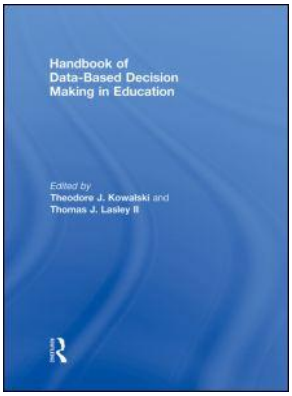

Handbook of Data-Based Decision Making in Education

Theodore J. Kowalski, Thomas J. Lasley

Formative versus Summative Assessments as Measures of Student Learning

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9780203888803.ch16
Robert J. Marzano
Published online on: 13 Oct 2008

Handbook of Data-Based Decision Making in Education

Edited by
Theodore J. Kowalski and
Thomas J. Lasley II

First published 2009
by Routledge
270 Madison Ave, New York, NY 10016

Simultaneously published in the UK
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

This edition published in the Taylor & Francis e-Library, 2008.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk."

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
Handbook of data-based decision making in education / Theodore J. Kowalski & Thomas J. Lasley II, editors.
p. cm.
Includes bibliographic references and index.
1. School management and organization–Decision making–Handbooks, manuals, etc. I. Kowalski, Theodor 1943– II. Lasley II, Thomas J. 1947–
LB2805 .H2862 2008
371.2 22

ISBN 0-203-88880-4 Master e-book ISBN

ISBN10: 0–415–96503–9 (hbk)
ISBN10: 0–415–96504–7 (pbk)
ISBN10: 0–203–88880–4 (ebk)

ISBN13: 978–0–415–96503–3 (hbk)
ISBN13: 978–0–415–96504–0 (pbk)
ISBN13: 978–0–203–88880–3 (ebk)

16

Formative versus Summative Assessments as Measures of Student Learning

Robert J. Marzano
Mid-continent Research Lab for Education and Learning

One can make a case that No Child Left Behind (NCLB) raised the emphasis on assessing and reporting student academic achievement to new levels. Guilfoyle (2006) chronicles the history of NCLB and its heavy reliance on testing. She notes: "The original law provided funding to school districts to help low-income students. Today, NCLB holds Title I schools that receive . . . federal money accountable by requiring them to meet proficiency targets on annual assessments" (p. 8). Guilfoyle (2006) describes the position of the U.S. Department of Education as follows:

> The law requires tests in reading and math for students annually in grades 3–8 and once in high school. In 2005–2006, 23 states that had not yet fully implemented NCLB needed to administer 11.4 million new tests in reading and math. Science testing began in 2007—one test in each of three grade spans must be administered (3–5, 6–9, and 10–12)—the number of tests that states need to administer annually to comply with NCLB is expected to rise to 68 million. (p. 8)

Assessment systems currently in use to fulfill the requirements of NCLB might best be described as "status oriented" in that they reflect the percentage of students who are at specific levels of achievement. Presumably the reason for using a status orientation is to provide no excuse for failure; regardless of the background characteristics of students, regardless of when students enter a particular school, all are expected to succeed. Theoretically, following the basic sentiment of NCLB, a district or school should have or at least approach 100% of students passing every state test at every grade level.

Typically a status approach utilizes summative assessments. McMillan (2007) describes summative assessment as "conducted mainly to monitor and record student achievement and . . . used for school accountability" (p. 1). While the logic behind a summatively based, status approach might be understandable, it is unfair as a method of determining the effectiveness of a district or school for a number of reasons. First, many districts and schools have highly transient populations. Consequently a district or a school that has a transiency rate of 50% is compared to a district or school that has a transiency rate of 5%. Quite obviously the districts and schools with the lower transiency rate will have had more time to work with students than the districts and schools with a rate of 50%. Relative standing in terms of percent of students at or above a specific criterion score on a summative assessment

might be more a function of the stability of the student population than it is the effectiveness of the district and school.

Second, districts and schools have student populations with very different demographics and those demographic differences are strongly related to differences in student achievement (Hedges & Nowell, 1998, 1999; Jacobsen, Olsen, Rice, Sweetland, & Ralph, 2001; Ladewig, 2006). Again a status orientation makes no allowances for differences in student demographics across districts and schools.

