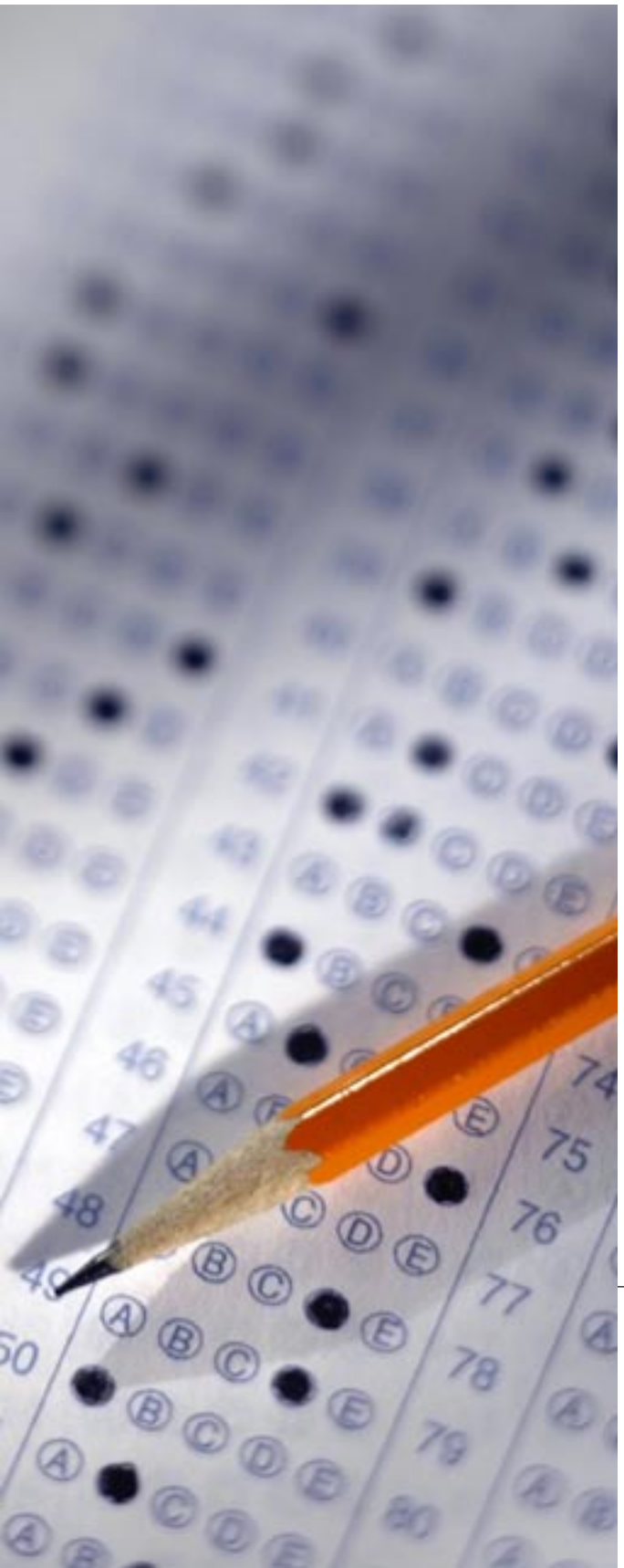
ISSUES&ANSWERS REL 2007–No. 017

REL MID-ATLANTIC
Regional Educational Laboratory
At Pennsylvania State University



The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education

ISSUES & ANSWERS REL 2007–No. 017

REL MID-ATLANTIC
Regional Educational Laboratory
At Pennsylvania State University

