Appendix C

Learning Mathematics for Teaching/Study of Instructional Improvement
Item Development
ITEM WRITING RECOMMENDATIONS
Revised 4.28.04

Item writing efforts have been under way at LMT/SII for four years. In that time, we have learned much about writing items which can be later useful in research and evaluation. In this document, we describe some of what we have learned.

Why we write. We create items to form scales which are then used for various research purposes. Currently, LMT uses these scales to assess changes in teachers' knowledge as a result of participation in mathematics-focused professional development. SII uses these scales to predict growth in student achievement from teachers' content knowledge for teaching. And other projects across the country – Math-Science Partnerships, in particular – are using these items to assess changes in teachers' knowledge as a result of programs or policies.

What we write. First priority for piloting this year are middle grades patterns, functions, and algebra and number and operations. We include proportional reasoning in the latter category.

How to write:

    Write content knowledge items which capture mathematics knowledge use <u>in teaching</u>. By this, we mean to differentiate items which capture "pure" or "adult" mathematical abilities (e.g., 8th grade NAEP items) from items which capture the additional ways teachers think about mathematics in the course of their work. Analyses of existing items suggest at least three categories within knowledge-for-teaching: making and using mathematical representations; providing explanations for common mathematical "rules" or statements; and evaluating student solutions or methods. There may be others, such as using mathematical definitions in the context of teaching, identifying adequate/inadequate proof of ideas or methods, choosing numeric examples to underscore or assess a point (e.g., ordering decimals item), etc.

Based on past experience, we find that when we aim to write only knowledge-in-teaching, we end up with about half knowledge-in-teaching items, half pure/adult items.

Write items to which teachers' responses will vary. We learn nothing from items that all respondents answer in the same way. This includes, for example, items which are too "difficult" for the average teacher, and those which are too "easy" for the average teacher. Even though these may be good items for other reasons, they yield minimal information regarding the capability of the majority of teachers in our sample. As a result, target items toward the 25th, 50th, and 75th percentile teachers.

Also, when developing items, pose questions that will yield a variety of responses from teachers. As you write, anticipate those responses, constructing choices that reflect what teachers might say in an interview. For instance, you might anticipate typical errors teachers might make mathematically, or typical misunderstandings they might have of student thinking.

Make distracters sound as reasonable as possible. One aid in writing medium to difficult items is to make the "distracters" – or wrong answers – as plausible as possible. For example, you might construct one of the following: answers that would make sense mathematically, but would not make sense in terms of students' learning; answers that correspond to popular but inaccurate theories about how students learn, or theories that would not help in this particular situation; answers that go with the reform "formats" – but don't actually have any mathematics inside (e.g., using manipulatives when there's really no mathematical point). To identify potential difficulty and distracters, it might help to pilot items with and without responses on your local teacher education students or with teachers with whom you work. Piloting items without responses can generate potential responses for multiple-choice items.

Do not cross constructs. Please do not write items that require teachers to draw on more than one domain or focal topic in formulating an answer. The example in Table 1 can help show why. In this question, the teacher must first know the mathematics, then figure out what these students might be thinking, and where that thinking went wrong. If a teacher answers this item incorrectly, we cannot know whether she does not know the students' mistake, or whether she does not know the mathematics. Because we cannot differentiate these two constructs, this item will not give us good information about either.

Table 1. Item requiring teachers to work in more than one domain.

Imagine that you decide to use base ten blocks to help your students understand decimals. When you ask them to compare .35 with .4, most students line up 35 little cubes next to 4 little cubes. When you ask them which is more, they reply that .35 is more than .4.

27. What is your interpretation of this?

a. The students think 35 is greater than 4. Because of this, they assume that .35 must also be more than .4
b. The students are paying attention to the blocks, which are misleading them, since these are not well-suited for use with decimals.
c. The students are right, even though it appears at first that they have it wrong. .35 is more than .4
d. The students are using the blocks correctly, but they are interpreting what they see incorrectly.

This item can be fixed by removing choice "c," indicating that .35 is LESS than .4 in the stem, and then configuring all response choices to be about possible student errors. To assess whether teachers do in fact understand whether .35 is less than .4, we would (and did) write a separate item, asking teachers to order decimals.

Cross (or ignore) ideological lines. The best items work consistently across multiple views of teaching and learning. Even the sharpest of partisans regarding subject matter should be able to agree that the items we use measure skills and knowledge important to teaching. Writing valid items across ideological lines is important, not only because LMT/SII's research program does not adhere to partisan thinking about teaching and learning, but also because these items ultimately should be able to garner general acceptance as measures of teacher knowledge.

