97%	98%
M10A	94%	95%	97%
M10B	60%	61%	82%
Mathematics Average	90%	90%	95%
Science Items			
S2	76%	74%	86%
S5	80%	78%	88%
S6	81%	79%	91%
S8	70%	70%	84%
S9	77%	74%	87%
S12	62%	59%	80%
S13	68%	66%	87%
S15	70%	70%	89%
Science Average	73%	72%	86%
Overall Average	80%	80%	90%

[1] Average of percent agreement between the two coders from each country for 18 of the countries in the international study in Table 5.5.

[2] Percent agreement from the international study based on all coder comparisons.

[3] Average of within-country percent exact agreement results from 26 countries reported in Tables 5.2 & 5.3.

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

These results show that the within-country and across-country percent agreement measures from the international study are comparable for all items, and that both are lower than the corresponding measure from the within-country study. The systematically lower percent agreement of the international study thus appears to be due primarily to situational and contextual differences in the way the two measures were obtained rather than to decreased across-country coding reliability. Based on these results, the reliability of the international free-response coding from the main study coding sessions would be expected to be no lower than that from the within-country results, and therefore, should be quite good for most items.

5.5.3 Comparing Code Agreement for Individual Countries

Another measure of across-country agreement is to compare how well the coders from each of the participating countries agree with the other coders in the same group. Table 5.11 presents the diagnostic code percent exact agreement between each participating coder and the coders from other countries in the same group.

These results, averaged across all items in the item set, reveal that there is a good level of consensus within each group of coders and that the agreement for individual coders is quite comparable across countries. For Item Set 1, the individual coder agreement ranges from 77% to 85% with an average of 82%. The agreement for Item Set 2 is somewhat lower, ranging from 71% to 80%, with an average of 77%. These differences in agreement within the two sets of coders, however, reflect the nature of the items assigned to them, with Item Set 1 being predominantly mathematics and Item Set 2 predominantly science, which were more complicated to code.

Table 5.11 Comparison of Exact Percent Diagnostic Code Agreement for Individual Countries[1]

Country	Item Set 1[2] 12 Math 4 Science	Item Set 2[3] 13 Science 2 Math
Australia	84%	77%
Bulgaria	79%	76%
Canada	84%	*
England	84%	78%
France	84%	*
Germany	82%	*
Hong Kong	83%	75%
Ireland	83%	76%
Lithuania	79%	79%
Latvia (LSS)	82%	75%
Norway	84%	80%
New Zealand	83%	80%
Philippines	77%	76%
Portugal	83%	74%
Romania	83%	76%
Russian Federation	79%	76%
Slovak Republic	80%	71%
Singapore	83%	79%
Sweden	81%	77%
Switzerland	79%	78%
United States	85%	79%
AVERAGE	82%	77%

[1] Percent agreement between each coder and the coders from all other countries averaged over all items in the item set.
[2] Items in Set 1: M2, M3, M4, M5, M6, M7, M8, M10, M11, M12, M13, M14, S2, S4, S13, S16
[3] Items in Set 2: M1, M9, S1, S3, S5, S6, S7, S8, S9, S10, S11, S12, S14, S15, S17
*No Item Set 2 coders from Canada, France and Germany.

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

5.5.4 FREQUENCIES OF DIAGNOSTIC CODE AGREEMENT

Contingency tables showing the cumulative frequencies of pairwise code combinations were computed for all items in the international study. These tables can be used to obtain detailed information about the nature of code discrepancies. Contingency tables and coding guides for all items in the international study are included in Appendix H. An example contingency table and the corresponding item and coding guide are shown in Figure 5.1 for item S11. This item was one for which the percent diagnostic code agreement was moderate, indicating that a number of code discrepancies are expected.

The cell counts in the contingency table in Figure 5.1 indicate the total number of times, over the entire set of student responses, that for the same student response one coder in a pairwise comparison gave the code corresponding to the row position, while another coder gave the code corresponding to the column position. The simple percent exact agreement for both the diagnostic code and the correctness score may be computed from these frequencies of nominal agreement. The diagnostic code percent exact agreement is computed from the sum of the diagonal cells, where the two paired codes match exactly, while the correctness score percent exact agreement is computed from the sum of all the shaded cells, where the first digits of the paired codes correspond to the same correctness score.

