Selective Testy Comments About Testing, Tests and Test Items

Werner Liedtke

Tests and testing are an integral part of teaching and learning. The results from tests are used for more than making instructional decisions and adjustments. Data from tests are used to rank schools and compare achievements in school districts, as well as provinces/territories and countries. Responses on tests are used to generate and publish statements about the development of student numeracy. These are important decisions, and testing and tests are therefore also important—the testing settings, the tests and the items on the tests yield the information that is both necessary and appropriate for these types of tasks.

We often assume that the tests we use provide us with the information we think they do, and that the items on these tests assess what we think they assess. Several examples will be presented in this article to illustrate that such an assumption may not always be true, and that greater care needs to be given to selecting test settings and designing test items. Without this care, the conclusions that are based on the results of tests may not be as meaningful and may come into question.

Test Settings

A valid comparison of test results is based on the premise that the tests were administered under similar circumstances for all or most students. However, this may not always be the case. Published rankings of schools by newspapers, and parents in British Columbia being allowed to send their children to the school of their choice puts a lot of pressure on having students do well on tests. The following examples indicate this.

A report from the United States included in a recent issue of *Maclean's* magazine mentioned that young students in a school were asked to change their answers on a test they just had written.

During the second part of a final practicum, one of my student teachers came to me to say that she was unable to teach a mathematics unit she had planned for her Grade 4 class. The school had decided to set aside three weeks to prepare these students for the topics that were part of the upcoming provincial test. Another student teacher told me that the students in her school were receiving specific instructions about how to write the provincial examination. I am sad to say that I was told of a school where someone looked at the actual test before it was administered and shared some information with students. This was done because of the low ranking from the previous year.

Imagine if a teacher takes the questions from the previous year's examination and displays them on the walls of the classroom at the beginning of the school year. These questions will become not only the focus of instruction but the year's curriculum. The students will score well on the tests, but will they be "Wise/Numerate and/or Test-wise and/or Otherwise" (Liedtke 2003)?

Some of these scenarios remind me of my Grade 12 experience in Alberta. The final mark was entirely based on the departmental examinations score. After a mad rush through the content, the last three months were spent writing old examinations.

Differences in settings can bring conclusions and comparisons into question. Settings that focus on test writing do not have any instructional value and do not contribute to fostering the development of number sense and numeracy.

Tests and Test Items

Tests Prepared by Teachers

No doubt many of us have been surprised, amused or perhaps a little disturbed at a student's response to an item on one of our tests. How could this happen after the great care we had taken to construct each item? No matter how much time and care are taken or how many times an item is revised, unpredictable responses will surface from time to time.

To collect assessment information about the strategies, ideas and procedures that students have learned and the conceptual understanding they have, appropriate test items need to be prepared or selected. This

task requires time, care and reflection. Items should be free of mathematical terminology, and instructions should be clearly identified and succinctly stated. Meeting these conditions can be quite a challenge. Without appropriate test items, student responses—correct, partially correct and incorrect—will not yield any meaningful information about student knowledge.

Some of the student teachers I work with are surprised when I express concern about some of the assessment items they have included in their first unit plan. They had used student textbooks as references, and many of the examples were selected from chapter and unit tests in these books. The fact is, I think that many items in these references are not appropriate and do not yield valuable assessment information for appropriate conclusions about students.

The following are examples of assessment items that were part of unit tests on measurement prepared by student teachers. Key questions that were posed and some of the points that were raised during the discussions about these items are identified.

The following examples illustrate the challenges of preparing and selecting test items, and raise awareness of inappropriate test items that sometimes appear in published references.

For figures showing six segments and unions of segments, the directions were: "Measure to the nearest centimetre. Record the results." Why were there so many items? What new information could become available about a student from the repetition of these tasks? How many items are needed to find out whether students know how to measure to the nearest unit? How can three examples tell us everything we need to know about a student's ability to round up and round down? Are complex figures that show the union of segments required? Why or why not?

