# Reflections on Norm-Referenced vs. Criterion-Referenced Testing in an NCLB Environment

Mary :Lyn Bourque, Ed.D.
Mid-Atlantic Psychometric Services, Inc.
Testimony before the Senate Committee on Education
October 4, 2006

I want to thank the Chair and members of the Senate Committee for inviting me here today to share some thoughts on the issue of norm-referenced testing (NRT) vs. criterion-referenced testing (CRT).

I will try to lay out what I think are the primary differences between these two kinds of assessments, the assets and deficits of each, and provide some examples of the issues I will raise. For purposes of discussion my comments may tend to make everything look black and white – believe me, that is not the case in the real world – there is a lot more gray than we would like to admit perhaps. You should be advised that there may be varying opinions on some of these issues. I will make every effort to be objective in my presentation, offering both sides of the argument if there is one.

Let me begin by saying that psychometrics in general and test development in particular are as much art as they are science. Although all the statistics may give test development the appearance of being very scientific, be assured, that the scientific aspect is often ameliorated by the art. As a test developer there are certain "industry standards" that we try to abide by: (1) reliability is one standard; (2) validity is another; and (3) generalizabilty or score interpretation is a third. Achieving these three standards will be embedded in my remarks today since I believe that meeting these standards is an ethical responsibility of all test developers, be they commercial publishers, private developers, or local, state or national agencies.

In the simplest definition, NRT measures achievement against a standard embodied in a normative group. For example, an NRT yields a measure of achievement that compares an individual student's performance against groups of students in the same classroom,

grade, school, district, state, or nation. The question being answered by the data is simply, "How does the proverbial Johnny's performance in mathematics compare to other students' performance in the reference group X?"

On the other hand, a CRT measures achievement against a standard based on content and embodied in the framework of the assessment. So, for example, a CRT provides a measure of achievement that compares a student's performance against the content standards and reflected in the performance standards (expectations) and the test questions. The question being answered by the data is, "How well is Johnny performing relative to the content standards?"

The obvious question is, can a single test accomplish both goals? Let's leave the answer for a bit later in the discussion. I would like to point out first some of the differences between and among these two general forms of assessment. One cannot tell simply by examining an assessment whether or not it is functioning as an NRT or a CRT. While they might look very much alike at one level, there are marked differences in the purposes of these two tests, in the content frameworks, in how the tests are constructed, and finally in how the results are interpreted. I will examine each of these areas in some detail.

***Test Purpose.*** First, let's explore *test purpose* as one of the four most distinguishing features between NRTs and CRTs. Building on the simple definition provided above for an NRT, it becomes clear that NRTs are designed to look at the overall results of a target population, and compare individual or group performance to that population. The best way to do that is to cover a broad swath of content that will ensure a "spread" of scores, from high to low, across the content of the test. This, of necessity, means that frequently the content domain is sampled for the assessment, and the amount of data that measures specific content is very thin (few test questions measuring that content). In other words, breadth, but not depth.

With CRTs, on the contrary, the purpose is to measure specific content against a standard. This may mean limiting some content, but what is measured is measured well. The results allow test users to make pass/fail decisions, or to answer the question above, "How well is Johnny performing relative to the standard?" This is also the kind of information that can be quite helpful to teachers in the classroom or instructional leaders in schools in judging whether or not their programs are working as intended and well.

An example of what these kinds of data look like can be found in the Appendix. These are real data based on the performance of TX students or a national group on one or more large scale assessments.

NAEP is a good example of an NRT with which you are all familiar. At the policy level NAEP is intended to provide the American public with a snapshot of student performance in a variety of academic subjects on a regularized basis. Exhibit 1 in the Appendix displays the results on the 2005 NAEP reading and mathematics assessments, comparing national and Texas student performance. These kinds of data are helpful in understanding where TX student performance stands in relationship to the Nation's performance.

