Lie, S., Taylor, A., and Harmon, M. (1996) "Scoring Techniques and Criteria" in M.O. Martin and D.L. Kelly (eds.), *Third International Mathematics and Science Study (TIMSS) Technical Report, Volume I: Design and Development*. Chestnut Hill, MA: Boston College.

7. Scoring Techniques and Criteria

Svein Lie
Alan Taylor
Maryellen Harmon

7.1 OVERVIEW

Free-response items play an important role in the TIMSS test design for measuring student achievement in mathematics and science. While many multiple-choice items effectively measure content and process outcomes over a range of cognitive behavior levels, they give little information about the procedures and thought processes students use in solving problems in mathematics and science. Free-response items are thus intended to supplement multiple-choice items, in an attempt to reflect the complex and multistage processes involved in mathematical and scientific thinking.

Analysis of student responses to free-response achievement items can provide valuable insights into the nature of student knowledge and understanding. The case for including free-response items in the international item pool was made by Taylor (1993a), who also noted the implications for coding reliability and the need for resources. He states:

> The inclusion of free-response items in the international item pool provides an opportunity to collect a rich source of data related not only to levels of student achievement but also to the method used by students in approaching a problem, and to the misconceptions or error types which may be demonstrated by them. Inherent in the collection of these data, however, are issues of reliability and need for additional resources in the coding process (p.1).

The TIMSS tests employed several item formats. These included multiple-choice, short-answer, and extended-response items, as well as performance tasks. The first three item types were included in the written assessment administered to all sampled students; performance tasks, were administered to subsamples of students in Populations 1 and 2. Test blueprints for the written tests allocated approximately 30 percent of testing time to free-response items for each of the three student populations. The distinction between the two free-response categories, short-answer and extended-response, was made mainly for reasons of time; the two did not differ sharply in rationale or information sought. The performance assessment component of TIMSS also required students to provide written answers to the test items in the tasks they completed. These were also considered to be free-response items, and were coded accordingly.

If student responses to free-response items were scored for correctness only, it would be sufficient for the purpose of aggregating results with corresponding multiple-choice items to develop test and subtest scores. But that would yield no information on how students approached problems. TIMSS therefore developed a special coding system that provides diagnostic information in addition to information about the correctness of student responses.

This chapter presents an overview of the development of the TIMSS coding systems for scoring free-response items. The development of the TIMSS two-digit coding system is discussed first, followed by exemplar coding rubrics for several free-response items in the TIMSS tests and for a performance task.

7.2 DEVELOPMENT OF THE TIMSS CODING SYSTEM

The TIMSS coding system was developed over several years, beginning with the early stages of the study when it was recognized that a coding system for both correctness and students' approaches and misconceptions was desirable.

7.2.1 THE 1991 PRE-PILOT STAGE

The first opportunity to code free-response items occurred with the 1991 Pre-Pilot test. The *Pre-Pilot Manual for National Research Coordinators* (Marshall et al., 1991) included directions for the translation of the items, administration of the instruments, data processing, and coding. Items were intended only as samples for the purpose of exploring the free-response scoring methodology. Codes used for free-response items at this stage of the project dealt with three types of information each for mathematics and science.

For mathematics, the three information categories were "answer," "implementation strategy," and "problem-solving strategy." The options within each category were the same across all items. For example, the "answer" category included four options: blank, correct, incorrect, and undetermined. "Implementation strategy," on the other hand, contained the following: no work shown, complete, incomplete, misinterpreted, and not clear. There were

11 options for "problem-solving strategy" (inspired by Polya's well known classification of such strategies), including typical approaches used by students such as systematic list, guess and check, work backwards, and the like.

Rubrics for science, on the other hand, were both generic and specific to each item. They included the number of ways students approached a question, what type of logic they demonstrated, how the approach was specific to the question, and the extent to which an answer was complete. The first two and the last of these were generic, whereas the second was specific to each question. For example, among the generic options under the type of logic were the following: no response, logical and appropriate, logical but not appropriate, not logical, and ambiguous.

