As we begin another millennium of educational testing, you’re probably scratching your head and asking yourself: “Which new millennium are you talking about? The one that began on January 1, 2000, or the one on January 1, 2001?” Who cares? That’s not the point. The topic of this chapter, as you can tell from the title, is Clint Eastwood, and of course, the similarities between standardized testing and a showdown. Let’s be serious.

Before the “official” substance of this chapter begins, we need to be clear on the terminology used in testing. The problem is that during the past decade we witnessed a new set of testing jargon, albeit “assessmenteze,” that has created confusion among educators, parents, and legislators. The terms used from the 1960s–1980s were replaced with new terms that were synonyms, more pretentious jargon, or words with no origins in measurement or even education literature. These changes are listed in Table 1.
Table 1
Assessment Jargon

<table>
<thead>
<tr>
<th>&quot;Out&quot; Terms 1960s–1980s</th>
<th>&quot;In&quot; Terms 1990s–2000s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goals</td>
<td>Content Standards</td>
</tr>
<tr>
<td>Objectives</td>
<td>Outcomes, Indicators,</td>
</tr>
<tr>
<td></td>
<td>Substandards</td>
</tr>
<tr>
<td>Items</td>
<td>Tasks</td>
</tr>
<tr>
<td>Exercises</td>
<td>Performance Tasks</td>
</tr>
<tr>
<td>Essay Items</td>
<td>Constructed-Response Items</td>
</tr>
<tr>
<td>Test</td>
<td>Assessment Tool</td>
</tr>
<tr>
<td>Multiple-Choice Test</td>
<td>Conventional Assessment</td>
</tr>
<tr>
<td>Standardized Test</td>
<td>Necessary Evil</td>
</tr>
<tr>
<td>Criterion-Referenced Test</td>
<td>Standards-Referenced Test</td>
</tr>
<tr>
<td>Non-Multiple-Choice Test</td>
<td>Alternative Assessment</td>
</tr>
<tr>
<td>Performance Test</td>
<td>Authentic Assessment</td>
</tr>
<tr>
<td>Performance Appraisal/ Assessment</td>
<td>Performance Assessment</td>
</tr>
<tr>
<td>Work Sample</td>
<td>Portfolio</td>
</tr>
<tr>
<td>Scoring Criteria (Protocol)</td>
<td>Rubric</td>
</tr>
</tbody>
</table>

Perhaps the most misunderstood and abused term was “standardized test.” It became synonymous with all commercially-developed, norm-referenced, multiple-choice tests. It was regarded as evil by many people outside of the measurement community, particularly those who made a hobby of bashing these tools. To give you a flavor—or, actually, a bad taste in your mouth—of how deeply entrenched this definition of standardized test is in our culture, consider the most recent indictment of testing by Kohn (2000): The Case Against Standard-
ized Testing. As an example of “standardized-test bashing,” it identifies 16 steps for fighting standardized tests:

1. Talk to friends and neighbors at every opportunity: in line at the supermarket; in the dentist’s waiting room; on airplanes; at the hairdresser’s and the playground; at dinner parties and children’s birthday parties.

2. Get in the habit of attending—and speaking out at—school board meetings and other events dealing with educational policy.

3. Let parents know they can write a letter to school administrators or board members expressing concern that test preparation is eclipsing more important learning activities.

4. Write to—or, better yet, get together a delegation of concerned citizens and then visit—your state legislators and other public officials.

5. Write letters to the editor—or, better yet, op-eds—for your local newspaper.

6. Organize a delegation of educators and/or parents and request a meeting with the education reporter and top editors of your paper.

7. Sponsor a forum or teach-in on testing.

8. Print up bumper stickers with slogans such as “STANDARDIZED TESTING IS DUMBING DOWN OUR SCHOOLS.”


10. Remind sympathetic school officials that under no circumstances should they brag about high (or rising) scores.
11. Speak to educational service agencies, universities, and administrators who offer events for teachers that provide advice on raising test scores and teaching to the standards.

12. Invite researchers in the area to commission a survey. When it’s completed, release the results at a press conference.

13. Challenge politicians, corporate executives, and others who talk piously about the need to “raise the bar,” impose “tougher standards,” and ensure “accountability,” to take the tests themselves.

14. Consider filing a lawsuit against the tests, which are potentially vulnerable in many ways.

15. Investigate whether your state has an “opt-out” clause that allows parents to exempt their children from testing just by notifying the authorities.

16. Boycott the tests even where there is no opt-out provision.

(Adapted from pp. 56–59)

Whether you agree or disagree with these recommendations, professionals with at least a basic knowledge of testing should have an idea of what standardized test means to know what they’re bashing. Unfortunately, Kohn doesn’t. An appropriate measurement definition of a standardized test is given below by Mehrens and Lehmann (1991):

A test that gathers performance information under uniform procedures, such as:

a. fixed set of questions;

b. administration with the same directions and time constraints; and
c. carefully specified and uniform scoring procedures. (Adapted from p. 290)

The above definition is applicable to most norm-referenced and criterion-referenced tests, regardless of item format, administered district-, state-, or nation-wide for educational or research purposes. Even computer-adaptive tests (CATs) satisfy all of the elements in the definition except “a.” Standardized has less to do with the type of test, item formats, and examinees, and more to do with administration and scoring procedures.