Since, in the case of NAEP, its purpose is to measure performance against a norming population, the reported data in Exhibit 1 are appropriate for meeting that purpose. The "norming population" is a sample of examinees from across the nation. TX appears to be above the norm in Grades 4 and 8 reading and mathematics. I suspect neither is significant, meaning that those differences are likely caused by chance. The appropriate interpretation of these data is that TX is very much like the nation in its performance on the mathematics and reading content assessed by NAEP.

Exhibit 2 looks somewhat different in that it details TX student performance relative to the content standards on the TAKS. According to the state's website, "TAKS was designed to measure core areas of the state-mandated curriculum, the Texas Essential Knowledge and Skills (TEKS)." In other words, the purpose of the TX assessment is

primarily criterion-referenced, that is, to measure school learning presumably described in a content standards document.

In Exhibit 2 the only "comparison" is that of examinee performance against a standard, reported as the percentage of students meeting the state standard. In the case of the TAKS assessment, the data reported are also appropriate given the test's purpose. TX has set a standard in both reading and mathematics, and the data merely report the percentage of examinees meeting that standard of performance.

***Assessment Framework.*** A second distinguishing feature between NRTs and CRTs is the manner in which the content specifications for the assessment are developed, and the resulting assessment framework. The test development process typically follows a specified protocol that begins with the development of test specifications which outline the test purpose, test content and specific characteristics of the test.

One of the differences between NRTs and CRTs in developing the content of the test is what I call "grain size." Because (as mentioned above) NRTs tend to be survey tests, the articulation of content on an NRT tends to be more global. Even if specific content objectives are part of the document, they still tend to be broad, thus giving greater latitude and flexibility to the item development process later on

On the contrary, the content specifications for CRTs are much more specific, narrow, and of smaller "grain size." Many times, the listing of objectives approaches the specificity of a test question, allowing only a finite sample of test items to be developed that could be aligned with the objective.

It is in this area of content specification that we see the differences between the NRTs survey nature (breadth but not depth) and the CRTs more diagnostic nature (depth, but not breadth). Again an example might help to illuminate these differences.

In Exhibit 3 there is a comparison of the NAEP Grade 4 assessment specifications and the TAKS objectives for one mathematics strand, ***algebra,*** and a single objective, namely, ***Patterns, relations, and algebraic thinking.*** Both assessments cover this one topic, but NAEP does it in a more global way covering a broader swath of content in this topic, while the TAKS is very specific in what is expected of students in grade 4. The reader should keep in mind that the TAKS is administered at sequential grades. Therefore it is likely that some of this same content may be found on the assessments at lower and higher grade levels.

***Item Development/Selection.*** A third area where one can observe differences between these two kinds of tests is in the manner in which the items are developed and/or selected for test forms. As mentioned above, NRTs are intended to report performance of Johnny against some target population. It is, therefore, important that tests be able to distinguish with some reliability the difference between high-performers and low-performers. In other words, the test attempts to "spread performance" along a continuum, sometimes called the normal curve. The "curve" is generated by administering the test to a representative sample of students for whom the test was developed. This is referred to as the "norming" population, and results in a test score scale that extends from low performance to high performance.

The kinds of test items placed on the test allow the spread of examinees on the scale in a reliable/stable way. Selections from the item bank are items that examinees are likely to answer correctly about half of the time (moderate difficulty), with some items easier, and some items more difficult (usually to put a ceiling and floor on the test). Therefore, when the test items are field tested and item selection is done, those building the test take into account not only the *content* of the test item, but how that item performs in the field, namely, the probability of a correct response and another statistic called discrimination. In other words, we want high performers to answer the more difficult items correctly, and low performers to answer the less difficult items correctly, and not the opposite.

If in NRT item selection it is a choice between content and item statistical characteristics, usually statistics win out.

On the contrary, CRTs are developed with a focus on content no matter what. Many items are field tested. Those that appear to be functioning correctly (good stats), are placed in a potential pool of active items. From the pool, those items that are aligned with the content specifications/objectives are selected for inclusion. Focus on content is paramount because score interpretation depends on referencing the student performance to the content criterion of interest, namely the content objectives.