Third, summatively based approaches can drain resources from the classroom. Zellmer, Frontier, and Pheifer (2006) analyzed the effects of NCLB reporting requirements on district resources in Wisconsin. They begin their treatment in the following way:

> How do the testing mandates of No Child Left Behind (NCLB) affect schools and students? Last November, while bipartisan politics and philosophical debates continued, 435,000 Wisconsin students sat down for an average of six and one-half hours each and took the expanded Wisconsin Knowledge and Concepts Exam (WKCE) required for NCLB accountability. As the dialogue about the 2007 reauthorization of the Elementary and Secondary Education Act (ESEA) unfolds this fall in the United States, it is imperative that we look beyond the rhetoric and consider the effect of NCLB testing on students and schools. (p. 43)

They explain that the tests require 4.75 to 8.66 hours of administration time annually for each student. This amounted to 1.4 million hours of testing in Wisconsin schools in 2004–2005. They note that when NCLB testing is fully implemented, 2.9 hours of test administration will be required. When special populations are considered, the impact of NCLB testing is even more dramatic. Specifically because teachers are involved in test administration special education students lose 8.5, 7.7, and 6.3 days of instruction at the elementary, middle school, and high school levels respectively. Title I students lose 8.6, 7.9, and 6.3 days at elementary, middle school, and high school. English language learners lose 7.4 days of instruction at elementary, middle school, and high school. Finally, cost of test administration paints an even bleaker picture. State tests cost an average of $33.91 per student. Extrapolating this cost across the 435,000 students in Wisconsin, they estimate that Wisconsin spent $14,700,000 on NCLB-related testing. Similar conclusions have been reported by Cizek (2007). Specifically, he notes that estimates of the cost of testing under NCLB range between $271 million and $757 million for the years 2002 through 2008.

Fourth, status systems that are based on large-scale summative assessments are not designed to be used to help individual students. Abrams (2007) reminds us: "It is important to note that the law [NCLB] only prescribes how schools—not students—should be held accountable" (p. 82). Cizek (2007) further makes the point that large-scale summative assessments are not designed to provide feedback on specific aspects of knowledge and skill within a subject area. He explains that the total score reliability across 40 items for the mathematics portion of the fourth-grade state test in a large midwestern state is .87—certainly an acceptable level of reliability. That test reports pupils' subarea performance using the National Council of Teachers of Mathematics categories: algebra, data analysis and probability, estimation and mental computation, geometry, measurement, number and number relations, patterns, relations, and functions, and problem-solving strategies. Unfortunately the reliabilities for these subscale scores range from .33 to .57. Perhaps even more striking

is the reliability of difference scores between those scales. Cizek provides the example of the reliability for the difference score between algebra and measurements. It is .015. He notes:

> it still might be that the dependability of conclusions about differences in subarea performance is nearly zero. In many cases, a teacher who flipped a coin to decide whether to provide the pupil with focused intervention in algebra (heads) or measurement (tails) would be making that decision about as accurately as the teacher who relied on an examination of subscore differences for the two areas. (p. 104)

For the reasons above as well as others, Barton (2006) has called for an accountability system built on a value-added or growth model:

> If we had an accountability system that truly measured student gain—sometimes called *growth* or *value added*—we could use whether students in any year have gained enough in that school year to show adequate progress. The end goal should not be achieving set scores by 2014. The goal should be reaching a standard for *how much* growth we expect during a school year in any particular subject. (p. 30)

The Role of Formative Assessments

The potential power of a value added or growth model is supported by a considerable amount of research on formative assessment. To illustrate, as a result of analyzing more than 250 studies British researchers Black and Wiliam (1998) report the following conclusions regarding formative assessments:

> The research reported here shows conclusively that formative assessment does improve learning. The gains in achievement appear to be quite considerable, and as noted earlier, amongst the largest ever reported for educational interventions. As an illustration of just how big these gains are, an effect size of 0.7, if it could be achieved on a nationwide scale, would be equivalent to raising the mathematics attainment score of an "average" country like England, New Zealand or the United States into the "top five" after the Pacific rim countries of Singapore, Korea, Japan and Hong Kong. (p. 61)

A value-added approach that is based on formative assessments appears to address many of the shortcomings of a summatively based, status approach. First, it addresses the issue of different transiency rates in that a school could estimate the unique effect it had on a student's learning regardless of when a student entered school. Even if a student were in school for a few months only, the student's knowledge gain could be estimated.