The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region

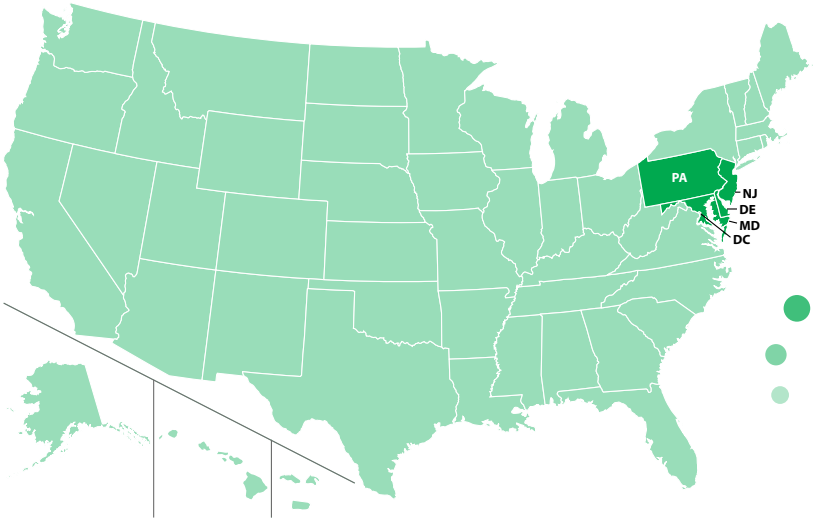
November 2007

Prepared by

Richard S. Brown
University of Southern California

Ed Coughlin
Metiri Group

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences
U.S. Department of Education



REL MID-ATLANTIC
At Pennsylvania State University

Issues & Answers is an ongoing series of reports from short-term Fast Response Projects conducted by the regional educational laboratories on current education issues of importance at local, state, and regional levels. Fast Response Project topics change to reflect new issues, as identified through lab outreach and requests for assistance from policymakers and educators at state and local levels and from communities, businesses, parents, families, and youth. All Issues & Answers reports meet Institute of Education Sciences standards for scientifically valid research.

November 2007

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-06-CO-0029 by Regional Educational Laboratory Mid-Atlantic administered by Pennsylvania State University. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Brown, R. S., & E. Coughlin. (2007). *The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region* (Issues & Answers Report, REL 2007–No. 017). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. Retrieved from http://ies.ed.gov/ncee/edlabs

This report is available on the regional educational laboratory web site at http://ies.ed.gov/ncee/edlabs.

Summary

The predictive validity of selected benchmark assessments used in the Mid-Atlantic Region

This report examines the availability and quality of predictive validity data for a selection of benchmark assessments identified by state and district personnel as in use within Mid-Atlantic Region jurisdictions. The report finds that evidence is generally lacking of their predictive validity with respect to state assessment tests.

Many districts and schools across the United States have begun to administer periodic assessments to complement end-of-year state testing and provide additional information for a variety of purposes. These assessments are used to provide information to guide instruction (formative assessment), monitor student learning, evaluate teachers, predict scores on future state tests, and identify students who are likely to score below proficient on state tests.

Some of these assessments are locally developed, but many are provided by commercial test developers. Locally developed assessments are not usually adequately validated for any of these purposes, but commercially available testing products should provide evidence of validity for the explicit purposes for which the assessment has been developed (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). But the availability of such information and its interpretability by district personnel vary across instruments. When the information is not readily available, it is important for the user to establish such evidence of validity. A major constraint on district testing programs is the lack of resources and expertise to conduct validation studies of this type.

As an initial step in collecting evidence on the validity of district tests, this study focuses on the use of benchmark assessments to predict performance on state tests (predictive validity). Based on a review of practices within the school districts in the Mid-Atlantic Region, this report details the benchmark assessments being used, in which states and grade levels, and the technical evidence available to support the use of these assessments for predictive purposes. The report also summarizes the findings of conversations with test publishing company personnel and of technical reports, administrative manuals, and similar materials.

The key question this study addresses is: What evidence is there, for a selection of commonly used commercial benchmark assessments, of the predictive relationship of each instrument with respect to the state assessment?

The study investigates the evidence provided to establish a relationship between district and state test scores, and between performance on district-administered benchmark assessments and proficiency levels on state assessments (for example, at what cutpoints on benchmark assessments do students tend to qualify as proficient or advanced on state tests?). When particular district benchmark assessments cover only a subset of state test content, the study sought evidence of whether district tests correlate not only with overall performance on the state test but also with relevant subsections of the state test.

While the commonly used benchmark assessments in the Mid-Atlantic Region jurisdictions may possess strong internal psychometric characteristics, the report finds that evidence is generally lacking of their predictive validity with respect to the required state or summative assessments. A review of the evidence for the four benchmark assessments considered—Northwest Evaluation Association's Measures of Academic Progress (MAP; Northwest Evaluation Association, 2003), Renaissance Learning's STAR Math/STAR Reading (Renaissance Learning, 2001a, 2002), Study Island's Study Island (Study Island, 2006a), and CTB/McGraw-Hill's TerraNova (CTB/McGraw-Hill, 2001b)—finds documentation of criterion validity of some sort for three of them (STAR, MAP, and TerraNova), but only one was truly a predictive study and demonstrated strong evidence of predictive validity (TerraNova).

