2015

ISS

Table of Contents

Summary of TIMSS 2015 Item Writing Process and Guidelines

Typically, participants will work in groups of two or three. Each group will be assigned specific content areas. Participants will be writing items in English and saving them as Microsoft Word files that will be collected at the end of each day.

When writing items, PLEASE:

1. Address the TIMSS 2015 or TIMSS Advanced 2015 Assessment Frameworks. Write questions that match the topics in each content domain, and pay particular attention to writing questions that cover the range of the three cognitive domains. In accordance with the frameworks, write questions that address the applying and reasoning domains, as well as the knowing domain.
2. Consider the best item format for the question. About half of the items you develop should be multiple-choice and the other half should be constructed-response items worth 1 or 2 score points.
3. For each item, consider the timing, grade appropriateness, difficulty level, potential sources of bias (cultural, gender, or geographical), and ease of translation. Make sure that item difficulty is not affected by factors that unnecessarily increase the difficulty of the item, such as unfamiliar or overly difficult vocabulary, grammar, directions, contexts, or stimulus materials.
4. For multiple-choice items, keep the guidelines for writing multiple-choice questions in mind. In particular, ask a direct question, make sure there is one and only one correct answer, and provide plausible distracters.
5. For constructed-response questions, write a full-credit answer to the question in terms of the language, knowledge, and skills that a student in the target grade could be expected to possess. This tests the clarity of the question and also provides guidance about whether to allocate 1 or 2 score points to the item.
6. Develop a specific scoring guide for each constructed-response item.

Introduction

These guidelines are to help ensure that the best possible items are developed for TIMSS 2015, TIMSS Numeracy 2015, and TIMSS Advanced 2015. The TIMSS & PIRLS International Study Center has developed these guidelines for writing and reviewing items and scoring guides to facilitate successful item development. It is important to follow some basic procedures so that the TIMSS assessments are uniform in approach and format. During the item-writing sessions, please ask staff or consult these guidelines if you have any questions.

General Issues in Writing Items for the TIMSS 2015 Assessments

Item writing is a task that requires imagination and creativity, but at the same time demands considerable disciplined thinking within the assessment framework and following the guidelines for item construction provided in this manual. These guidelines pertain to good item and test development practices in general, and have been collected from a number of sources. They are designed to help produce items that measure achievement in mathematics and science fairly and reliably, and that enhance the validity of the TIMSS assessments. All of the following issues must be considered in judging the quality and suitability of an item for inclusion in the TIMSS field tests.

Alignment with the Frameworks

Consistent with the principles of evidence-centered design (e.g., Mislevy, Almond, & Lukas, 2003), the TIMSS 2015 assessment is based on:

- Detailed content and cognitive domain descriptions organized into frameworks for each assessment;
- Items aligned with the content topics and cognitive domains and designed to collect evidence about what students know and are able to do; and

S

domain, but should also produce evidence that a student can employ skills associated with practicing science.

Types of Items

TIMSS includes two types of items—multiple-choice items where the student chooses the correct answer from four options, and constructed-response items that require students to write their answers.

use grade-appropriate vocabulary and terms. The reading level of items should be at an elementary level for the target grade. In general, the amount of reading should be kept to a minimum, given the context of the problem. Write questions in the active voice (i.e., doer of action (subject) before action (verb)) and avoid conditional words, clauses, and tenses (e.g., if, suppose, when).

The context for the item may relate only to the discipline of mathematics or science, or to aspects of those subjects encountered in everyday life. However, if the item involves a "real-world" setting make sure the setting is familiar to students. Avoid using context-specific vocabulary that may not be familiar to all students. Avoid unnecessarily complicated item context or unfamiliar context-specific vocabulary may artificially increase the difficulty of the item and pose a threat to item validity.

Item Difficulty

Information from individual TIMSS 2015 items should provide valuable insight into student learning by providing evidence about what the student knows or is able to do. Additionally, each of the items needs to contribute to the overall mathematics or science assessment. It is desirable that there be some relatively easy items and some challenging items. However, items that almost all students or almost no students are able to answer correctly reduce the effectiveness of the test to discriminate between groups with high achievement and groups with low achievement. Typically, the majority of items used in the final test will be ones that are answered correctly by 30 to 70 percent of the students on average internationally.