Second, a formatively based, value-added system also addresses the issue of different demographics. A school with a majority of students from higher income homes will most likely have a greater proportion of students at or above a specified level of proficiency than a school with a majority of students from low income families. This situation notwithstanding, the knowledge gain in the low income school might be greater than that in the higher income school.

Third, a formatively based, value-added system might even address some of the resource problems of a status system. This is because it relies on classroom-level

assessments that do not detract from instructional time as do the high stakes state-level tests that are characteristic of status approaches. Assessment data can be gleaned as a regular part of the instructional process as opposed to taking time away from instruction as do state-level tests.

Fourth, a formatively based, value-added system addresses the characteristic inability of large-scale status systems to provide guidance regarding instructional practices for individual students. To this end Marzano and Haystead (in press) have recommended that state standards documents be reconstituted into parsimonious "measurement topics" that form the basis for formative assessment. For example, they suggest the list of measurement topics in mathematics depicted in Figure 16.1.

While it seems evident that a formatively based, value-added system is superior to a summatively based, status system, just how to implement the former is not evident. Some districts and schools use "off-the-shelf" formative assessments developed by standardized test makers. In his article entitled "Phony formative assessments: Buyer beware," Popham (2006) harshly criticizes the unquestioning use of commercially prepared formative assessments. He notes that:

> As news of Black and Wiliam's conclusions gradually spread into faculty lounges, test publishers suddenly began to relabel many of their tests as "formative." This name-switching sales ploy was spurred on by the growing perception among educators that formative assessments could improve their students' test scores and help their schools dodge the many accountability bullets being aimed their way. (p. 86)

He further explains that the impressive results regarding formative assessment relate to classroom assessments—those designed and administered by classroom teachers during their daily interactions with teachers—not to external benchmark assessments. In effect, any external assessment that is not intimately tied to the

Numbers and Operations:
1. Number Sense and Number Systems
2. Basic Addition and Subtraction
3. Basic Multiplication and Division
4. Operations, Computation, and Estimation

Algebra:
5. Basic Patterns
6. Functions and Equations
7. Algebraic Representations and Mathematical Models

Geometry:
8. Lines, Angles, and Geometric Objects
9. Transformations, Congruency, and Similarity

Measurement:
10. Measurement Systems
11. Perimeter, Area, and Volume

Data Analysis and Probability:
12. Data Organization and Interpretation
13. Probability

Figure 16.1 Sample measurement topics.

Source: Adapted from Marzano and Haystead (in press).

classroom by definition violates the tenets of formative assessment. Shepard (2006) makes the same criticism of external formative assessments:

> The research-based concept of formative assessment, closely grounded in classroom instructional processes, has been taken over—hijacked—by commercial test publishers and is used instead to refer to formal testing systems called "benchmark" or "interim assessment systems."
>
> (as cited in Popham, 2006, p. 86)

A similar criticism might be leveled at district-made formative assessments. Specifically, they violate one basic tenet of formative assessment which is that they must allow for both formal and informal judgments of student achievement. As McMillan (2007) explains:

> [Benchmark] assessments, which are typically provided by the district or commercial test publishers, are administered on a regular basis to compare student achievement to "benchmarks" that indicate where student performance should be in relation to what is needed to do well on end-of-year high stakes tests. . . . Although the term *benchmark* is often used interchangeably with *formative* in the commercial testing market, there are important differences. Benchmark assessments are formal, structured tests that typically do not provide the level of detail needed for appropriate instructional correctives. (pp. 2–3)

Clearly, then, a formatively based, value-added system cannot be populated exclusively by district- or school-designed assessments nor can it be populated by commercially prepared assessments. They simply do not satisfy the defining features of formative assessment. What, then, is necessary to develop a comprehensive system of formative assessments?