Moreover, nearly all of the criterion validity studies showing a link between these benchmark assessments and state test scores in the Mid-Atlantic Region used the Pennsylvania State System of Assessment (CTB/McGraw-Hill, 2002a; Renaissance Learning, 2001a, 2002) as the object of prediction. One study used the Delaware Student Testing Program test as the criterion measure at a single grade level, and several studies for MAP and STAR were related to the Stanford Achievement Test–Version 9 (SAT–9) (Northwest Evaluation Association, 2003, 2004; Renaissance Learning, 2001a, 2002) used in the District of Columbia. None of the studies showed predictive or concurrent validity evidence for tests used in the other Mid-Atlantic Region jurisdictions. Thus, no predictive or concurrent validity evidence was found for any of the benchmark assessments reviewed here for state assessments in Maryland and New Jersey.

To provide the Mid-Atlantic Region jurisdictions with additional information on the predictive validity of the benchmark assessments currently used, further research is needed linking these benchmark assessments and the state tests currently in use. Additional research could help to develop the type of predictive validity evidence school districts need to make informed decisions about which benchmark assessments correspond to state assessment outcomes, so that instructional decisions meant to improve student learning as measured by state tests have a reasonable chance of success.

November 2007

TABLE OF CONTENTS

This report examines the availability and quality of predictive validity data for a selection of benchmark assessments identified by state and district personnel as in use within Mid-Atlantic Region jurisdictions. The report finds that evidence is generally lacking of their predictive validity with respect to state assessment tests.

THE IMPORTANCE OF VALIDITY TESTING

In a small Mid-Atlantic school district performance on the annual state assessment had the middle school in crisis. For a second year the school had failed to achieve adequate yearly progress, and scores in reading and math were the lowest in the county. The district assigned a central office administrator, "Dr. Williams," a former principal, to solve the problem. Leveraging Enhancing Education Through Technology (EETT) grant money, Dr. Williams purchased a comprehensive computer-assisted instruction system to target reading and math skills for struggling students. According to the sales representative, the system had been correlated to state standards and included a benchmark assessment tool that would provide monthly feedback on each student so staff could monitor progress and make necessary adjustments. A consultant recommended by the publisher of the assessment tool was contracted to implement and monitor the program. Throughout the year the benchmark assessments showed steady progress. EETT program evaluators, impressed by the ongoing data gathering and analysis, selected the school for a web-based profile. When spring arrived, the consultant and the assessment tool were predicting that students would achieve significant gains on the state assessment. But when the scores came in, the predicted gains did not materialize. The data on the benchmark assessments seemed unrelated to those on the state assessment. By the fall the assessment tool, the consultant, and Dr. Williams had been removed from the school.[1]

This story points to the crucial role of predictive validity—the ability of one measure to predict performance on a second measure of the same outcome—in the assessment process (see box 1 for definitions of key terms). The school in this

BOX 1

Key terms used in the report

Benchmark assessment. A benchmark assessment is a *formative assessment,* usually with two or more equivalent forms so that the assessment can be administered to the same children at multiple times over a school year without evidence of practice effects (improvements in scores resulting from taking the same version of a test multiple times). In addition to formative functions, benchmark assessments allow educators to monitor the progress of students against state standards and to predict performance on state exams.

Criterion. A standard or measure on which a judgment may be based.

Criterion validity. The ability of a measure to predict performance on a second measure of the same construct, computed as a correlation. If both measures are administered at approximately the same time, this is described as *concurrent validity.* If the second measure is taken after the first, the ability is described as *predictive validity.*

Formative assessment. An assessment designed to provide information to guide instruction.

Predictive validity. The ability of one assessment tool to predict future performance either in some activity (success in college, for example) or on another assessment of the same construct.

Reliability. The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group.

example had accepted the publisher's claim that performance on the benchmark assessments would predict performance on the state assessment. It did not.

Many districts and schools across the United States have begun to administer periodic assessments to complement end-of-year state testing and provide additional information for a variety of purposes. These assessments are used to provide information to guide instruction (formative assessment), monitor student learning, evaluate teachers, predict scores on future state tests, and identify students who are likely to score below proficient on state tests.