In a review of their results from the pre-pilot, the Scandinavian TIMSS groups (Brekke et al., 1992) identified several issues stemming from the free-response codes proposed at that time. Among their recommendations for free-response item coding were the following.

- As a criterion for selecting of free-response items, the time required to code an item should be related to the value of information obtained
- The set of codes for an item should be based on empirical evidence
- Codes should relate to specific answers or strategies rather than generic types or categories of response
- Diagnostic information should in some cases be included in the rubrics themselves, to help coders understand how students might have reasoned
- The coding guide should in many cases give precise examples of responses that belong to a certain code

Information from this review helped to provide direction for the further development of coding rubrics in TIMSS.

7.2.2 THE 1992 PILOT

As a preparation for the subsequent item pilot, a group of countries voluntarily reviewed and rated a number of extended-response items. The selected items were piloted in early 1992 in these countries. Each country then grouped the responses to each task into a response "typology," or category, for classification of the most common responses. Thus, the different countries' typologies could be compared. A rather complicated meta-analysis was undertaken based on these data (Wiley, 1992), with the goal of obtaining codes that could provide rich information on student thinking throughout the world. However, the construction of such codes seemed to be a very complex task, and development along these lines was discontinued.

7.2.3 THE 1993 ITEM PILOT AND REVIEW

As a preparation for the 1993 item pilot, Taylor (1993a) proposed that information gathered from free-response items focus on three aspects of student response: degree of correctness, method or approach, and misconception or error type. The rubric for correctness ranged from zero to the number of score points for an item. Numbers for the other rubrics corresponded to major approaches or error types identified for each question. An illustration of these rubrics is shown below in Table 7.1.

Table 7.1 Coding Rubrics in the 1993 Item Pilot

Degree of Correctness	Method or Approach	Misconception/Error Type
0 - no work leading to answer	0 - no work	0 - no error type
1 - step 1 toward answer	1 - approach 1	1 - error type 1
2 - step 2 toward answer	2 - approach 2	2 - error type 2
x - correct answer	y - approach y	z - error type z
	y+1 - other	z+1 - other

The number of points for each rubric varied by item since each was unique in terms of answer, approach, and types of misconceptions generated. Although these aspects are similar to those used at the final stage of data collection for TIMSS, the design at this point suggested a separate set of codes for each aspect. Taylor proposed that the descriptions of codes within each rubric be based on student responses from the item pilot. He also made suggestions for the composition of coding committees, for training to improve inter-rater reliability, and for feedback on item responses.

Following acceptance by NRCs of the direction proposed by Taylor (1993a), a manual for coding free-response items was developed for use in the 1993 item pilot (Taylor 1993b). The manual included directions for establishing coding committees, training procedures, and coding reports. Instruments included exemplar rubrics for coding, item review forms, mark allocation forms, and correctness rubrics for all free-response items. The national centers coded student responses collected in the item pilot according to the correctness rubrics. In addition, some countries volunteered to report for each free-response item the most common responses, approaches and/or error types. Further, sample student papers were used to develop the coding rubrics.

7.2.4 THE 1994 FIELD TRIAL AND THE NORWEGIAN INITIATIVE

Plans for the 1994 field trial did not include construction of coding rubrics for more than just correctness. Due to a shortage of time and resources, student responses from the field trial were coded by score points only. At this time the Norwegian national center initiated an effort to develop a set of richer coding rubrics for the main survey, in line with the earlier work. Individuals at the TIMSS International Coordinating Center (Alan Taylor, Ed Robeck, Ann Travers, and Beverley Maxwell) were involved in many discussions that led to the Norwegian proposal.