In their meta-analytic review of research on assessment, Black and Wiliam (1998) defined formative assessment in the following way: "all those activities undertaken by teachers and/or by students which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged" (pp. 7–8). Wiliam and Leahy (2007) describe formative assessment as follows:

> the qualifier *formative* will refer not to an assessment or even to the purpose of an assessment, but rather to the function it actually serves. An assessment is formative to the extent that information from the assessment is fed back within the system and actually used to improve the performance of the system in some way (i.e., that the assessment *forms* the direction of the improvement). (p. 31)

At face value these sentiments seem to characterize formative assessment as involving a wide array of formal and informal techniques for designing and interpreting assessments. This places the classroom teacher clearly at the center of effective formative assessment. Unfortunately, many teacher-designed assessments are not adequate to the task of formative assessment. This is because of the century-old practice of using the 100-point scale.

The Problem with Points

Clearly the most common way teachers design assessments is to use a point or percentage approach. Bock (1997) traces the point system to World War I and the Army Alpha Test. The test required a quick and efficient scoring system that could be applied to multiple-choice items scored as correct or incorrect. Correct items were assigned one point; incorrect items were assigned no points. The summary score on the test was easily computed by forming the ratio of the number of correct items divided by the total number of items and multiplying by 100 to obtain a percentage score.

While the point system has a long history in K-12 education, opinions from experts, common sense, and research indicate that it simply does not work well in a formative approach. Thorndike (1904) commented indirectly on the lack of utility in the point system in the following way:

> If one attempts to measure even so simple a thing as spelling, one is hampered by the fact that there exist no units in which to measure. One may arbitrarily make up a list of words and observe ability by the number spelled correctly. But if one examines such a list one is struck by the inequality of the units. All results based on the equality of any one word with another are necessarily inaccurate. (p. 7)

By inference assigning points to something seemingly as straightforward as spelling words is still highly subjective. If spelling the word *cat* correctly receives one point, how many points are assigned to the word *octopus*?

More recently, Thissen and Wainer (2001) commented on the use of points for large-scale assessments:

> In classroom examinations, combinations of selected-response and constructed-response items have often been scored by arbitrary assignment of a certain number of points for each, but that procedure may not be acceptable for a large-scale testing program, in which scoring may be subject to extensive public scrutiny, professional standards of precision are expected to be met (p. 2).

The problem with points is illustrated in a study by Marzano (2002). In that study five teachers were asked to score the tests of ten students. Prior to scoring those tests each teacher was asked to assign points to each item representing the relative importance of the items. Because of the differential weighing of items students received vastly different scores from the five teachers. For example, one student received a total score of 50 from one teacher and a score of 91 from another teacher.

In short, the practice of differential weighting of items for teacher-designed assessments creates a situation in which a student can receive high scores on one test regarding a specific topic and receive low scores on a subsequent test even though the student has learned during the interim. Weighting easy items higher in the first test than in the second test would create this apparent contradiction—a student has learned but his or her scores have gone down. In short, when teachers design assessments using the 100-point scale, each scale for each assessment will likely be very different due to differential weighting of items.

Moving Away from Points

Conceptually one might say that the solution to the problem of points is a scale that remains constant across all formative assessments. In large-scale assessments this issue is addressed using latent trait or Item Response Theory (IRT) models. Such models postulate a continuum of latent scores (somewhat related to the classical test theory concept of true score) and then mathematically estimate each student's score on the latent continuum. Theoretically, if uni-dimensional, parallel assessments were administered, students' progress on the latent continuum could be computed and tracked. This is the essence of formative assessment—demonstrating progress over time. However, Hattie (1984, 1985) has noted that the ideal of uni-dimensional parallel tests is rarely if ever attained. This conclusion notwithstanding, IRT models are designed to allow for multiple parallel assessments. Thissen and Orlando (2001) explain that

> because the probability of a correct response is a function of the ratio of the proficiency of the person to the difficulty of the item, the item parameters cancel for the ratios of probability-correct for two persons, leaving an *item-free* comparison of their proficiencies. Thus the model makes *objective* or *item-free* statements about the relative likelihood that two persons will respond correctly to an item or a set of items, without any reference to the items themselves. (p. 75)