Avoid ambiguity. Ambiguity arises when insufficient information is provided in the item or stem, when a choice is not unambiguously wrong, and when it is not clear what the teacher is being asked to do. We should strive to eliminate ambiguity in items and stems, for it cannot be undone analytically; once we determine an item is ambiguous, we can no longer use it.

Stem 22 item d suffers from ambiguity (see Table 2). Item d is ambiguous because zero may or may not be the smallest number depending on the number system being referenced. Neither the item nor the stem referenced a number system, and, therefore, the item has no clear

answer. For example, if a teacher marked "No," her answer was correct if she was thinking of whole numbers and wrong if she was thinking of integers. Therefore, items need to reference a context that is absolutely clear to teachers or if not, clarified in the item or stem, and the content being measured by the item must be true regardless of the context unless the context is stated.

Table 2. Ambiguous Item: Stem 22 *item d*.

Ms. Dominguez was working with a new textbook and she noticed that it gave more attention to the number 0 than her old book. She came across a page that asked students to determine if a few statements about 0 were true or false. Intrigued, she showed them to her sister who is also a teacher, and asked her what she thought.

Which statements should the sisters select as being true? (Mark an "X" in the YES, NO, or I'm NOT SURE box for each item below.)

	Yes	No	I'm not sure
a) 0 is an even number.	1	2	3
b) 0 is not really a number. It is a placeholder in writing big numbers.	1	2	3
c) The number 8 can be written as 008.	1	2	3
d) 0 is the smallest number.	1	2	3

Avoid partial credit by making answers clearly right or wrong (not better or worse). Partial credit arises when items are ordered best-to-worst, best-not great-bad, best-possible-wrong, and so forth. Don't write these, as they introduce new analytic wrinkles we'd prefer not to deal with.

Keep items independent (local independence). Independence of items occurs when knowing an answer to one item does not affect the probability of correctly answering another item, controlling for underlying ability. Unfortunately, some stems with multiple items have dependence, since items are linked by the stem. A teacher answering Stem 26 (see Table 3), need only determine what she thinks is wrong with Jill's response—then can mark the other two responses as incorrect interpretations. The probability of answering a and c correctly depend

upon the probability the teacher identifies b correctly, as well. These three items, then, are not independent.

To avoid this problem, try as much as possible not to link multiple items under one stem. Stem 26, for instance, can be converted to a multiple-choice item. Likewise, we should avoid partial credit stems (see below) as this creates dependence.

Table 3. Example item showing dependence and non-multiple choice response –
Stem 26

<u>Local Dependence – more than one answer, three stems that can not stand alone</u>

Consider Jill's response to a subtraction problem. How might she have gotten an answer like this?

```
      51
    - 18
      47
```

(Mark an "X" in the YES, NO, or I'M NOT SURE box for each item below.)

	Yes	No	I'm Not Sure
	1		5
		8	
a. She does not know basic subtraction facts.	☐	🟨	☐
b. Instead of regrouping, she subtracted 1 from 8.	🟨	☐	☐
c. She does not understand place value.	☐	🟨	☐

<u>Local Independence – only one answer, only one item</u>

Consider Jill's response to a subtraction problem. How might she have gotten an answer like this?

```
      51
    - 18
      47
```

Circle the answer that indicates why Jill arrived at an answer of 47 instead of 33.

a. She does not know basic subtraction facts.

b.   Instead of regrouping, she subtracted 1 from 8.

c. She does not understand place value.

Another form of dependence occurs when all correct answers to multiple items under one stem are the same. That is, all items within a stem are all 'true', 'yes', or 'b', for example. Stem 28 (see Table 4) was written with the correct answer to all items as 'yes.' This creates a form of "psychological dependence" because the respondent may force an answer to be different, not because she thinks the answer to an item is 'no' in this case but because she doubts that all the correct answers for a stem will be 'yes.' Psychological dependence reduces our reliability. Correct answers

should vary across yes, no, true, false, a, b, c, etc when multiple items are linked to a stem.

Table 4. Psychological Dependence –Item 28.

28.   Which of the following is a common problem among first graders?
(Mark an "X" in the YES, NO, or I'M NOT SURE box for each item below.)