A series of discussion papers on free-response coding was prepared by Carl Angell and Truls Kobberstad (Angell, 1993; Kobberstad, 1993; Angell & Kobberstad, 1993). The first two papers were prepared for a meeting of the TIMSS Subject Matter Advisory Committee in September 1993, and the third for the October 1993 meeting of the TIMSS NRCs. In these papers it was proposed that a system of two-digit coding be employed for all free-response items. The first digit, ranging between 1 and 3, would be used for a correctness score, and the second digit would relate to the approach used by the student. Numbers between 70 and 79 would be assigned for different categories of incorrect response attempts, while 90 would be used if the student did not respond. The papers also presented a number of exemplar mathematics (Population 2) and physics (Population 3) items. The rubrics were described and applied on student responses. Further, some promising results were reported on inter-rater reliability using this method of coding.

The Subject Matter Advisory Committee supported the proposal for two-digit coding and recommended that an international coding committee be established to develop final versions of coding rubrics prior to final administration of the instruments. Subsequently, in 1994 the International Study Director established the Free-Response Item Coding Committee (FRICC), the purpose of which was to develop coding rubrics for the free-response items in the TIMSS tests. The FRICC included, representatives from 11 countries: Jan Lokan, (Australia); Alan Taylor, (Canada); Peter Weng, (Denmark); Josette Le Coq, (France); Nancy Law, (Hong Kong); Algirdas Zabulionis, (Lithuania); Svein Lie (chair), (Norway); Galina Kovalyova, (Russian Federation); Vladimir Burjan, (Slovak Republic); Kjell Gisselberg, (Sweden); Maryellen Harmon, (USA); Curtis McKnight, (USA); and Senta Raizen, (USA). In addition to the formal members, the following individuals made substantial contributions to the FRICC activities: Truls Kobberstad, Carl Angell, Marit Kjaernsli, and Gard Brekke from Norway, and Anna Hofslagare from Sweden.

7.3 DEVELOPMENT OF THE CODING RUBRICS FOR FREE-RESPONSE ITEMS

In order to capture the richness of the intended information efficiently and reliably, the FRICC established a set of criteria to which the coding rubrics should adhere. The TIMSS rubrics should do the following.

- Permit scoring for correctness and capture the analytical information embedded in student responses.
- Be clear, distinct, readily interpretable, and based on empirical data (student responses obtained from pilot or field trials) so as to account for the most common correct responses, typical errors, and misconceptions.
- Be capable of encoding the adequacy of an explanation, justification, or strategy as well as the frequency with which it is used.

- Be simple, in order to get high reliability and not to impose unreasonable time or resource burdens.
- As far as possible, allow for nuances of language and idiosyncratic features of various countries but avoid being so complex that coders are overwhelmed and tend to limit themselves to a few stereotypical codes.
- Have a number of codes that is not excessive, but sufficient to reduce coding ambiguity to a minimum.

The task of the FRICC was to develop scoring rubrics that could be efficiently and consistently applied, and that were based on empirical evidence in a number of countries. The Norwegian team, on the basis of a detailed review of student responses to items from the field trial in Norway, developed a draft set of rubrics for consideration by the FRICC (Angell et al., 1994). Committee members began their work by analyzing Population 1 and 2 results from the field trial in each of their countries, applying the draft rubrics prepared by researchers at the Norwegian national center (Kjaernsli, Kobberstad & Lie, 1994). In July 1994, the committee arrived at descriptors by response category for each rubric for the items in those tests. The criteria used in the development of the codes and the draft coding rubrics were presented to and approved by the NRCs in August 1994. A number of achievement items were modified following the 1994 field trial, and some of these were administered to a convenience sample in the Scandinavian countries in August 1994. Student responses were then used in the development of coding rubrics for those items.

After the coding rubrics had been developed, the International Study Center assembled the coding manuals for distribution to the participating countries (TIMSS 1995a, 1995b). The manuals included the coding rubrics developed by the FRICC and, for many items, example student responses corresponding to the appropriate codes.