In an earlier statement Thissen and Orlando note:

> This aspect of IRT means that comparable scores may be computed for examinees who did not answer the same questions, without intermediate equating steps. As a result, an extremely large number of alternate forms of a test may be used. (p. 73)

While it is theoretically possible to use IRT models to construct a wide array of parallel tests to be used as classroom formative assessments, such a task would take an enormous amount of time and resources. To illustrate, consider the 13 measurement topics for mathematics depicted in Figure 16.1. Multiple IRT-based assessments would have to be designed for each topic at each grade level. Even if such assessments could be designed they would not address the need for informal formative assessments that teachers might construct and administer on an ad hoc basis. This seems to be a staple of effective formative assessment. Recall Black and Wiliam's (1998) comment that formative assessment involves a wide variety of activities undertaken by teachers and/or by students which provide feedback to modify the teaching and learning activities.

Another shortcoming of formative assessments based on IRT models is that they provide no information that could guide teachers and students regarding how to improve teaching and learning. Again, this is a staple of formative assessment. While the scores generated by an IRT model line up nicely in terms of a mathematical continuum they are meaningless in terms of specific components of knowledge. One can assume that a student with a latent trait score of 2.5 knows more than a student with a latent trait score of 1.5. However, little can be said about what type of knowledge is possessed by one student versus another. Therefore little can be said about how a student might improve.

A more flexible option to designing IRT-based formative assessments is to articulate a continuum of knowledge as opposed to assuming a latent continuum. At first blush, this seems incompatible with current test theory since a basic assumption underlying both classical test theory and latent trait theory is that the underlying continuum of knowledge need not or cannot be articulated. As Thissen and Orlando (2001) note: "Item response theory is concerned with the measurement of such hypothetical constructs as *ability* and *proficiency*. Because such constructs have no concrete reality, their measurement is by analogy with some directly observable variable" (p. 78). Fortunately, for the purposes of formative assessment, the last decade of the twentieth century witnessed a movement that was designed in part to identify the components of continuums of knowledge within specific subject areas. That movement is the standards movement. Discussing the movement's impact, Glaser and Linn (1993) explain:

> In the recounting of our nation's drive toward educational reform, the last decade of this century will undoubtedly be identified as the time when a concentrated press for national educational standards emerged. The press for standards was evidenced by the efforts of federal and state legislators, presidential and gubernatorial candidates, teachers and subject-matter specialists, councils, governmental agencies, and private foundations. (p. xiii)

Glaser and Linn made their comments at the end of the twentieth century. There is no indication that the standards movement has lost any momentum at the beginning of the twenty-first century. Indeed, over a dozen national standards documents have been created and transformed into state-level standards documents in virtually every state in the union. In effect, the standards movement has provided an unprecedented opportunity to enhance measurement theory in that it provides guidance as to how to construct continuums of knowledge within various subject areas. With continuums of knowledge articulated for a given subject area formal and informal formative assessments could be designed that reference those explicit continuums.

Articulating a Continuum of Knowledge

It is probably impractical to articulate a complete continuum of knowledge within a given subject area. However, in a series of works Marzano (2006; Marzano & Haystead, in press) has offered the scale in Figure 16.2 as a tool for articulating partial continuums of knowledge.

The lowest score value on the scale in Figure 16.2 is a 0.0 representing no knowledge of a given topic—even with help the student demonstrates no understanding or skill relative to the topic. A score of 1.0 indicates that *with help* the student shows partial knowledge of the simpler details and processes as well as the more complex ideas and processes. To be assigned a score of 2.0 the student independently demonstrates understanding of and skill at the simpler details and processes but not the more complex ideas and processes. A score of 3.0 indicates that the student demonstrates understanding of and skill at all the content—simple and complex—*that was explicitly taught in class.* Finally, a score of 4.0 indicates that the student demonstrates inferences and applications that *go beyond what was taught in class.*

Using the scale depicted in Figure 16.2, subject-specific measurement topics at

Score 4.0: In addition to Score 3.0 performance, in-depth inferences and applications that go beyond what was taught.

Score 3.5: In addition to Score 3.0 performance, partial success at inferences and applications that go beyond what was taught.