The TIMSS free-response coding guides are specific to each item; the coding guide for item S11 is shown as an example. This coding guide has 8 valid diagnostic codes, 3 that correspond to a correct response (first digit of 1) and 5 that correspond to an incorrect response (3 codes with a first digit of 7 for an incorrect response and 2 codes with a first digit of 9 for a nonresponse). A second digit of 9 (code 19 or 79) is used for responses that are judged to be within the level of correctness indicated by the first digit but do not fit any of the other specific diagnostic codes. The level of disagreement about specific diagnostic codes varies substantially from item to item, but there are some patterns of code disagreement that are common to many of the items. Some of these types of patterns can be observed with the S11 example item.

Item S11 had a diagnostic code percent agreement of 72% and a correctness score percent agreement of 86%. The frequencies of matched codes indicate that approximately 10% of the code comparisons reflect diagnostic code disagreements where two paired codes on a student response are either both correct or both incorrect, but only one coder used a specific diagnostic code (10,11 or 70,71), while the other used an Other code (19 for Other Correct or 79 for Other Incorrect). Another 3% of code discrepancies are due to student responses that were coded as correct but where there was disagreement on whether the 10 or the 11 diagnostic code was given. The most common code discrepancies contributing to the lack of correctness score agreement, approximately 12%, is due to student responses that one coder scored as correct (10, 11, or 19), while another coder gave a code of Other Incorrect (79). This types of code discrepancy was found to be fairly common across many of the items investigated in the international study, with the use of the Other codes accounting for a substantial portion of disagreement in both the diagnostic code and the correctness score. These code disagreements did not usually result in low overall correctness score agreement, however. Another type of discrepancy that can reduce diagnostic code agreement is the interpretation of what constitutes a nonresponse (code 90 or 99). Although the 99 code was to have been reserved for absolutely blank responses, sometimes very brief partial responses also were given a code of 99. In most instances, however, the disagreements were between the 90 and the 79 codes. The extent of that disagreement is understandable given that both codes reflect types of incorrect responses that can be difficult to interpret. For all items, less than 3% of code comparisons reflected disagreements involving the 90 or 99 codes.

Figure 5.1
Item Description, Frequencies of Diagnostic Code Agreement and Coding Guide for Example Item S11

Carbon Dioxide Fire Extinguishers **Item S11**

Carbon dioxide is the active material in some fire extinguishers. How does carbon dioxide extinguish a fire?

	FREQUENCIES OF MATCHED CODES								
	10	**11**	**19**	**70**	**71**	**79**	**90**	**99**	**ROW SUM**
10	3508	297	340	2	42	542	18	0	4749
11		490	116	29	59	365	1	0	1060
19			43	3	16	165	4	0	231
70				296	10	126	3	0	435
71					265	286	10	0	561
79						1086	89	0	1175
90							111	36	147
99								764	764

TOTAL VALID COMPARISONS	**9122**

Coding Guide	
Code	**Response**
Correct Response	
10	Mentions that carbon dioxide keeps oxygen away; response includes explicit reference to oxygen.
11	Mentions that carbon dioxide keeps “air” away.
19	Other correct.
Incorrect Response	
70	Mentions that carbon dioxide cools down the fire.
71	Refers to a material in carbon dioxide.
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.

5.5.5 Generalizability of Free-Response Item Scores

An analysis of variance of the international reliability study data was used to estimate generalizability coefficients for both the country-level average scores and the student-level scores on each of the free-response items in the international study. The generalizability coefficients computed are a measure of the reliability of the free-response item scores in that they reflect the proportion of observed variance due to true score variance for the object of measurement. In the computation of generalizability coefficients, specific sources of error variance are identified according to the design of the study and the definition of the object of measurement. Generalizability coefficients computed for each of the items are shown in Table 5.12 for mathematics items and Table 5.13 for science items.

For the country-level averages, the object of measurement is the average score for each of the seven English-test countries contributing student responses. The relative error variance has contributions from both the variance due to students within countries and the variance due to rater effects (both main and interaction effects). The generalizability coefficient reflects 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 and the total number of raters involved in rating the entire set of student responses in each country. Generalizability for all items increases as each of these levels increases, but for the typical sample sizes used in the TIMSS study, 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 contribution to total variance due to country, student, and rater effects.

In Tables 5.12 and 5.13, generalizability coefficients 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 5.12
Generalizability of Scores on Free-Response Mathematics Items Based on the International Reliability Study Sample

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 5.13

Generalizability of Scores on Free-Response Science Items Based on the International Reliability Study Sample

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	
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.