This process was repeated for the Population 3 items in November-December 1994. For this effort, the work of the FRICC was based on draft codes prepared by Vladimir Burjan (advanced mathematics), Carl Angell (physics), Truls Kobberstad (mathematics literacy), and Kjell Gisselberg (science literacy). Again, additional piloting was carried out for modified items in order to ensure that the coding rubrics would represent common student responses, approaches, and misconceptions.

7.4 DEVELOPMENT OF THE CODING RUBRICS FOR THE PERFORMANCE ASSESSMENT TASKS

While the FRICC established the TIMSS two-digit coding system and developed coding rubrics for the free-response items, the Performance Assessment Committee (PAC) collaborated to develop the coding guides for the performance assessment tasks. Led by Maryellen Harmon (United States) and Per Morten Kind (Norway), in 1994 the PAC developed the initial coding rubrics for the performance assessment tasks administered in the performance assessment field trial. Countries participating in the field trial coded the student responses to the items within each of the tasks. The ensuing data were used to evaluate the tasks for suitability for use in the main survey. These initial coding rubrics served as the basis for the codes developed for the main performance assessment study.

Using the responses from the field trial and from additional piloting of the tasks in Norway and the United States, Harmon and Kind, with the assistance of the PAC, developed rubrics for each item within the tasks selected for the main survey. Like the codes for the free-response items, the codes for the performance assessment tasks were developed to include the common correct responses and the common misconceptions of students. An additional feature of the performance assessment coding rubrics was a set of criteria for what a correct response should include, as well as additional information to aid the coder in evaluating the response.

Following the development of the codes, the International Study Center assembled the *Coding Guide for Performance Assessment* (TIMSS, 1994a), which included the possible codes and examples of the most common responses to each item in all tasks. To facilitate the coding effort in the participating countries, the International Study Center also prepared the *Supplement to the Coding Guide for Performance Assessment* (TIMSS, 1995c). This included a full set of example student responses to all tasks.

7.5 THE NATURE OF FREE-RESPONSE ITEM CODING RUBRICS

The TIMSS coding system is demonstrated in Table 7.2 with a generic example of the coding scheme for a free-response item worth one score point. Actual coding rubrics for actual items are presented later, as is an example of a performance task.

Table 7.2 TIMSS Two-Digit Coding Scheme

Code	Text
10	correct response, answer category / method #1
11	correct response, answer category / method #2
12	correct response, answer category / method #3
19	correct response, some other method used
70	incorrect response, common misconception / error #1
71	incorrect response, common misconception / error #2
76	incorrect response, information in stem repeated
79	incorrect response, some other error made
90	crossed out / erased, illegible, or impossible to interpret
99	blank

Student responses coded as 10, 11, 12, or 19 were correct and earn one score point. The type of response in terms of the approach used or explanation provided is denoted by the second digit. A response coded as 10 demonstrates a correct response of answer type #1 or method #1. For items worth more than one score point, rubrics were developed to allow partial credit and to describe the approach used or explanation provided.

Student responses coded as 70, 71, 76, or 79 were incorrect and earned zero score points. The second digit in the code represents the type of misconception displayed, incorrect strategy used, or incomplete explanation given. A code of 76 was assigned to an incorrect response in which the student merely repeated information from the item stem. In addition, countries had the option of assigning country-specific codes for correct and incorrect responses in cases where the international rubrics failed to allow for common responses. For the international analyses, the country-specific codes were recoded to 19 and interpreted as “other correct” for items worth one point (to 29 and 39 for items worth two and three points respectively), or to 79 and interpreted as “other incorrect” response.

Student responses coded as 90 or 99 also earned zero score points. A 90 indicates that a student attempted the item but did not provide a coherent response. A 99 indicates that the student did not attempt the item. The differentiation between 90 and 99 allows for the identification of a series of totally blank items towards the end of the test (deemed “not reached”) versus items a student has attempted but failed to answer.

The three examples of free-response items shown below illustrate how these rubrics corresponded to specific items and provided diagnostic information on item-specific features. The fourth example demonstrates the application to an item in a performance task.