The instructions for several segments were: "Estimate the length of each path. Record your estimate." What marking scheme is used for the response? What, if anything, does a student's response tell us about her ability to estimate or about the estimation strategies she has at her disposal? During discussions about estimation, the word *reasonable* is almost sure to surface. How is *reasonable* defined? Different levels of number and measurement sense exist, and if students estimated, should not all of their responses be considered reasonable?

The instructions for several segments were: "First record an estimate. Measure and record the results." How many students will complete the task in this order? If students do follow the instructions, how can we know if they changed their estimates after carrying out the measurements? What, if anything, can we learn about students from items of this type?

For five irregular figures of different shapes and with straight sides, enough information was provided to determine the dimensions of each of the sides. For one of the figures, two different units are used for the dimensions. The instructions were: "Find and record the perimeter for each." What is assessed? Why are so many items used? What do we find out about a student's knowledge of perimeter if they answered incorrectly? Why are two different units used for one of the figures? Would anyone ever use different units for sketches or diagrams? If students answer incorrectly, what can we possibly say about their knowledge of perimeter? What types of questions would give us insight into students' conceptual understanding of perimeter?

The same or similar questions can be posed for assessment items for other areas of measurement and other topics as well. These examples and questions illustrate that appropriate assessment items are required to make meaningful evaluative statements about students. Without appropriate items, it may not be possible to plan for effective instruction or intervention (individualized education plans [IEPs]).

As one might expect, one of the major goals of the Diagnosis and Intervention course (available from the division of continuing studies at the University of Victoria) is teaching the creation and selection of effective assessment. Effective intervention is not possible if we do not know what students know and do not know, or how they think. A major part of the course consists of designing questions and assessing questions designed by others. In one assignment, teachers enrolled in the course take seven sample items from one of the readings for the course (Charles and Lobato 1998) and collect reactions from their students. I was amazed, to say the least, to see the results. Three are described below.

The first item:

> Circle the number that is closest to ¼:
> a. 0.4 b. 1.4
> c. 0.14 d. 0.25
> e. 0.5

Two things became obvious. Many students did not know how to deal with 0.25. Should it be one of the choices? If so, why? Some students chose 1.4 because it was "closest" to ¼. How can one argue with that response? How many markers would think of that response?

The importance of using correct language is illustrated in the following example:

> Lakiesha says that you can add zero to a decimal number (like 0.2) without it changing the value of the number. Is she correct? Explain why or why not.

Were the authors thinking about adding a zero (0.2 + 0)? I have my doubts, but that is what the item calls for. Many of the students understood the item this way.

One item was about two boys fishing. A couple of girls asked why boys were fishing and not girls.

If these sorts of things can appear in a monograph published by the National Council of Supervisors of Mathematics, then my concerns for greater care in designing test items are warranted and reinforced.

Published Tests

For the sake of discussion, let's imagine that a test to establish provincial comparisons, establish national or provincial norms, rank schools or make international comparisons is about to be administered to students who have been exposed to a sequential, research-based curriculum. If topics or items are included that are not part of this curriculum, such items should be blacked out. No matter how well-known or important a test, its items should not be a basis for curriculum revision. It may seem a little far-fetched for this to happen, but I have heard the suggestion many times for introducing topics simply because they are part of a well-known or important test. Tests should not determine curriculum; it should be the other way around.

The many tests I have examined over the years have led me to conclude that people in measurement and testing who design and construct these instruments need to take much greater care. Not only are editors required, referees are needed as well. Without appropriate test items, interpretations of the results come into question.

The mathematical language used in tests has to be correct. *Number* is too often confused with *numeral*, and *more* is used when *greater than* is correct. The language is often unclear, uncommon or even incorrect.

The following examples illustrate some of the reasons for my concerns. (More examples could have been included. At least half of the test items could be edited, especially for inappropriate spelling or representation.)

Trends in International Mathematics and Science Study (TIMSS) sample elementary school mathematics test (Grades 3 and 4):

Question 6) 25 × 18 is more than 24 × 18. How much more?
A) 1 B) 18 C) 24 D) 25

Not only is the language incorrect, but no one actually talks like that, especially not Grade 3 or 4 students. Several items on this test have sentences that are too long and complex. However, this item requires a few more appropriate words. As the item stands, there are at least two possible interpretations. The answers could be 1 (1 more group) or 18 (18 more items). The instructions need to tell students to consider the answer or product, and the comparison should be discussed in terms of *greater*.