Score 3.0: No major errors or omissions regarding any of the information and/or processes (simple or complex) that were explicitly taught.

Score 2.5: No major errors or omissions regarding the simpler details and processes and partial knowledge of the more complex ideas and processes.

Score 2.0: No major errors or omissions regarding the simpler details and processes but major errors or omissions regarding the more complex ideas and processes.

Score 1.5: Partial knowledge of the simpler details and processes but major errors or omissions regarding the more complex ideas and processes.

Score 1.0: With help, a partial understanding of some of the simpler details and processes and some of the more complex ideas and processes.

Score 0.5: With help, a partial understanding of some of the simpler details and processes but not the more complex ideas and processes.

Score 0.0: Even with help, no understanding or skill demonstrated.

Figure 16.2 A scale designed for articulating a partial continuum.

every grade level can be written in scale format. To illustrate this, consider Figure 16.3 from Marzano and Haystead (in press).

Figure 16.3 depicts the measurement topic of "atmospheric processes and the water cycle" at the eighth-grade level. Similar scales would be designed for multiple topics in grades kindergarten through high school. Using these scales teachers would design and score formative assessments. That is, teachers would construct score 2.0 items, score 3.0 items, and score 4.0 items. Together, these items would constitute a single formative assessment on a particular topic. Assessments would be scored using the same logic articulated in Figures 16.2 and 16.3. If a student answered all score 2.0 items correctly she would receive a score of 2.0; if the student answered all score 2.0 items correctly and received partial credit on score 3.0 items she would receive a score of 2.5 and so on.

Designing and scoring formative assessments as described above is in keeping with the Fuchs and Fuchs (1986) finding that scoring assessment according to an explicit rule has an effect size of .91. That is, the scales depicted in Figures 16.2 and 16.3 are explicit rules for scoring that can be communicated directly to students, thus providing students with explicit guidance regarding how to improve. Marzano (2002) has found that the reliability of scoring assessments using this system is about three times greater than the reliability of teachers designing and scoring assessments using a 100-point scale.

Tracking Student Progress on Formal and Informal Assessments

With a system of measurement topics in place like that in Figure 16.1 and their accompanying scales like those in Figure 16.3, teachers can design formal and

	Atmospheric Processes and the Water Cycle	
	Grade 8	
Score 4.0	**In addition to score 3.0, in-depth inferences and applications that go beyond what was taught, such as:** • how climatic patterns differ between regions • how people living today impact Earth's atmosphere	
	Score 3.5	In addition to score 3.0 performance, in-depth inferences and applications with partial success.
Score 3.0	**While engaged in tasks that address atmospheric processes and the water cycle, the student demonstrates an understanding of important information such as:** • how water cycle processes impact climatic patterns (temperature, wind, clouds) • the effects of temperature and pressure in different layers of Earth's atmosphere (troposphere, stratosphere, mesosphere, thermosphere) **The student exhibits no major errors or omissions.**	
	Score 2.5	No major errors or omissions regarding the score 2.0 elements and partial knowledge of the score 3.0 elements.
Score 2.0	**No major errors or omissions regarding the simpler details and processes such as:** • recognizing and recalling specific terminology, such as: — climate/climatic pattern — troposphere — stratosphere — mesosphere — thermosphere • recognizing and recalling isolated details, such as: — precipitation can cause temperature to change — the atmosphere of the Earth is divided into five layers **However, the student exhibits major errors or omissions with score 3.0 elements.**	
	Score 1.5	Partial knowledge of the score 2.0 elements but major errors or omissions regarding the score 3.0 elements.
Score 1.0	**With help, a partial understanding of some of the score 2.0 elements and some of the score 3.0 elements.**	
	Score 0.5	With help, a partial understanding of some of the score 2.0 but not the score 3.0 elements.
Score 0.0	**Even with help, no understanding or skill demonstrated.**	

Figure 16.3 Science: Atmospheric processes and the water cycle (Grade 8).