Some of these assessments are locally developed, but many are provided by commercial test developers. Locally developed assessments are not usually adequately validated for any of these purposes, but commercially available testing products should provide validity evidence for the explicit purposes for which the assessment has been developed (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). But the availability of this type of information and its interpretability by district personnel vary across instruments. When such information is not readily available, it is important for the user to establish evidence of validity. A major constraint on district testing programs is the lack of resources and expertise to conduct validation studies of this type.

The most recent edition of *Standards for Educational and Psychological Testing* states that predictive evidence indicates how accurately test data can predict criterion scores, or scores on other tests used to make judgments about student performance, obtained at a later time (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, pp. 179–180). As an initial step in collecting evidence on the validity of district tests, this study focuses on use of benchmark assessments to predict performance on state tests. It investigates whether there is evidence of a relationship between district and state test scores and between performance on locally administered benchmark assessments and proficiency levels on state tests (for example, at what cutpoints on benchmark assessments do students tend to qualify as proficient or advanced

TABLE 1
Mid-Atlantic Region state assessment tests

State or jurisdiction	State assessment test	Source
Delaware	Delaware Student Testing Program (DSTP)	*Retrieved March 14, 2007, from http://www.doe.k12.de.us/AAB/DSTP_publications_2005.html*
District of Columbia	District of Columbia Comprehensive Assessment System (DC CAS)[a]	*Retrieved March 14, 2007, from http://www.k12.dc.us/dcps/data/dcdatahome.html and www.greatschools.com*
Maryland	Maryland High School Assessments (HSA)[b]	*Retrieved March 14, 2007, from http://www.marylandpublicschools.org/MSDE/testing/*
	Maryland School Assessment (MSA)	
New Jersey	New Jersey Assessment of Skills and Knowledge (NJ ASK 3–7)	*Retrieved March 14, 2007, from http://www.state.nj.us/njded/assessment/schedule.shtml*
	Grade Eight Proficiency Assessment (GEPA)	
	High School Proficiency Assessment (HSPA)	
Pennsylvania	Pennsylvania State System of Assessment (PSSA)	*Retrieved March 14, 2007, from http://www.pde.state.pa.us/a_and_t/site/default.asp?g=0&a_and_tNav=|630|&k12Nav=|1141|*

a. In the 2005/06 school year the District of Columbia replaced the Scholastic Aptitude Test–Version 9 with the District of Columbia Comprehensive Assessment System.

b. Beginning with the class of 2009, students are required to pass the Maryland High School Assessments in order to graduate. Students graduating before 2009 must also take the assessments, but are not required to earn a particular passing score.

on state tests?). (Table 1 lists the state assessment tests for each Mid-Atlantic Region jurisdiction.) When a district benchmark assessment covers only a subset of state test content, the study looks for evidence that the assessment correlates not only with overall performance on the state test but also with relevant subsections of the state test.

PURPOSES OF ASSESSMENTS

Assessments have to be judged against their intended uses. There is no absolute criterion for judging assessments. It is not possible to say, for example, that a given assessment is good for any and all purposes; it is only possible to say, based on evidence, that the assessment has evidence of validity for specific purposes. Furthermore, professional assessment standards require that assessments be validated for all their intended uses. A clear statement of assessment purposes also provides essential guidance for test and assessment item developers. Different purposes may require different content coverage, different types of items, and so on. Thus, it is critical to identify how assessment information is to be used and to validate the assessments for those uses.

This study examines the availability and quality of predictive validity data for a selection of benchmark assessments that state and district personnel identified to be in use within the Mid-Atlantic Region jurisdictions (Delaware, District of Columbia, Maryland, New Jersey, and Pennsylvania).

For this review a benchmark assessment is defined as a formative assessment (providing data for instructional decisionmaking), usually with two or more equivalent forms so that the assessment can be administered to the same children at multiple times over a school year without evidence of practice effects (improvements in scores resulting from taking the same version of a test multiple times). In addition to formative functions, benchmark assessments allow educators to monitor the progress of students against state standards and should predict performance on state exams.

Frequently, benchmark assessments are used to identify students who may not perform well on

the state exams or to evaluate how well schools are preparing students for the state exams. These uses may require additional analysis by the districts. The predictive ability of an assessment is not a use but rather a quality of the assessment. For example, college admissions tests are supposed to predict future performance in college, but the tests are used to decide who to admit to college. Part of the evidence of predictive validity for these tests consists of data on whether students who perform well on the test also do well in college. Similar correlation evidence should be obtained for the benchmark assessments used in the Mid-Atlantic Region. That is, do scores on the benchmark assessments correlate highly with state test scores taken at a later date? For example, is there evidence that students who score highly on a benchmark assessment in the fall also score highly on the state assessment taken in the spring?