The numbers of items selected for a content objective will vary from NRTs to CRTS as well. For example, if a 5th grade test has one or two items on fractions we will not know in a very reliable way whether or not Johnny has met the standard for fractions. If the test is an NRT and has lots of other items that measure 5th grade math content in general such as decimals, percents, mixed numbers, and rounding, then maybe two questions on fractions is fine. What is really of interest here is whether Johnny is performing as well as other 5th graders. If this were a CRT, and we were measuring Johnny's in-depth knowledge of fractions we would want questions that measured order of magnitude (½ > ⅓), or conversion (½ = 0.5), or operations (½ x ½ = ¼), or other objectives related to the topic of fractions. In other words, we would want to sample the objective in depth with a variety of test questions.

Some examples of item collections from the tests may be helpful.

The NAEP item map in Exhibit 4 is an example of the "spread" that most NRTs will try to achieve as they identify and select items for inclusion. Notice there are only a few items in the map that fall below the Basic level, and slightly more that appear above the Advanced level. The preponderance of the items fall somewhere in between. This is as it should be in order to "spread" the performance of examinees on the scale. It is also as it should be in order to ensure reliable estimates at the NAEP *Proficient* and *Advanced* achievement levels.

In Exhibit 5 two NAEP items are reproduced to demonstrate differences (and some of the vagaries) in item difficulty. The first item falls at 232 on the item map [*Determine next number in given pattern.*] Judged by NAEP to be an item of moderate complexity, it turned out to be a fairly easy item, falling in the middle of the *Proficient* region on the NAEP scale.

The second item comes form the same sub domain as item #1. This item falls at 294 on the item map [*Identify equation to describe pattern given in table.*] It was judged by NAEP to be an item of low complexity (partly because it is very concrete). In the field, it is a more difficult item falling in the *Advanced* region of the NAEP scale.

The point of this display is simply to demonstrate the range of items on an NRT. These items were chosen by NAEP because they met the parameters set down by NAEP for inclusion, including their statistical characteristics and content alignment with the NAEP assessment framework.

The TAKS item specifications in Exhibit 6 demonstrate the focus on content in developing the TAKS assessment. Exhibit 6 reprints the TAKS blueprint for Grade 4 mathematics. There are 7 test questions measuring objective #2 (*Patterns, Relationships, and Algebraic Reasoning*) – more than the NAEP uses for that same content objective. Similarly, there are 11 measuring Objective #1, etc. The TAKS measures the content more in depth and that is appropriate for CRTs. There may be some TALS released items measuring these objectives, but I was unable to locate them on the website.

***Score Interpretation.*** That leads to the final critical distinction between NRTs and CRTs, namely score interpretation. Interpretation of performance on NRTs is referenced back to the norming group, thus the name, norm-referenced tests. Unless the agency using the NRT is very lucky, the coverage of the content in an NRT interpretation will not be very helpful at the local level. If I am a 4$^{th}$ grade teacher, how helpful is it to know that Johnny performed at the 35$^{th}$ percentile in mathematics? I certainly cannot make any

meaningful instructional decisions for the student based on that information. I do not know whether the student is having trouble with basic operations (addition, subtractions, etc.), or with measurement concepts (perimeter problems), or with reading a graph.

However, at the **policy** level norm-referenced score interpretation can be quite helpful. Programs are selected and evaluated, funding is appropriated, textbooks are selected, based not on what Johnny does, but on the performance of the whole. So, school districts may find it quite helpful to have NRT results to look at when trying to evaluate a new math program. Or states certainly want to know whether or not to appropriate funding for a specific professional development effort for teachers. Having broad NRT results to examine for the state's population of $4^{th}$ graders is helpful in making these kinds of decisions.

NRT score interpretations may generalize to broad domains such as $4^{th}$ grade mathematics, or, in some case, to subdomains, such as numbers and operations, measurement, geometry, algebra, etc. These kinds of generalizations are important if looking at the big picture.

CRT score interpretations allow the user to generalize first to the objectives aligned with the test questions, second to the aggregate of those objectives in a subdomain (e.g., numbers and operations), and third perhaps to the whole domain (e.g., mathematics), depending on the size of the item bank and on how the test questions are selected.