Source: Adapted from Marzano & Haystead (in press).

informal formative assessments and track students' progress. Within such a system assessments could employ traditional formats such as forced choice and constructed response items. Marzano (2006) has observed that score 2.0 items tend to employ forced-choice formats, whereas score 3.0 and 4.0 items tend to employ constructed response formats. Assessment formats could also be quite nontraditional. For example, the scales in Figures 16.2 and 16.3 allow teachers to use a discussion with a particular student as a form of assessment. The teacher would ask questions of the student making sure that score 2.0, score 3.0, and score 4.0 questions were included in the discussion. A final score would be assigned to the informal oral examination again using pattern of responses across item types. Valencia, Stallman, Commeyras, Pearson, and Hartman (1991) have found that discussions like this provide three

times the information about a student's knowledge of academic content as assessments that employ forced-choice and constructed response formats. As Valencia and colleagues note: "On average, 66 percent of the typically relevant ideas students gave during interviews were not tested on any of the . . . [other] measures" (p. 226). This is a rather startling finding from an assessment perspective. It implies that more traditional classroom assessment formats like forced-choice items and essays might not allow students to truly show what they know about a given topic. One of Valencia et al.'s final conclusions was that "a comprehensive view of a person's topical knowledge may well require multiple measures, each of which contributes unique information to the picture" (p. 230).

A system of formal and informal assessments allows a teacher to generate multiple scores for students on measurement topics. Multiple scores allow for a tracking system like that depicted in Figure 16.4.

The first column in Figure 16.4 represents an assessment given by the teacher on October 5. This student received a score of 1.5 on that assessment. The second column represents the assessment on October 12. This student received a score of 2.0 on that assessment and so on. Having each student keep track of his or her scores on learning goals in this fashion provides him or her with a visual tracking of his or her progress. It also allows for powerful discussions between teacher and students. The teacher can discuss progress with each student regarding each learning goal. Also, in a tracking system like this the student and teacher are better able to communicate with parents regarding progress in specific areas of information and skill. Of course, one of the most powerful aspects of tracking as depicted in Figure 16.4 is that it allows for value-added interpretations; students see their progress over time. In a value system

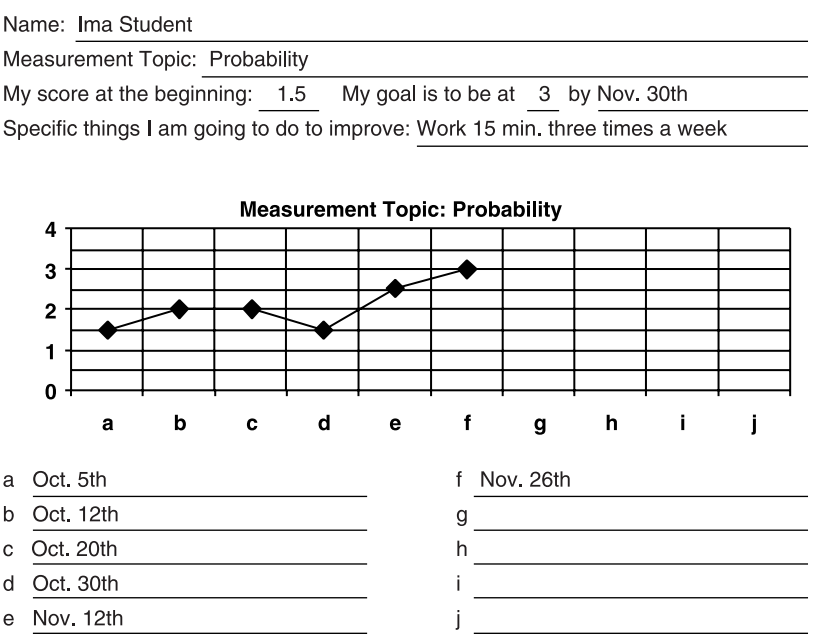

Figure 16.4 Student progress chart.
Source: Adapted from Marzano (2006).

virtually every student will "succeed" in the sense that each student will increase his or her knowledge relative to specific learning goals. One student might have started with a score of 2.0 on a specific learning goal and then increased to a score of 3.5; another student might have started with a 1.0 and increased to a 2.5—both have learned. "Knowledge gain," then, is the currency of student success in an assessment system that is formative in nature. Focusing on knowledge gain also provides a legitimate way to recognize and celebrate success. Covington (1992) has noted that reporting knowledge gain has the potential of stimulating intrinsic motivation for virtually every student.