Another term that won the Emmy Award for “Most Confusing Term Used by Essay Examiners” is rubric, which is Pig Latin for “yellow brick road.” Its dictionary definitions in order of frequency of use are as follows:

1. A class or category; a title;
2. A part of a manuscript or book that appears in decorative red lettering or is otherwise distinguished from the rest of the text;
3. A title or heading of a statute or chapter in a code of law;
4. A direction in a missal, hymnal, or other liturgical book;
5. An authoritative rule or direction;
6. A short commentary or explanation covering a broad subject; and

May I ask you a question? How did we move from those definitions to “explicit scoring criteria for constructed-response item formats?” Is it any wonder there is confusion over the terminology we create? I guess I should be happy that educators who “rubricate” are not called “rubricians” yet. Or are they?
In an effort to lend some order to this chaos, the remainder of this chapter is devoted to the most basic questions asked about testing and my no-nonsense, straightforward, weird, nonrubricated answers. There are two purposes: (1) to suggest some answers to basic testing questions that may provide a context or perspective for various testing issues and (2) to arm you with possible answers to questions that you may be asked about testing. The Q&A format that was so cute in last year’s volume from National Evaluation Systems is repeated again here because I couldn’t think of anything more clever.

Q1. Why do we test?

A1. We test to provide information (i.e., quantitative evidence) to make informed decisions about individuals or programs. (I bet you thought it was to make the examinee’s life miserable.) This is not to say that there are tests administered without a clear purpose or decision in mind. Tests must be decision-driven. The titles of some tests indicate the decisions for which their scores will be used. These are frequently admission, licensure, and certification tests in professions such as teaching, nursing, medicine, and law. With the exception of the National Assessment of Educational Progress (NAEP), the titles of tests designed for students usually designate what they measure (i.e., construct), such as achievement, attitude, ability, intelligence, basic skills, performance, proficiency, and/or understanding of the content domain, rather than the decisions for which they will be used. If you don’t need a particular type of test information to make a decision, don’t test. The test is just a tool, not an end in itself.
Q2. **What types of information do tests provide?**

A2. There are two basic types of information:

1. *Individual* measures of performance
   - score (raw or standard)
   - pass-fail status
2. *Group* measures of performance
   - mean
   - pass-fail percentages

Given the variety of scores that can be reported from a single test, it would be most effective and efficient to specify in advance the types of scores required for decision making. The shotgun approach to score reporting is fun, but undirected and meaningless without a purpose for interpretation and use.

Q3. **When can test information be gathered?**

A3. Test data may be collected at one or more points in time:

- Gathering data at 1 point measures current performance (snapshot).
- Gathering data at 2 points measures change between pre- (baseline) and post-instruction/intervention performance.
- Gathering data at 3 or more points using multiwave, longitudinal data measures trends to evaluate progress over time.

Reports can be prepared for any of the above time points. Each has a particular purpose. Multiwave data are especially useful in evaluating trends in student performance and program effectiveness as well as the financial
value of education reform. That is, are we getting our money’s worth in terms of educational outcomes? Such data are commonplace in daily stock market analysis and economic forecasting. It is even possible to conduct a market analysis of individual performance. Consider the following example:

NES has been providing me with a decent sum of money every year to make a presentation at their annual Chicago FunFest. Bottom line: *Have they gotten their money’s worth?*

**Figure 1**

*Market Analysis of Presentations at NES Conferences Based on the NESDAQ-100 Index*

*NESDAQ - 100 INDEX*
A graph of my performance over the past seven years based on the NESDAQ-100 Index is shown in Figure 1. Here’s a market analysis to answer the question:

NES’s investment strategy from 1994 to 1996 seemed relatively stable when I presented on measurement topics, such as standard setting and multiple-choice item formats, although there was a slight dip in 1996 on the topic of content standards. In 1997 when I spoke on preservice training, the bottom almost fell out. That was “Black Wednesday” at the Chicago conference. NES analysts turned deeply bearish due to my plunging productivity and obviously overvalued assets. Mid-conference corrections were badly needed. In 1998, there was a slight upturn on the topic of teacher preparation, but delivering the presentation with a fever of nearly 102° didn’t generate much confidence. A major upswing occurred in 1999 on the topic of reliability. Back on measurement track, our hero delivered again in 2000. Investments soared. NES analysts turned bullish entering the new millennium.

Q4. What decisions are made on the basis of test results?

A4. A wide range of decisions are made about students, teachers, administrators, and programs on the basis of test results. Examples of these decisions are given below:

Students
- Diagnosis
- Screening
- Placement
- Monitoring Progress (formative)
- Grading (summative)
• Promotion
• High School Graduation

Teachers
• Evaluation/Appraisal
• Promotion
• Salary Increase
• Merit Pay
• Licensure
• Certification

Administrators
• Evaluation/Appraisal
• Certification

Programs
• Instructional Effectiveness
• Curriculum Evaluation
• School Effectiveness
• District Effectiveness
• State Effectiveness
• National Effectiveness
• Institutional Accreditation

Q5. What would happen if we didn’t test?