REVIEW OF PREVIOUS RESEARCH

A review of the research literature shows few published accounts of similar investigations. There is no evidence of a large-scale multistate review of the predictive validity of specific benchmark assessments (also referred to as curriculum-based measures). Many previous studies were narrowly focused, both in the assessment area and the age of students. Many have been conducted with only early elementary school students. For example, researchers studied the validity of early literacy measures in predicting kindergarten through third grade scores on the Oregon Statewide Assessment (Good, Simmons, & Kame'enui, 2001) and fourth grade scores on the Washington Assessment of Student Learning (Stage & Jacobsen, 2001). Researchers in Louisiana investigated the predictive validity of early readiness measures for predicting performance on the Comprehensive Inventory of Basic Skills, Revised (VanDerHeyden, Witt, Naquin, & Noell, 2001), and others reviewed the predictive validity of the Dynamic Indicators of Basic Early Literacy Skills for its relationship to the Iowa Test of Basic Skills for a group of students in Michigan (Schilling, Carlisle, Scott, & Zheng, 2007). McGlinchey and Hixson (2004) studied the predictive validity of curriculum-based measures for student reading performance on the Michigan Educational Assessment Program's fourth-grade reading assessment.

While the commonly used benchmark assessments in the Mid-Atlantic Region jurisdictions may possess strong internal psychometric characteristics, evidence is generally lacking of their predictive validity with respect to the required summative assessments in the Mid-Atlantic Region jurisdictions

Similar investigations studied mathematics. Clarke and Shinn (2004) investigated the predictive validity of four curriculum-based measures in predicting first-grade student performance on three distinct criterion measures in one school district in the Pacific Northwest, and VanDerHeyden, Witt, Naquin, & Noell (2001) included mathematics outcomes in their review of the predictive validity of readiness probes for kindergarten students in Louisiana. Each of these studies focused on the predictive validity of a given benchmark assessment for a given assessment, some of them state-mandated tests. Most of these investigations dealt with the early elementary grades. Generally, these studies showed that various benchmark assessments could predict outcomes such as test scores and need for retention in grade, but there was much variability in the magnitude of these relationships.

ABOUT THIS STUDY

This study differs from the earlier research by reviewing evidence of the predictive validity of benchmark assessments in use across a wide region and by looking beyond early elementary students.

The key question addressed in this study is: What evidence is there, for a selection of commonly used commercial benchmark assessments, of the predictive validity of each instrument with respect to the state assessment?

Based on a review of practices within the school districts in the Mid-Atlantic Region, this report details

the benchmark assessments being used, in which states and grade levels, and the technical evidence available to support the use of these assessments for predictive purposes. The report also summarizes conversations with test publishing company personnel and the findings of technical reports, administrative manuals, and similar materials (see box 2 and appendix A on methodology and data collection).

While the commonly used benchmark assessments in the Mid-Atlantic Region jurisdictions may possess strong internal psychometric characteristics, the report finds that evidence is generally lacking of their predictive validity with respect to the required summative assessments in the Mid-Atlantic Region jurisdictions. A review of the evidence for the four benchmark assessments considered (table 2)—Northwest Evaluation Association's Measures of Academic Progress (MAP; Northwest Evaluation Association, 2003), Renaissance Learning's STAR Math/STAR Reading (Renaissance Learning, 2001a, 2002), Study Island's Study Island (Study Island, 2006a),[2] and CTB/McGraw-Hill's TerraNova (CTB/McGraw-Hill, 2001b)—finds documentation of concurrent or predictive validity of some sort for three of them (STAR, MAP, and TerraNova), but only one was truly a predictive study and demonstrated strong evidence of predictive validity (TerraNova).