Summary and Conclusions

This chapter has provided a case for the superiority of formative assessments as measures of student learning. While NLCB has created an emphasis on assessment, a status-oriented, summatively based approach has been the default value. For a variety of reasons a formatively based, value-added approach is superior. By definition formal and informal assessments are needed for a comprehensive system of formative assessments. Unfortunately, use of the 100-point scale has proved to be insufficient to the task of effective formative assessment. A scale was provided that allows for a partial articulation of the specific continuums of knowledge within subject areas. This scale allows teachers to design and score formal and informal formative assessments so that learning can be tracked and knowledge gain quantified and celebrated.

References

Abrams, L. M. (2007). Implications of high-stakes testing for the use of formative classroom assessment. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 79–98). New York: Teachers College Press.

Barton, P. E. (2006). Needed: Higher standards for accountability. *Educational Leadership*, *64*(3), 28–31.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7–75.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement, Issue and Practice*, *16*(4), 21–33.

Cizek, G. J. (2007). Formative classroom and large-scale assessment: Implications for future research and development. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 99–115). New York: Teachers College Press.

Covington, M. V. (1992). *Making the grade: A self-worth perspective on motivation and school reform.* New York: Cambridge University Press.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta analysis. *Exceptional Children*, *53*(3), 199–208.

Glaser, R., & Linn, R. (1993). Foreword. In L. Shepard, R. Glaser, R. Linn, & G. Bohrnstedt (Eds.), *Setting performance standards for student achievement: A report of the National Academy of Education Panel on the evaluation of the NAEP trial state assessment: An evaluation of the 1992 achievement levels* (pp. xiii–xiv). Stanford, CA: National Academy of Education, Stanford University.

Guilfoyle, C. (2006). NCLB: Is there life beyond testing? *Educational Leadership*, *64*(3), 8–13.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, *19*, 49–78.

Hattie, J. (1985). Methodology review: Assessing the unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139–164.

Hedges, L. V., & Nowell, A. (1998). Black–White test score convergence since 1965. In C. Jencks & M. Phillips (Eds.), *The Black–White test score gap* (pp. 149–181). Washington, DC: Brookings Institution Press.

Hedges, L. V., & Nowell, A. (1999). Changes in the Black–White gap in test scores. *Sociology of Education*, *72*, 111–135.

Jacobsen, J., Olsen, C., Rice, J. K., Sweetland, S., & Ralph, J. (2001). *Educational achievement and Black–White inequality*. Washington, DC: National Center for Educational Statistics, Department of Education.

Ladewig, B. G. (2006). *The minority achievement gap in New York State suburban schools since the implementation of NCLB*. Unpublished doctoral dissertation, University of Rochester.

Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education*, *15*(3), 249–268.

Marzano, R. J. (2006). *Classroom assessment and grading that work*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., & Haystead, M. W. (in press). *Making standards useful in the classroom*. Alexandria, VA: Association for Supervision and Curriculum Development.

McMillan, J. H. (2007). Formative assessment: The key to improving student achievement. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 1–7). New York: Teachers College Press.

Popham, W. J. (2006). Phony formative assessments: Buyer beware. *Educational Leadership*, *64*(3), 86–87.

Shepard, L. (2006, June). Panelist presentation delivered at the National Large-Scale Assessment Conference sponsored by the Council of Chief State School Officers, San Francisco, CA.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Erlbaum.

Thissen, D., & Wainer, H. (2001). On overview of *Test Scoring*. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 1–19), Mahwah, NJ: Erlbaum.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurement*. New York: Teachers College Press.

Valencia, S. W., Stallman, A. C., Commeyras, M., Pearson, P. D., & Hartman, D. K. (1991). Four measures of topical knowledge: A study of construct validity. *Reading Research Quarterly*, *26*(3), 204–233.

Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 29–42). New York: Teachers College Press.

Zellmer, M. B., Frontier, A., & Pheifer, D. (2006). What are NCLB's instructional costs? *Educational Leadership*, *64*(3), 43–46.