Figure 7.1 presents a science item and its coding guides. Because this item was administered to all three student populations, the coding rubrics were developed to accommodate a wide range of responses. Correct responses were coded 10, 11, 12, or 13 depending on the type or response or method employed. The most common misconceptions are covered by codes 70 (drinking makes us cool down), 71 (you dry out, particularly in your throat), and 72 (you drink to get energy).

The coding guide for this item allow a detailed study of students' conceptions, at different ages and in different countries, of water balance and temperature regulation of the human body.

Figure 7.1 Exemplar Coding Guides — Thirsty on a Hot Day

O16. Write down the reason why we get thirsty on a hot day and have to drink a lot.

Code	Response
Correct Response	
10	Refers to perspiration and its cooling effect and the need to replace lost water.
11	Refers to perspiration and only replacement of lost water. *Example: Because when we are hot, out body opens the pores on our skin and we lose a lot of salt and liquid.*
12	Refers to perspiration and only its cooling effect.
13	Refers to perspiration only. *Examples: We are sweating.* *Your body gives away much water.* *We are sweating and get drier.*
19	Other acceptable explanation.
Incorrect Response	
70	Refers to body temperature (being too hot) but does not answer why we get thirsty. *Example: You cool down by drinking something cold.*
71	Refers only to drying of the body. *Examples: Your throat/mouth gets dry.* *You get drier.* *The heat dries everything.*
72	Refers to getting more energy by drinking more water. *Example: You get exhausted.*
76	Merely repeats the information in the stem. *Examples: Because it is hot.* *You need water.*
79	Other incorrect: *Example: You loose salt.*
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret
99	BLANK

The mathematics item displayed in Figure 7.2 was administered to students in Population 2. It is a simple equation, the solution of which is straightforward for those who know the algorithm. Calculation errors are covered by codes 70 (clear indication of student confusing addition and subtraction) and 71 (other calculation errors), whereas code 72 covers responses that reach no numeric solution for x.

With this set of codes, in spite of its simplicity, one can analyze not only students' knowledge of and ability to apply the algorithm for solving a linear equation, but also the frequency of the most common errors.

Figure 7.2 Exemplar Coding Guides — Solve for X

L16. Find x if $10x - 15 = x + 20$

Answer: ______________________________

Code	Response
Correct Response	
10	7
Incorrect Response	
70	1 OR 2.33 OR 3
71	Other incorrect numeric answers.
72	Any expression or equation containing *x*.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret
99	BLANK

Figure 7.3 presents the set of codes for a science item that deals with the conservation of mass during melting. This item was administered to students in Population 2. One score point is given for a correct answer (no change of mass), and two score points are given for an explanation that refers to the principle of conservation of mass. Among the incorrect responses there are codes for increase of mass (70 and 71) and decrease of mass (72 and 73). Further, the codes allow us to compare countries on how many responses include any explanation, and to determine whether a response is fully correct (code 20) or not (codes 10, 70, or 72).

Figure 7.3 Exemplar Coding Guides — Melting Ice Cubes

A glass of water with ice cubes in it has a mass of 300 grams. What will the mass be immediately after the ice has melted? Explain your answer.

Note: For this question do not distinguish if the student substitutes kg for g: that is, accept 300 kg as the same as 300 g.