Question 7) The given numerals in column A are 10, 15, 25 and 50. Corresponding numerals in column B are 2, 3, 5 and 10, respectively. The question posed is: "What do you have to do to each number in column A to get the number next to it in Column B?"

A) Add 8 to the number in column A.
B) Add 8 to the number in Column A.
C) Subtract 8 from the number in Column A.
D) Multiply the number in Column A by 5.
E) Divide the number in Column A by 5.

The first two choices are the same—well, almost. In the first choice, the *c* in *column* is lower case. Also, each choice uses the singular form of *number*. This could make choice C correct because 10 – 8 = 2. Even if this problem can easily be solved, the test in the present form should never be displayed for anyone or, even worse, used by teachers.

Question 15) Four children measured the width of a room by counting how many paces it took them to cross it. The chart shows their measurements.

Name	Number of paces
Stephen	10
Erlane	8
Ana	9
Carlos	7

Who has the longest pace?

A) Stephen
B) Erlane
C) Ana
D) Carlos

Important information is missing, and the language could be much simpler. Should students be told that an attempt was made to take steps of the same size? What exactly is meant by "their measurements?" Is that a common expression?

Programme for International Student Assessment (PISA)—National Centre for Education Statistics (NCES) (Grade 4):

Question 3) Which of these is largest?
1 kilogram 1 centigram 1 milligram 1 gram

My curiosity is killing me. What is the person who designed this item trying to assess? What would anyone possibly learn about a student's knowledge if they made a correct or incorrect choice? What about the choice of centigram? Who uses this unit?

One major reason we teach about measurement is that our eyes may deceive us. For example, is it possible for one gram of cotton candy to be larger than one kilogram of lead? Items like these are of as little value as examples that ask students to record an estimate, or even worse, to first record an estimate and then to perform a calculation or measurement.

Question 6) Here is a number sentence $4 \times \_ < 17$
Which number could go in the _ to make the sentence true?
• 4 • 5 • 12 • 13

This is the exact language and punctuation used on the test. Not only is the presentation awkward and somewhat unusual, but the instructions are difficult to follow as well.

Question 14) What is 3 times 23?
• 323 • 233 • 69 26

The last choice lacks an indicator and the question is curt and unusual. Students who have a conceptual understanding of multiplication could supply all kinds of responses to the request. The question needs to be clearer by asking students to think about the answer.

Question 19) In which pair of numbers is the second number 100 more than the first number?
• 199 and 209 • 4236 and 4246
• 9635 and 9735 • 51,863 and 52, 863

Some people will object to the use of numerals or number names in items like these. However, when numerals are compared, the terms *greater* and *less* should be used. The use of commas in the last item looks odd and could confuse. For some students in Canada and in other countries, commas are decimal points. Spaces and half-spaces should be used.

Question 20) Figure 1 - A 1by3 rectangle showing 3 squares
Figure 2 - A 2by3 rectangle showing 6 squares
Figure 3 - A 3by3 rectangle showing 9 squares
Here is the beginning of a pattern of tiles. If the pattern continues, how many tiles will be in Figure 6?
• 12 • 15 • 18 • 21

Test items about patterns require a lot of essential information. Any repeating pattern can be changed into a growing pattern and vice versa. Because growing patterns can be extended in many different ways, students need to be told if the pattern is repeating or growing, and if it is the latter, if it is growing in the same way as indicated by the examples. It is possible, for example, that the given pattern could continue and that 21 could be a logical choice for Figure 6. Would that be marked as incorrect? It likely would, even though the answer can be justified.

Mathematics Assessment (Grade 6)

While my daughter was teaching in northern Manitoba, she shared with me an examination that the Grade 6 students in the district had to write. Quite a few items caught my eye, and I will comment on a couple of examples.