For the student-level scores, the object of measurement is the individual student score, and the relative error variance is due to the effect of the interaction between raters and students. The generalizability coefficient reflects the reliability of an individual student's score on an item, given that each student response receives only one rating. The person-by-rater interaction effect was found to vary substantially from item to item, particularly for the science items. The variance due to the person-by-rater interaction ranged from as low as 1% to as high as 50% of the total variance in student scores. This is reflected in the generalizability coefficients observed across the science items, which range from 0.42 to 0.99. Despite a low generalizability for a few items, for 11 out of 18 science items the student score generalizability was above 0.70. For mathematics items, the rater effects were much lower and the generalizability of the student scores was quite high for all but one of the items. Even for some of the science items with low individual score generalizability, however, the generalizability of the country-level averages was still quite high as it is based on a large number of student responses and raters. 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 free-response items. These results serve as a caution, however, in performing secondary analyses that involve making any generalizations from individual student scores on specific items.

5.6 SUMMARY

Within resource constraints facing both the individual countries and the International Study Center, TIMSS has put considerable energy into the use of free-response items. Approximately one-third of the students' response time is devoted to free-response questions, which across the TIMSS tests encompasses about 300 free-response items. To provide diagnostic information about achievement, test development included an extensive effort to design scoring guides tailored to each of these questions. In the TIMSS two-digit scoring approach, the first digit indicates the correctness score (including levels of partial credit) and the second digit provides diagnostic information about the specific type of response.

Considering the number of items, number of countries, number of testing languages, and number of students involved, the scope of the free-response scoring effort was complex by anyone's standard. Therefore it was important for TIMSS to emphasize the importance of reliable scoring procedures. This includes very careful attention both to using reliable scoring procedures and to conducting studies to document the success of the procedures used.

Planning for TIMSS data collection included an ambitious series of training sessions for participating countries. The International Study Center conducted regional training sessions of essentially one week each to assist representatives of the national centers who would then be responsible for training personnel in their countries to apply the two-digit codes reliably. Nine sessions were held in total to accommodate participation by all countries in accordance with the different schedules in the Southern and Northern hemispheres. During the training sessions, participants were given detailed information about how to conduct free-response scoring and opportunities to practice the procedures, including substantial time in practicing scoring actual student responses according to the TIMSS guides.

The results from the within-country scoring reliability studies indicate that the percent of exact agreement among coders was very high, especially considering the many challenges underlying the effort. Each country was required to collect information about the reliability of its scoring procedures by having 10% of the student responses scored independently by two coders. Not all countries were able to afford this effort, but 26 countries provided data about the reliability of their scoring procedures. The average percent of exact agreement for the correctness score within each of the countries ranged from 97% to 100% on the mathematics items and from 88% to 100%. Average percentages of exact agreement for the diagnostic codes also were quite respectable, ranging from 89% to 99% for mathematics items and from 73% to 98% for science items.

The results of the international reliability study conducted using student responses from Population 2 also revealed a very high degree of across-country agreement among coders. Based on 350 student responses to each of 31 mathematics and science items, a total of 39 coders from 21 countries participated in the cross-country reliability study conducted by the International Study Center. The student responses used were randomly sampled from the within-country reliability samples of seven English-test countries: Australia, Canada, England, Ireland, New Zealand, Singapore, and the United States. A high overall average percentage of exact agreement of 92% for correctness scores and 80% for diagnostic codes was obtained.

In addition to documenting the high quality of the TIMSS free-responses scoring, various comparisons of the results from the TIMSS within-country and cross-country reliability studies reveal some interesting findings. Agreement was systematically higher for the mathematics items than for the science items. This seems reasonable, given that the coding guides for the mathematics items tended to be more straightforward. The results also indicate somewhat less agreement across countries than within countries, although further analyses reveal that these differences may be attributed primarily to differences in the conditions of the two types of studies. For example, for the international study many coders were not evaluating student responses in their native languages, so translation and cultural issues most likely made interpretation of responses more difficult. Also, for some coders several months had passed since the scoring effort in their own countries, and the coding task might not have been as familiar during the international study despite refresher training.

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 based on using data from the TIMSS free-response items. The generalizability of country-level averages is quite high for most of the items, 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 and not individual-level results, the lower generalizability for individual scores is not a concern for reporting free-response item averages. In fact, all the TIMSS data from the reliability studies indicate that the scoring procedures were very robust both within and across countries. At least from the perspective of the quality of the free-response scoring, TIMSS can report the international achievement results with confidence.

REFERENCES

Mullis, I.V.S., Jones, C., and Garden, R.A. (1996). "Training Sessions for Free-Response Scoring and Administration of Performance Assessment" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995). *Guide to Checking, Coding, and Entering the TIMSS Data* (Doc. Ref.: ICC918 / NRC449). Chestnut Hill, MA: Boston College.