Avoiding Bias

When preparing assessment items, be sensitive to the possibility of unintentionally placing particular groups of students at an unfair disadvantage. An international study requires special attention to the diversity of environments, backgrounds, beliefs, and cultures among students in the participating countries.

Considering National Contexts

Be particularly aware of issues related to nationality, culture, ethnicity, and geographic location. Items requiring background knowledge confined to a subset of participating countries are unlikely to be suitable.

Geographic location has an effect on students' learning experiences, as aspects of the local environment have an impact on schooling. Even though television and the Internet can provide students with some knowledge of remote places, firsthand experience of some phenomena enhances understanding and can give some students an advantage over others.

Gender

A gender

a common article in one country may be a fraction of the base unit of currency, while the same article in another country may cost thousands of the base unit. In some countries, the cost of an article may never include a decimal point. If the inclusion of costs is an essential part of a problem, use “Zeds”. This is the TIMSS fictitious unit of currency, which enables each country to work with the same numbers.

Graphics

Take special care to ensure that diagrams and graphs are drawn accurately (to scale unless otherwise noted), and are correctly and fully labeled. Any graphics included in an item should be necessary, i.e., in order to solve the problem or to answer the question, and should be adequately explained and referred to directly within the item text as indicated by the principles of universal design for assessment items. In particular, visual elements should

TIMSS2015any copyrighted stimulus material must be acknowledged appropriately. For example, statistical graphs from publications or extracts from articles in publications that are used in an item must be identified appropriately, and full details about the sources must be submitted with the item.

Pattern Items —TIMSS 2015 Mathematics ONLY

mathematics and physics notation, selected formulas from advanced mathematics and physics, and selected physics constants. When writing items for which a formula is necessary to solve the problem or answer the question, please also include the formula(s) with the distracteranalysis (multiple-choice items) or the scoring guide (constructed-response items).

Writing Multiple-choice Items

A multiple-choice item asks a question or establishes the situation for a response. For the TIMSS 2015 assessment, this type of item includes four response choices, or options, from which the correct answer is

The Stem

For the TIMSS 2015 assessment, since clarity is of vital importance, please phrase all stems as a direct question. The following is an example of a stem formulated as a question:

1. Provide sufficient information in the stem to make the question clear and unambiguous to students. In nearly all cases, the question must be able to stand alone, and be answerable without the response options. An exception would be items asking students to choose the best estimate of a quantity.
2. The stem should not include extraneous information. Extraneous information is liable to confuse students who might otherwise have determined the correct answer.
3. Avoid using negative stems—those containing words such as NOT, LEAST, WORST, EXCEPT, etc. If it is absolutely necessary to use a negative stem, highlight the negative word, (e.g., capitalize, underline, or put in bold type so that it stands out for the student). If the stem is negative, use only positive response options—do not use double negatives.
4. If there is not one universally agreed upon answer to the question, it is best to include “of the following” or some similar qualifying phrase in the stem.

5. Avoid questions for which a wrong method yields the correct answer (e.g., a question about a circle with a radius of 2, because $2r = r^2$, students computing either the area or the circumference get 4).

Structure of the Response Options (or Alternatives)

1. Write multiple-choice items with four response options, labeled A–D (as shown in the example item about distance traveled, above).
2. Make sure that one of the four response options or alternatives is the key or correct answer. Make sure there is only one correct or best answer.
3. Make sure that the four response options are independent. For example, response options should not represent subsets of other options. Also, do not include pairs of response options that constitute an inclusive set of circumstances (e.g., day or night, does or does not).
4. Make sure that the grammatical structure of all response options "fit" the stem. Inconsistent grammar can provide clues to the key or eliminate incorrect response options. Avoid writing items where the options complete a sentence begun in the stem, because these can separate problems with translation.
5. Make sure all (or sets) of the response options are parallel in length, level of complexity, and grammatical structure. Avoid the tendency to include more details or qualifications in the correct response, thus making it stand out. If the options are not parallel in length, please order the options short to long if at all possible.
6. Do not use words or phrases in the stem that are repeated in one of the response options and, therefore, act as a clue to the correct response.
7. Do NOT use "none of these" and "all of these" as response options.
8. Arrange the response options in a logical order if this makes sense and saves the student time in reading the options (e.g., years in chronological order, numbers from least to greatest).

9. Avoid writing items where students can work backwards from the response options to find the correct answer (e.g., solving for x in an equation). Sometimes described as “plug and chug” items, such questions or problems will not be included in any of the TIMSS 2015 assessments. In such cases, a constructed-response item may be more appropriate than a multiple-choice item.