Moreover, nearly all of the criterion validity studies showing a link between these benchmark assessments and state outcome measures used the Pennsylvania State System of Assessment (CTB/McGraw-Hill, 2002a; Renaissance Learning, 2001a, 2002) as the criterion measure. One study used the Delaware Student Testing Program test at a single grade level as the criterion measure, and several studies for MAP and STAR were related to the Stanford Achievement Test–Version 9 (SAT–9) (Northwest Evaluation Association, 2003, 2004; Renaissance Learning, 2001a, 2002) used in the District

TABLE 2
Benchmark assessments with significant levels of use in Mid-Atlantic Region jurisdictions

Benchmark assessment	Publisher	Publisher classification	State or jurisdiction
4Sight Math and Reading[a]	Success For All	Nonprofit organization	New Jersey Pennsylvania
Measures of Academic Progress (MAP) Math and Reading	Northwest Evaluation Association	Nonprofit organization	Delaware Maryland New Jersey Pennsylvania District of Columbia
STAR Math and Reading	Renaissance Learning	Commercial publisher	Delaware Maryland New Jersey
Study Island Math & Reading	Study Island	Commercial publisher	Maryland New Jersey Pennsylvania
TerraNova Math and Reading	CTB/McGraw-Hill	Commercial publisher	Maryland New Jersey Pennsylvania

a. The 4Sight assessments were reviewed for this report but were subsequently dropped from the analysis as the purpose of the assessments, according to the publisher, is not to predict a future score on the state assessment but rather "to provide a formative evaluation of student progress that predicts how a group of students would perform if the PSSA [Pennsylvania State System of Assessment] were given on the same day." As a result, it was argued that concurrent, rather than predictive, validity evidence was a more appropriate form of evidence of validity in evaluating this assessment. Users of the 4Sight assessments, as with users of other assessments, are strongly encouraged to use the assessments consistent with their stated purposes, not to use any assessments to predict state test scores obtained at a future date without obtaining or developing evidence of validity to support such use, and to carefully adhere to the Standards for Educational and Psychological Testing (specifically, Standards 1.3 and 1.4) (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999).

Source: Education Week, Quality Counts 2007, individual test publishers' web sites, and state department of education web sites.

BOX 2

Methodology and data collection

This report details the review of several benchmark assessments identified to be in widespread use throughout the Mid-Atlantic Region. The report is illustrative, not exhaustive, identifying only a small number of these benchmark assessments.

Some 40 knowledgeable stakeholders were consulted in the identification process, yielding a list of more than 20 assessment tools. Three criteria were used to make the final selection: the assessments were used in more than one jurisdiction, the assessments were not developed for a single district or small group of districts but would be of interest to many schools and districts in the jurisdictions, and there was evidence, anecdotal or otherwise, of significant levels of use of the assessments within the region.

While not all of the assessments selected are widely used in every jurisdiction, each has significant penetration within the region, as reported through the stakeholder consultations. Short of a large-scale survey study, actual numbers are difficult to derive as some of the publishers of these assessments consider that information proprietary. For the illustrative purposes of this report the less formal identification process is sufficient.

This process yielded four assessments in both reading and mathematics: Study Island's Study Island Math and Reading assessments, Renaissance Learning's STAR Math and STAR Reading assessments, Northwest Evaluation Association's Measures of Academic Progress (MAP) Math and Reading assessments, and CTB/McGraw-Hill's TerraNova Math and Reading assessments.[1]

Direct measures of technical adequacy and predictive validity were collected from December 2006 through February 2007. Extensive efforts were made to obtain scoring manuals, technical reports, and predictive validity evidence associated with each benchmark assessment, but test publishers vary in the amount and quality of information they provide in test manuals. Some test manuals, norm tables, bulletins, and other materials were available online. However, since none of the test publishers provided access to a comprehensive technical manual on their web site and because critical information is often found in unpublished reports, publishers were contacted directly for unpublished measures, manuals, and technical reports. All provided some additional materials.

A benchmark assessment rating guide was developed for reviewing the documentation for each assessment, based on accepted standards in the testing profession and recommendations in the research literature (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Rudner, 1994). Ratings were provided for each element on the rating guide. First, the lead author, a trained psychometrician, rated each element based on the information collected, with scores ranging from 3 ("yes or true") through 1 ("no or not true"). Next, the assessment publishers were asked to confirm or contest the ratings and invited to submit additional information. This second phase resulted in modifications to fewer than 10 percent of the initial ratings, mostly due to the acquisition of additional documentation providing previously unavailable evidence.

Note

1. 4Sight Math and Reading assessments, published by Success for All, were reviewed for this report, but were subsequently dropped from the analysis as the purpose of the assessments, according to the publisher, is not to predict a future score on the state assessment but rather "to provide a formative evaluation of student progress that predicts how a group of students would perform if the PSSA [Pennsylvania State System of Assessment] were given on the same day." As a result, it was argued that concurrent, rather than predictive, validity evidence was a more appropriate form of validity evidence in evaluating this assessment.

of Columbia. None of the studies showed predictive or concurrent validity evidence for tests used in the other Mid-Atlantic Region jurisdictions. Thus, no predictive or concurrent validity evidence was found for any of the benchmark assessments reviewed here for state assessments in Maryland and New Jersey.