At the **instructional** level CRT score generalizations are very helpful. They can re-direct teaching and review; they can provide clues as to what is working in the classroom and what isn't, they can (at least initially) point in the direction of further diagnosis of students with special needs, they can also help a teacher understand where he/she is not strong in particular teaching skills.

***The Conundrum.*** From my perspective, it looks as if the states could be better assisted in their decision-making by both kinds of tests. CRTs are needed to respond with

reliability and validity to all the instructional decisions that must be made almost on a daily basis in order to improve the quality of classroom instruction. How can teachers know if they are meeting the needs of their students unless there is some hard evidence to that effect? On the other hand, how do states make sound policy decisions without seeing the whole picture? Further, how do states know that the "state picture" they see is the "truth?" So I believe that NRTs help to provide that broad brush-stroke vision for the future of education in the state, and it also keeps the states honest, ensuring that the vision is as close to the "truth" as we can get. In other words, NRTs can serve the function of "benchmarking" state standards in an NCLB environment.

***The Temptation.*** Can a single test accomplish both goals? Yes, but usually not well. There are practical constraints on all tests, not the least of which is test length. NRTs are generally considered "survey" tests, that is, a lot of content gets covered, but no content gets covered in depth. CRTs are, by design, intended to measure the test content well, or at least well enough to make important instructional decisions, such as who passes or fails, or who graduates or does not, or who gets promoted or not. It is very difficult, if not impossible, to use an NRT for these types of critical decisions. Although, I must add, that this difficulty does not preclude test publishers from providing users with criterion-referenced interpretations and reports.

The temptation is to try to use a single test for multiple purposes. This is what some in NAEP call the "Christmas Tree" approach. We have a single test (a tree), and we try to hang as many ornaments on it (purposes) as it will hold before falling down under its own weight. It is a very powerful temptation, bearing on testing resources, testing time vs. instructional time, convenience, simplicity of public reporting, and a host of other compelling issues. My advice is, "Don't fall prey." It is far better to do fewer things very well, than a lot of things not so well. Identify your ***Purpose***; prepare the appropriate ***Assessment Frameworks;*** support quality ***item development/selection***; and provide valid ***Score Interpretation*** reports, and in all likelihood Texas will have a systemic program of instruction/assessment it can be proud of.

Whether it is an NRT or CRT assessment will be determined by each choice made along the way.

# Appendix[1]

## Exhibit 1

**2005 NAEP Results in TX and the Nation: Grades 4 and 8 Reading and Mathematics.**

| | | Scale Score | | **Nat**/TX Achievement Level Percent At or Above | | |
|---|---|---|---|---|---|---|
| Subject | Grade | **Nation** Avg. | TX Avg. | Basic | Prof | Adv |
| Math | 4 | **238** | 242 | **80**/87 | **36**/40 | **5**/5 |
| | 8 | **279** | 281 | **69**/72 | **30**/31 | **6**/6 |
| Reading | 4 | **219** | 219 | **64**/64 | **31**/29 | **8**/6 |
| | 8 | **262** | 258 | **73**/69 | **31**/26 | **3**/2 |

## Exhibit 2

**2005 TAKS Results in Grades 4 and 8 Reading and Mathematics: Percent Met Standard**

| | Grade 4 | | Grade 8 | |
|---|---|---|---|---|
| Content | No. Tested | Percent Met | No. Tested | Percent Met |
| | | | | |
| Reading | 273,508 | 79% | 291,845 | 83% |
| | | | | |
| Mathematics | 278,466 | 81% | 291,433 | 61% |

[1] All data are retrieved from the NAEP website for National and TX data, or from the TEA website for state of TX data.