Plausibility of Distracters

Use plausible distracters (incorrect response options) that are based on likely student errors or misconceptions. This reduces the likelihood of students arriving at the correct response by eliminating other choices and, equally important, may allow identification of widespread student misunderstandings or tendencies that could lead to curricular or instructional improvements. If there are no plausible errors or misconceptions, still make the options “reasonable.” For example, they should be from the same area of content. However, avoid the use of “trick” distracters.

Distracter Analysis

Please include a brief analysis of each response option or rationale for inclusion of specific response options with your item (one sentence at the most for each response option). For example:

Distracterrationale:

A. [Key]

B. Assumes that boiling water heats up the rock and results in its separation into two pieces

C. Associates the density difference between water and rock with water acting to split the rock.

D. Assumes that water dissolves the rock in such a way that the two pieces result.

Writing Constructed-response Items and Scoring Guides

For some desired outcomes of mathematics and science education, constructed-response items provide more valid measures of achievement than do multiple-choice items. The quality of constructed-response items depends largely on the ability of scorers to assign scores consistently and reliably within and across countries. Thus, it is essential that each constructed-response item and its scoring guide be developed together.

PLEASE keep the guidelines for writing constructed-response questions in mind. In particular, ask a clear question, and develop a scoring guide for the question at the same time as the question is developed.

Constructed-response items usually require students to give a numerical result, provide a short explanation or description given in one or two phrases or sentences, complete a table, or provide a sketch. They are scored as either 1 or 2 points for fully

6. Produce a scoring guide (see below). This action usually results in amendments to the item to clarify its purpose and improve the quality of information that can be obtained from student responses.

Writing Scoring Guides

To ensure reliability, constructed-response items need scoring guides with well-defined categories for allocating score points. It also is important to collect information of value for educational improvement. Students' answers can provide insights into what they know and are able to do, including common misconceptions.

The TIMSS Generalized Scoring Guidelines

The generalized scoring guidelines used for 1- and 2-point constructed-response items are described in Table 1.

Table 1: TIMSS Generalized Scoring Guidelines for Constructed-response Items

Score Points for 1-point system
1 Point (Full credit) A one-point response is correct. The response indicates that the student has completed the task correctly.
0 Points (No credit) A zero-point response is incorrect, irrelevant, or incoherent.
Score Points for 2-point system
2 Points (Full credit) A two-point response is complete and correct. The response demonstrates a thorough understanding of the concepts and/or procedures embodied in the task. • Indicates that the student has completed all aspects of the task, showing correct application of concepts and/or procedures • Contains clear, complete explanations, supporting work, or evidence when required
1 Point (Partial credit) A one-point response is only partially correct. The response demonstrates only a partial understanding of the concepts and/or procedures embodied in the task. • Addresses some elements of the task correctly but may be incomplete • May contain a correct answer but an incomplete explanation when required • May contain an incorrect answer with an explanation or supporting work indicating a correct understanding of the concepts
0 Points (No credit) A zero-point response is inaccurate or inadequate, irrelevant, or incoherent.

The TIMSS Two-digit Diagnostic Scoring System

The TIMSS two-digit diagnostic scoring system uses two digits. For example, 10, 11, or 20.

The first digit is the score indicating the degree of correctness of the response as described in the generalized scoring guidelines.

The second digit is used to classify the method used in solving a problem, or perhaps to track common errors or misconceptions. The information from the second digit addresses questions such as: Do approaches that lead to correct responses to the item vary across countries? Is there one approach that students have more success with than others? What are the common misconceptions that students have about the matter being tested? What common errors are made?

The First Digit

The first digit for correct or partially correct responses signifies the number of score points given to the response. Thus:

The first digit for correct responses is 1 for one point or 2 for two points. When TIMSS started in the early 1990s, it was decided not to use 0 for the first digit. Thus:

The first digit for incorrect responses is 7.

The first digit for a blank response is 9.

The Second Digit

The second digit for correct or incorrect responses provides diagnostic information. Thus:

The second digits used for diagnostic purposes where correct or incorrect responses can be 0 through 2 (codes 20–22, 10–12, and 70–72).