	Yes	No	I'm Not Sure
	1	**5**	**8**
a. Writing 23 as 203.....................................	☐	☐	☐
b. Mixing up 23 with 32.	☐	☐	☐
c. Having trouble reading two-digit numbers.........	☐	☐	☐
d. Counting objects accurately up to 50.	☐	☐	☐
e. Writing numerals correctly....................................	☐	☐	☐

Ensure each item can be taken by teachers in all target grades. We cannot predict how these items will be used; it may be that second-grade teachers take our third-to-eighth grade geometry items, or third grade teachers take middle-grade patterns, functions, and algebra items. Writing across grades implies not identifying grade level in the stem or the item itself (i.e., not writing "Two third-grade teachers were talking….") so that some teachers do not decide that the item does not pertain to them.

Write items that look as little as possible like test items. Instead, please embed each item you write in a realistic problem of teaching, or what we call a "scenario." We use scenarios for two reasons. The first reason is based in how teachers perceive the items; in past research, teachers have objected to efforts to test their content knowledge. They also argued that most teacher-test items have little to do with the way in which they use knowledge in teaching. When respondents do not think items are fair or valid, they may skip them, creating serious missing data problems for the research. Items using well-constructed scenarios address this concern. By tapping into legitimate teaching practices, scenarios provide a more credible approach to assessment, focusing on content knowledge as it is situated in the work of teaching. Scenarios also serve to disguise the test-like quality of the items. A second reason for using scenarios relates to our interest in a particular kind of teacher content knowledge. We want to measure knowledge that teachers use in their

practice. By placing knowledge items in context, we are working at the intersection of teacher knowledge and teaching practice.

Possible scenarios include:

- A teacher designing or creating curriculum materials or assessments for her own classroom or for colleagues.
- A teacher reflecting on her own classroom instruction.
- A teacher reviewing, grading, reflecting on the work of her own students.
- A teacher holding a discussion with colleagues, principals, parents, or students.
- A teacher thinking about a presentation or discussion in a professional development setting.
- A teacher appraising a specific task in her textbook.
- A teacher examining an item on an assessment.
- A teacher evaluating the validity of a representation.
- A teacher considering what a reader might have to know or do to interpret a particular text.
- A teacher evaluating a mathematical claim, or considering what it would take to evaluate the claim.

Write as briefly as possible. These items can take teachers a significant amount of time to read, understand, and answer. Whenever you can, however, please try to keep item length to a minimum.

Advice on how to formulate choices:

Write items with multiple-choice responses. When compared to true/false and yes/no response choices, items containing multiple-choice responses are more reliable. Multiple-choice items are preferred because they have more distracters (false propositions) and distracters reduce the chance that respondents will guess the answer correctly. Multiple-choice items have four or five response choices with only one correct response --the remaining responses are distracters. A binary response (examples: true/false, yes/no) only has one distracter, and a respondent has 50% chance of guessing the correct response. Therefore, questions with only two response choices (i.e. true/false, yes/no) leads to the increased likelihood of guessing an answer correctly.

Use "don't know" responses carefully. "Don't know" responses too often provide little or no information about what teachers know or how their understanding relates to "yes" or "no." On the other

hand, we have found that providing the response choice of "don't know" reduces the likelihood that teachers will skip the item, particularly for very specialized knowledge domains. For example, in a test designed to measure reading knowledge for teachers, phoneme items were piloted without the option to mark "don't know" and many teachers skipped the item. These items were then used on the SII Teacher Questionnaire with an option to mark "don't know" and the response rate improved. As a result, we suggest writing multiple-choice items when possible. When writing difficult items, do provide a "don't know" response choice when the possibility of teachers skipping the items seems high.

Avoid rank-ordering. Typical rank-order response type items ask the respondent to put something in order, in order of importance or the order in which something happened. Rank-order items require additional cognitive demand than required by a multiple-choice item. For the respondent to answer a rank-order item, she must first do the work to find the answer and order the answers. Incorrect ordering can be due to either not knowing the correct order or mis-ordering. Additionally, analysis of this type item requires a different statistical model than do binary or multiple-choice response items.

The "who" in what we write:

Get clearance to use copyrighted text. If you use passages from mathematics texts, basal readers, or other printed and copyrighted materials, please *notify Jenny Lewis as soon as possible*. We can attempt to get copyright permissions, but it takes time (and some expense). Permission for copyright varies by publisher and author. An alternative strategy (and one we strongly prefer) is to write your own problems, texts, tests, and so forth.

Provide each item with an intellectual history. As you near completion on each item, please take time to stop and reflect on its development, then record those thoughts. Our experience tells us that having some sense of each item's history can be helpful during analysis, or revisions should they be necessary. For instance, we might want to know, for each item: what that item is intended to measure; what research or experience led you to write this item; what teachers' responses might mean (e.g., if a teacher chooses option A, we might conjecture she thinks in this particular way); what the "right" answers are. We will provide a form to help organize this information.