# Exhibit 3

**Comparison of Content Specifications Found in NAEP and TAKS**

| NAEP Grade 4 | TAKS Grade 4 |
|---|---|
| **Objective 1:** Patterns, relations, and functions | **Objective 2:** Patterns, relations, and algebraic thinking |
| a. Recognize, describe, or extend numerical patterns | (4.6) Uses patterns in multiplication and division<br><br>A. Use patterns to develop strategies to remember basic multiplication facts;<br><br>B. Solve division problem related to multiplication facts (fact families) such as $9 \times 9 = 81$ and $81 \div 9 = 9$<br><br>C. Use patterns to multiply by 10 and 100. |
| b. Given a pattern or sequence, construct or explain a rule that can generate the terms of the pattern or sequence. | 4.7 Uses organizational structures to analyze and describe patterns and relationships. Student is expected to describe the relationship between two sets of related data such as ordered pairs in a table. |
| c. Given a description, extend or find a missing term in a pattern or sequence. | |
| d. Create a different representation of a pattern or sequence given a verbal description. | |
| e. Recognize or describe a relationship in which quantities change proportionally. | |

Legend
MC = Multiple Choice
CR = Constructed Response

# Exhibit 4

500

340

- 330

**325** Arrange tiles in different ways to satisfy given condition—**Extended Response** (CR)

- 320

**317** Solve a story problem involving comparison of unit costs—**Extended Response** (CR)

- 310

**304** Identify the appropriate title for a graph **(MC)**

**304** Solve a story problem involving comparison of unit costs—**Satisfactory Response** (CR)

- 300

**294** Identify equation to describe pattern given in table **(MC)**

- 290

**287** Arrange tiles in different ways to satisfy given condition—**Satisfactory Response** (CR)

**286** Identify given measurements on a ruler—**Correct Response** (CR)

**286** Solve a story problem involving comparison of unit costs—**Partial Response** (CR)

**286** Identify number sentence matching a situation **(MC)**

**284** Subtract fractions with common denominators **(MC)**

282 *Advanced*

- 280

**276** Approximate fraction of an hour given minutes **(MC)**

**275** Solve a story problem involving comparison of unit costs—**Minimal Response** (CR)

**273** Solve a story problem involving large numbers (calculator available)—**Correct Response** (CR)

**272** Determine missing numbers in number sentence—**Correct Response** (CR)

**272** Solve a story problem involving multiplication (calculator available) **(MC)**

- 270

**260** Determine the width of a rectangle after it is folded **(MC)**

**260** Arrange tiles in different ways to satisfy given condition—**Partial Response** (CR)

- 260

**258** Represent a situation with an algebraic expression **(MC)**

**254** Identify which figure on grid has greatest area **(MC)**

**253** Complete a bar graph from a description of data **(CR)**

250

- 249 *Proficient*

**247** Determine missing numbers in number sentence—**Partial Response** (CR)

**245** Determine the value of a point on a number line **(CR)**

**242** Identify given measurements on a ruler—**Partial Response** (CR)

**240** Arrange tiles in different ways to satisfy given condition—**Minimal Response** (CR)

- 240

**239** Solve a story problem involving large numbers (calculator available)—**Partial Response** (CR)

**232** [illegible] given [illegible] **(CR)**

- 230

**228** Classify numbers as even or odd—**Correct Response** (CR)

**223** Determine which attribute could be measured with a meter stick **(MC)**

- 220

**219** Subtract two digit numbers to solve a story problem **(MC)**

- 214 *Basic*

**211** Identify which shapes are cylinders **(MC)**

**211** Subtract two digit number from three-digit number **(CR)**

- 210

**203** Identify a number given in expanded notation **(MC)**

**203** Classify numbers as even or odd—**Partial Response** (CR)

- 200

**197** Determine the most likely outcome in a story problem **(MC)**

190

180

170

## NAEP Item Example #1

**Exhibit 5**

3, 6, 5, 8, 7, 10, 9, ?

1. In the number pattern above, what number comes next?

Answer: _____

Question 1

**Mathematical Content Area**: *Algebra* (Sub content classification: )
**Mathematical Complexity**: *Moderate Complexity*

### Mathematical Complexity

#### *Low Complexity*

This category relies heavily on the recall and recognition of previously learned concepts and principles. Items typically specify what the student is to do, which is often to carry out some procedure that can be performed mechanically. It is not left to the student to come up with an original method or solution.