Code	Response
Correct Response	
20	300 g with a good explanation. *Examples: 300 g. The ice changes into the same amount of water.* *The same. The ice only melts.* *The same weight. Nothing disappears.*
Partial Response	
10	300 g. Explanation is inadequate.
11	300 g. No explanation.
Incorrect Response	
70	More than 300 grams with explanation. *Examples: More. Water has higher density.* *More. Water is heavier than ice.*
71	More than 300 g. No explanation.
72	Less than 300 g. With explanation. *Examples: Less. Ice is heavier than water.* *Less. There will be water only.*
73	Less than 300 g. No explanation.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret
99	BLANK

Figure 7.4 presents the possible codes for one item in the performance assessment task Solutions, administered to Population 2 students. The TIMSS two-digit coding system is the basis for the coding rubric. Coders were also provided with a list of criteria for a complete response. The solutions task required students to investigate what effect different temperatures have on the speed with which a soluble tablet dissolves. Students were required to develop and record their plan for an experiment to investigate this, carry out their proposed tests on the tablets and record their measurements in a table. They were asked to explain what effect different temperatures have on the speed with which a soluble tablet dissolves, according to their investigation. The following coding guide was used to score students responses to this.

Figure 7.4 Exemplar Coding Guides – Item 3, Solutions Performance Assessment Task

Q3. According to your investigation, what effect do different temperatures have on the speed with which a tablet dissolves?

Criteria for a complete response:

i) Conclusion must be consistent with data table or other presentation of data (graph or text).

ii) Conclusion must describe the relationship presented in the data.

NOTE:

- A wrong direction in the data has been coded for in Q2. However, if an anomaly has occurred and the student recognizes it as such and identifies it as such s/he should receive credit in this question.
- If the student says that temperature has no effect on rate of solution, and if this conclusion is consistent with the student's data, code 29.

Code	Response
Complete Response	
20	Correctly describes trend in the data. *Example: As the temperature increases the tablet dissolve faster.*
21	Describes explicitly only what happens in hot or in cold water but not both. *Examples: In hot water the tablet dissolves faster.* *In cold water the tablet dissolves more slowly.* *In hot water the tablet dissolves twice as fast.*
29	Other complete summaries of data.
Partially Correct Response	
10	Describes trend in the data but the temperatures are not in a reasonable range and student fails to recognize or account for this.
19	Other partially correct.
Incorrect Response	
70	Conclusion not consistent with student's data, and no explanation of the inconsistency offered.
71	Mentions that temperature has an effect but does not describe the effect. *Example: The temperature has a big effect.*
72	Conclusion erroneous: that is, temperature "does not affect rate of solution." *Example: All temperatures have the same effect.*
76	Repeats data but does not draw a conclusion or generalization.
79	Other incorrect.
Nonresponse	
90	Crossed out/erased, illegible, or impossible to interpret.
99	BLANK

In performance assessment almost no item measures a single trait. For example, in order to answer many of the questions, students had to recall and synthesize two or more concepts and use a number of different skills. In addition, items within a task were interdependent both in the sense of being clustered around a common investigatory question (although calling for various skills) and in the sense that some responses depended on data collected and analyzed in previous responses. This richness within a single task or scenario is characteristic of "authentic" problems, and was intentionally structured into the tasks, even though it would render interpretation of results complex and difficult. The revised coding system attempts to reduce these levels of complexity to quantifiable, interpretable data.

In the coding example above, compromises had to be made between expanding the number of codes to capture additional alternative approaches and/or misconceptions, and limiting the coding time per item. Therefore not all possible responses were included in the codes. Decisions about which codes to include were based on empirical data: an alternative approach or an error had to have been made in at least 5% of the field trial responses to be included in the final set of codes.

The similarities in approach and application of the coding systems for performance assessment and free-response items does not imply that the two genres are equal in difficulty for coders. In fact, because of the complexity of measuring several entangled traits simultaneously, it is essential that those who code performance assessment tasks have actually done the tasks themselves or observed students doing these tasks. This is necessary for coders to understand fully what the task is intended to measure, the functioning of equipment, and possible "alternative" perceptions of the tasks by students.

7.6 SUMMARY

In this chapter we have discussed the importance of free-response items and explained how the TIMSS coding guides are used to collect information on how students respond to these items. To illustrate the application of the rubrics to actual TIMSS items, and to demonstrate the potential for analysis, some specific examples were given.