However, it is unusual for an item to give rise to more than two commonly used correct methods, or more than one common error or misconception. Frequently no more than one or two categories are required. In other words, the specific diagnostic codes should capture only the predominant correct and incorrect approaches/strategies used by students. Scoring of constructed-response items is a significant cost factor for national centers, so care should be taken not to provide codes for response types that do not have apparent value for educational improvement.

Since not all incorrect student responses should be categorized into pre-defined categories, for codes with a first digit of 7, the second digit of 9 is used to designate a response that is "other" than any specific diagnostic codes included in the guide. Thus, an incorrect response not fitting a pre-defined incorrect code is given a 79 for "other incorrect." If no diagnostic categories are defined, all incorrect responses are coded 79.

Code 99 means a completely BLANK response.

Examples of Scoring Guides

The following examples are given to illustrate the diagnostic scoring guides used in TIMSS (Grade 4 and Grade 8) and TIMSS Advanced (advanced mathematics and physics).

Grade 4 Mathematics Item (1 point):

Code	Response	Item: M051601
	Correct Response	
10	13	
	Incorrect Response	
70		
79	Incorrect (including crossed out, erased, stray marks, illegible, and off task)	
	Non response	
99	Blank	

Grade 8 Science I tem (2 points):

	" "	# * **1 3** / 3
	! ! # " "	
1 /	" ! " # ! " " " # (! ! # & # ! % ! - # " # $ # " " $! # # ! + ' " * & # ! ! # " ! # & # ! % ! # ! & " $ & # ! & # # $ " " $! + " $! # " " $ " # " " # # # & # ! + # ! # & # ! % ! # " " $! " " # ! +	
	! # (! ! # " "	
0 /	" ! " # ! " " " # (! ! # & # ! % ! - # " & # $ # # # " " # # ! + ' " * $! & # ! % ! " +	
0 0	# # " " # & # $ # ! ! ! # & # ! % ! + ' " * " # + # " ! # ! +	
	! ! # " "	
6 8	! ! # - $! " " $ #) ! ! " ")) " # ! ( ) ! # " . ' " * $! # " (+ # ! # $ " +	
	! " "	
8 8		

Advanced Mathematics I tem (1 point):

Code	Response	Item: MA13027
	Correct Response	
20	Any of 2 pi, 2π , 6.28, 6.3, or $2\pi = 6.28$	
	Partially Correct Response	
10	$\lim_{n\to\infty}$ —	

Physics Item (2 point s):

	" "	# * 1 2 / 1 1
	! ! # " "	
1 /	! " " # # $ " # & "# " 0 + # # " # # & & " # # & # , " & * # & $F = \frac{GMm}{r^2}$ # (* 1 + " # ! $! #! # $a = \frac{v^2}{r}$)! # * 2 + ! % " # ! $! $v = \sqrt{\frac{GM}{r}}$ # (!) $ % # . $" " # " # " & # #$". " ! # ! # - !# . +	
	!# (! ! # " "	
0 /	# 0 1 # $# # # 2	
	! ! # " "	
6 /	# 0 (# +	
6 8	# ! ! ! # - $! "" $#) ! ") "#! (! ") # " .	
	! " "	
8 8		

If it happens that items are written after the NRC meeting, the item

Reviewing Constructed

Appendix A: Multiple-choice Item Review Checklist

Item Characteristic	Yes	No
Is the mathematics/science correct?		
Task clear to students?		

Appendix B: Constructed-response Item and Scoring Guide Review Checklist

Item Characteristic	Yes	No
Is the mathematics/science correct?		
Task clear to students?		
Free of cultural, gender, or geographical bias?		
Seems to be OK for translation ?		
No unfamiliar factors contributing to difficulty?		
Clear expectations for full-credit response?		
Task can be completed in a reasonable time?		
Scoring guide has appropriate correct and incorrect categories?		
Scoring guide has appropriate number of score points?		
Scoring guide descriptors clear?		
Content classification correct?		
Cognitive classification correct?		

References

Dolan, R.P., & Hall, T.E. (2007). Developing accessible tests with universal design and digital technologies: Ensuring we standardize the right things. In L. L. Cook & C. J. Cahalan (Eds.), *Large-scale assessment and accommodations: What works* (pp. 95-111). Arlington, VA: Council for Exception Children.

Mislevy, R.J., Almond, R.G., & Lukas, J.F. (2003). *A brief introduction to evidence-centered design* (ETS Research Report RR-03-16). Princeton, NJ: Educational Testing Service.