#### *Moderate Complexity*

Items in the moderate-complexity category involve more flexibility of thinking and choice among alternatives than do those in the low-complexity category. They require a response that goes beyond the habitual, is not specified, and ordinarily has more than a single step. The student is expected to decide what to do, using informal methods of reasoning and problem-solving strategies, and to bring together skill and knowledge from various domains.

## NAEP Item Example #2

| □ | △ |
|---|---|
| 4 | 9 |
| 5 | 11 |
| 6 | 13 |
| 7 | 15 |

1. Which rule describes the pattern shown in the table?

A) □ + 5 = △
B) □ + □ = △
C) □ + □ + 1 = △
D) □ + □ + 2 = △

Question 1

**Mathematical Content Area**: *Algebra* (Sub content classification: )
**Mathematical Complexity**: *Low Complexity*

## Exhibit 6

**Texas Assessment of Knowledge and Skills (TAKS) Blueprint for Grade 4 Mathematics**

| TAKS Objectives | Number of Items |
|---|---|
| Objective 1: Numbers, Operations, and Quantitative Reasoning | 11 |
| Objective 2: Patterns, Relationships, and Algebraic Reasoning | 7 |
| Objective 3: Geometry and Spatial Reasoning | 6 |
| Objective 4: Measurement | 6 |
| Objective 5: Probability and Statistics | 4 |
| Objective 6: Mathematical Processes And Tools | 8 |
| **Total Number of Items** | **42** |

## Summary of Major Distinctions between NRTs and CRTs

| | *NRTs* | *CRTs* |
|---|---|---|
| ***Definition*** | Measures achievement against a target group | Measures achievement against content and performance standards |
| ***Test Purpose*** | Measures a broad cross-section of the whole content domain (breadth)<br>*Example: NAEP Exhibit 1* | Measures a narrower cross-section of the content domain<br>*Example: TAKS Exhibit 2* |
| ***Assessment Framework*** | Level of specificity of content more global (larger grain size)<br>Greater flexibility in item development/selection<br>*Example: Exhibit 3* | Level of specificity of content more focused (smaller grain size)<br>More limited flexibility in item development/selection<br>*Example: Exhibit 3* |
| ***Item Development/Selection*** | Items developed/selected to maximize distribution of performance and discrimination among high/low ability examinees<br><br>Employs items of moderate difficulty and maximum discrimination<br><br>Number of items for any specific objective low<br>*Example: NAEP Exhibits 4 & 5* | Items developed/selected that align with content and performance standards<br><br>Eliminates poor items and selects on content<br><br>Number of items for any specific objective can be substantial<br>*Example: TAKS Exhibit 6* |
| ***Score Interpretation*** | Big picture results at level of content domain or subdomian<br>Report by derived scores such as scale scores<br><br>Important for ***policy*** decisions | Focused results reporting both "percentage meeting overall standards" or "percentage mastering a specific objective" or cluster of objectives<br>Important for ***instructional*** decisions |

## Author's Biography

Mary Lyn Bourque (Ed.D., University of Massachusetts, 1979) is currently the Director of Mid Atlantic Psychometric Services, Inc., a group of measurement associates providing consulting services to state and local departments of education, licensing agencies, and other educational testing agencies. Dr. Bourque specializes in all aspects of test development and full-services standard setting, from technical planning to training and execution. From 1989 to 2001, she was chief psychometrician for the National Assessment Governing Board, responsible for policy-related technical issues, particularly standard setting on the National Assessment. Formerly she was on the faculty of the University of Maryland in the Department of Measurement, Statistics and Evaluation. She has also served as Director of Testing for the city of Providence RI, and taught secondary mathematics and sciences in the Boston area. Dr. Bourque is a member of the National Council for Measurement in Education and the American Educational Research Association, and has authored numerous technical reports and articles on applied measurement issues. She has published in *Reading Research Quarterly, Educational Measurement: Issues and Practices, Education,* and was a contributor to recent edited books including *Defending Standardized Testing, A History of NAEP, Setting Performance Standards*, and the *Handbook of Educational Policy.* Her research interests focus on large-scale assessment, standard-setting, and applied measurement issues.