A5. Could you even imagine a world without tests? Our greatest critics certainly can. Probably the most appropriate form in which we should view that hypothetical situation is that of a multiple-choice item:
If there weren’t any tests, which of the following would be the most serious consequence?

A. Many of us wouldn’t have a job, at least not our present one.

B. NES would be NBS: National Bagel Shoppe of Amherst, starting a *new* Gorth family business.

C. Instead of visiting you during the year, Dick Allan would be making flavored cream cheeses in Amherst.

D. I might be doing stand-up comedy or locked up in a home for the PSYCHOmatically disoriented.

E. We wouldn’t be partying at the Chicago Westin year after year.

Although these are super-serious consequences from our perspectives, what would happen from a student, teacher, administrator, or program perspective? Most likely, all of the types of decisions listed previously in the answer to question 4 would be made anyway, based on less objective evidence, perhaps by direct observation or anecdotal records, or by none at all. Yup, that’s what I said. None! If a decision must be made and no testing information is available, it still will be made.

Interestingly, even when the information is available, people may not be aware of it or choose not to use it for one reason or another when making decisions. In education, this practice is called fact-free or evidence-free evaluation. This also occurs frequently in the realm of consumer purchasing. For example, some people will buy a car spontaneously based on appearance only, while
others who admittedly may be obsessive-compulsive and anal-retentive will research car models extensively for months using Consumer Reports and the Internet before making a decision.

Q6. How do you pick the right test for the decisions to be made?

A6. Once you know the decision to be made, there are two steps:

1. Determine the *type of test results* or data you need for the specific decisions, and

2. *Select or design* the appropriate test (measurement tool) to provide those results. Tests may be either:
   - norm-referenced tests (NRTs) or criterion-referenced tests (CRTs)
   - high-stakes tests or low-stakes tests

These steps are humanly illogical and unnatural. This process essentially requires that we visualize the form and substance of the results that are needed for a decision before the data are collected. The last step is to select or develop the test.

More often, educational practice occurs in the reverse direction. Select a test. Present the scores. Make the decision. Since a test is simply a tool, the purpose must be determined before we know what tool to pick and how to use it. As noted in the answer to question 1, the decision drives everything.
Q7. How do NRTs and CRTs compare?

A7. Norm-referenced and criterion-referenced tests have certain characteristics in common, such as appearance and standardization. However, there are also several important differences: purposes, uses, score interpretations, and psychometric properties. Table 2 outlines the differences in the first three areas; the psychometric differences have been covered in previous NES volumes (e.g., Berk, 2000).

Table 2
Comparison of Test Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Norm-Referenced Test</th>
<th>Criterion-Referenced Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Purpose</td>
<td>a. To spread out examinees along a normal curve</td>
<td>a. To identify examinees who have mastered the content/skills and those who haven’t</td>
</tr>
<tr>
<td></td>
<td>b. To discriminate among individuals</td>
<td>b. To discriminate between examinees who know the content and those who don’t</td>
</tr>
<tr>
<td>2. Use in Instruction</td>
<td>a. Overall performance in a single content domain</td>
<td>a. Pinpoints specific strengths and weaknesses</td>
</tr>
<tr>
<td></td>
<td>b. Encourages competition among examinees and reduces cooperation</td>
<td>b. Eliminates competition and fosters cooperation; everyone works toward achieving the same goal (mastery)</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>3. <strong>Score Interpretation</strong></th>
<th><strong>Norm-Referenced Test</strong></th>
<th><strong>Criterion-Referenced Test</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a. Score is referenced to broad content domain</td>
<td>a. Score is referenced to objectives and well-defined content domain</td>
</tr>
<tr>
<td></td>
<td>b. Score is referenced to norm group or normative score; each examinee is ranked along a normal curve</td>
<td>b. Score is referenced to competency standard (cutoff score)</td>
</tr>
<tr>
<td></td>
<td>c. Individual performance is evaluated in terms of how the norm group performed</td>
<td>c. Performance is evaluated in terms of the standard (e.g., mastery/nonmastery, competent/incompetent), irrespective of how everyone else performed</td>
</tr>
<tr>
<td></td>
<td>d. Encourages use of a quota to determine how many examinees fall into different areas of the curve</td>
<td>d. Every examinee has the same opportunity to learn the content and achieve competence</td>
</tr>
</tbody>
</table>
Q8. What have we learned?

A8. We have probably learned to think in an unnatural way. The order of thinking is depicted below:

```
DECISION

↑

DETERMINE TEST RESULTS

↑

PICK/DESIGN TEST
```

Thinking in this mode for many years can make us a little looney, as evidenced by the style of this chapter. However, it’s psychometrically sound and will increase your survival quotient in the trenches. Should you be tempted to think “naturally” in the reverse, you can easily be blindsided by the decision makers or whacked by the testing curmudgeons. And you certainly do not want to be branded as a “whackee.”
References