Two items dealt with rolling a die. In both instances, the rather elaborate illustrations show two dice. In one instance, the action diagram indicates that they are being rolled. This can create some interesting scenarios, especially because many people have difficulty distinguishing between *die* and *dice*.

A question labelled "Number Sense" asks students to "Write equations that equal 180." After students are asked to use each of the four operations, they are requested to "use 3 or more numbers in each equation," "use 2 or more operations in each equation" and "use decimals." However, what if students used only zeroes and/or ones for the equations? How might their answers be assessed? How might students define *number sense* and why might they think it was used as the title to this question?

One of the items dealt with patterns. Squares with sides 1, 2 and 3 are shown, and the respective perimeters 4, 8 and 12 are provided. The first request is: "Draw the next 2 models in this pattern." The word *draw* could result in some students attempting to produce some time-consuming constructions. How might responses of drawing the squares with sides 1, 2 and 3, or with sides 3, 2 and 1 be assessed? My concerns are the same as for Question 20 on the PISA, which dealt with pattern. The term *model* seems somewhat unusual.

The item that brought a smile to my face listed four fractions—2/10, 2/3, 2/4 and 2/6. A number line that shows zero, one-half and one-whole is given. The students are requested to "Place the following fractions in order on the number line." How might students interpret "in order," and how will the markers tell if the directions were followed?

I found it somewhat satisfying to see items labelled "Number Sense" and "Mental Computations," even if I was curious about the assessment criteria used for the latter. One item about a menu for lunch puzzled me by being labelled "Nonroutine Problem." Could this identification be considered a contradiction?

Conclusion

I could go on discussing specific items to illustrate my major points, but there is no need. It is sad and sobering that these types of items were or are used to assess students' mathematical ability. Items that lack clarity and are inappropriate are not fair because they

do not allow students to show what they know and understand. Much greater care needs to be taken in designing test items. Without that care, the conclusions reached about students and shared with parents lack validity and general statements about students' performance may not be as meaningful as we would like.

After many years of devising test questions, we tend to feel as if we have perfected the craft. But inevitably a student will interpret a question in a way we never anticipated. I am reminded of the sobering comments in a report by Peck, Jencks and Connell (1991). Incorrect answers on tests were followed up with brief interviews, and it was concluded that "52 per cent of a group of students would have been misjudged had test results not been supplemented with these brief interviews." It can be concluded that with poor questions, that percentage will be even higher.

On a lighter note, there are students who do not know the answer to a question but will not hesitate to provide one anyway. I read a newspaper report that included a Grade 4 social studies exam question asking students to state the population of a province. One student wrote, "As for the completely total population of that province I would estimate that I distinctly do not know." The teacher's challenge is to award a mark to this possible future politician.

One of my examination questions asked teachers-to-be to use their own words to define Cartesian product. One lady who had obviously been absent during the discussion of the topic had no idea, but she drew a three-dimensional box with a wide ribbon around it and a big bow. Printed in the corner of the box in bold letters was "Product of Cartesia." I still recall the surprise! It still makes me smile!

Most of the items I designed for one examination must have been appropriate because one creative teacher-to-be wrote the following at the end:

Thank God this test is over
I think I am going to die,
I didn't know the answers
And I couldn't even lie!

Perhaps the creator of the next rhyme could have partially blamed her difficulties on the inappropriateness of some of the examination items. However, with an option like hers, who would want to examine the appropriateness of items?

There was a girl who saw teaching
As a goal for which she was reaching.
After writing a test
Which wasn't her best
She decided, "It's Hawaii for beaching!"

References

Charles, R. and J Lobato. 1998. *Future Basics: Developing Numerical Power*. Monograph. Golden, Colo: National Council of Supervisors of Mathematics.

Liedtke, W. 2003. "Wise/Numerate and/or Test-Wise and/or Otherwise." *Vector* 45, no 1: 13–15.

Peck, D, S Jencks and M Connell. 1991. "Improving Instruction Through Brief Interviews." *Arithmetic Teacher* 37, no 3: 15–17.

Werner Liedtke is a professor emeritus at the University of Victoria in Victoria, British Columbia.