To provide a context for the rubrics, a historical overview of their conception and development was included. Given the interest and expectation from the early stages of the study, it was desired that the information gathered via the free-response items not be limited to correctness only. As a result, coding rubrics were designed to measure three aspects of student response: correctness, method or approach or type of explanation/example given, and misconception or error-type. Through use of a two-digit system it was possible to collect information on all of these aspects.

The analysis of data collected for free-response items will answer several questions of interest, in addition to their contribution to the correctness score. Analyses of students' approaches and conceptions around the world will be of great interest to researchers in mathematics and science education. Furthermore, such data can provide valuable diagnostic information for mathematics and science teachers. We hope that not only the data themselves, but also the methods of analysis that have been briefly described here, will turn out to be useful tools for a better understanding of student thinking in science and mathematics.

REFERENCES

Angell, C. and Kobberstad, T. (1993). *Coding Rubrics for Free-Response Items* (Doc. Ref.: ICC800 / NRC360). Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Angell, C. (1993). *Coding Rubrics for Short-answer and Extended-response Items*. Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Angell, C., Brekke, G., Gjortz, T., Kjaernsli, M., Kobberstad, T., and Lie S. (1994). *Experience with Coding Rubrics for Free-Response Items* (Doc. Ref.: ICC867). Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Brekke, G., Kjaernsli, M., Lie, S., Gisselberg, K., Wester-Wedman, A., Prien, B., and Weng P. (1992). *The TIMSS Pre-Pilot Test: Experience, Critical Comments and Recommendations from Norway, Sweden, and Denmark* (Doc. Ref.: ICC310 / NPC083). Paper prepared for the Third International Mathematics and Science Study (TIMSS).

Kjaernsli, M., Kobberstad, T., and Lie S. (1994). *Draft Free-response Coding Rubrics– Populations 1 and 2* (Doc. Ref.: ICC864). Document prepared for the Third International Mathematics and Science Study (TIMSS).

Kobberstad, T. (1993). Discussion Paper for the Third International Mathematics and Science Study Subject Matter Advisory Committee Meeting, September 1993.

Marshall, M., Koe, C., and Donn, S. (1991). *Pre-Pilot Manual for National Project Coordinators* (Doc. Ref.: ICC158 / NPC032). Prepared for the Third International Mathematics and Science Study (TIMSS).

Taylor, A. (1993a). *Coding Rubrics for Free-Response Items* (Doc. Ref.: ICC648 / NPC249). Discussion Paper for the Third International Mathematics and Science Study National Project Coordinators Meeting, March 1993.

Taylor, A. (1993b). *Manual for Coding Free-Response (Open-Ended) Items for Achievement Review* (Doc. Ref.: ICC662 / NPC260). Document prepared for the Third International Mathematics and Science Study (TIMSS).

Third International Mathematics and Science Study (TIMSS). (1994a). *Coding Guide for Performance Assessment* (Doc. Ref.: ICC885 / NRC422). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1994b). *Coding Rubrics for Free-Response Items.* Paper prepared by the TIMSS Free-Response Item Coding Committee for the National Research Coordinators Meeting, August, 1994.

Third International Mathematics and Science Study (TIMSS). (1994c). *Manual for Coding Free-Response Items–Population 3 Field Trial* (Doc. Ref.: ICC844 / NRC397). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995a). *Coding Guide for Free-Response Items–Populations 1 and 2* (Doc. Ref.: ICC897 / NRC433). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995b). *Coding Guide for Free-Response Items–Population 3* (Doc. Ref.: ICC913 / NRC446). Chestnut Hill, MA: Boston College.

Third International Mathematics and Science Study (TIMSS). (1995c). *Supplement to the Coding Guide for Performance Assessment* (Doc. Ref.: ICC933 / NRC456). Chestnut Hill: Boston College.

Wiley, D. (1992). *Response Typologies From the First TIMSS Achievement Pilot: Implications for Scoring Extended-Response Tasks*. Paper prepared for the Third International Mathematics and Science Study (TIMSS).