REVIEW OF BENCHMARK ASSESSMENTS

This study reviewed the presence and quality of specific predictive validity evidence for a collection of benchmark assessments in widespread use in the Mid-Atlantic Region. The review focused

on available technical documentation along with other supporting documentation provided by the test publishers to identify a number of important components when evaluating a benchmark assessment that will be used for predicting student performance on a later test. These components included the precision of the benchmark assessment scores, use of and rationale for criterion measures for establishing predictive validity, the distributional properties of the criterion scores, if any were used, and the predictive accuracy of the benchmark assessments. Judgments regarding these components were made and reported along with justifications for the judgments. While additional information regarding other technical qualities of the benchmark assessments is provided in appendix C, only a brief description of the assessment and the information on predictive validity evidence is described here.

A rating guide was developed for reviewing the documentation for each benchmark assessment, based on accepted standards in the testing profession and sound professional recommendations in the research literature (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education,1999; Rudner, 1994). The review occurred in multiple stages. First, the lead author, a trained psychometrician, rated each element of the rating guide based on a review of information collected for each assessment. Each element was rated either 3, indicating yes or true; 2, indicating somewhat true; 1, indicating no or not true; or na, indicating that the element was not applicable. In most cases the judgment dealt primarily with the presence or absence of a particular type of information. Professional judgment was employed in cases requiring more qualitative determinations.

To enhance the fairness of the reviews, each profile was submitted to the assessment publisher or developer for review by its psychometric staff. The publishers were asked to confirm or contest the initial ratings and were invited to submit additional information that might better inform the evaluation of that assessment. This second phase resulted in modifications to fewer than 10 percent of the ratings, mostly due to the acquisition of additional documentation providing evidence that was previously unavailable.

Only TerraNova provided evidence of predictive validity, and that was limited to a single state assessment in only a few grades

For each benchmark assessment below a brief summary of the documentation reviewed is followed by two tables. The first table (tables 3, 5, 7, and 9) describes the assessment and its use, and the second table (tables 4, 6, 8, and 10) presents judgments about the predictive validity evidence identified in the documentation. Overall, the evidence reviewed for this set of benchmark assessments is varied but generally meager with respect to supporting predictive judgments on student performance on the state tests used in the Mid-Atlantic Region. Although the MAP, STAR, and TerraNova assessments are all strong psychometrically regarding test score precision and their correlations with other measures, only TerraNova provided evidence of predictive validity, and that was limited to a single state assessment in only a few grades.

Northwest Evaluation Association's Measures of Academic Progress (MAP) Math and Reading assessments

The Northwest Evaluation Association's (NWEA) Measures of Academic Progress (MAP) assessments are computer-adaptive tests in reading, mathematics, and language usage. Several documents were consulted for this review. The first is a 2004 research report by the assessment developer (Northwest Evaluation Association, 2004). The others include the MAP administration manual for teachers (Northwest Evaluation Association, 2006) and the MAP technical manual (Northwest Evaluation Association, 2003). These reports provide evidence of reliability and validity for the NWEA assessments, including reliability coefficients derived from the norm sample (1994, 1999, and 2002) for MAP. With rare exceptions these measures indicate strong interrelationships among the test items for these assessments.[2]

TABLE 3
Northwest Evaluation Association's Measures of Academic Progress: assessment description and use

Item	Description
Publisher	Northwest Evaluation Association
What is measured	Math, reading, language usage
Scoring method	Computer scored—using item response theory (IRT)
Type	Computerized adaptive
Target groups	All students in grades 2–10
Mid-Atlantic Region jurisdictions where used	Delaware, District of Columbia, Maryland, New Jersey, and Pennsylvania
Intended uses	"Both MAP and ALT are designed to deliver assessments matched to the capabilities of each individual student" (technical manual, p. 1).

TABLE 4
Northwest Evaluation Association's Measures of Academic Progress: predictive validity

Criterion	Score[a]	Comments
Is the assessment score precise enough to use the assessment as a basis for decisions concerning individual students?	3	Estimated score precision based on standard error of measurement values suggests the scores are sufficiently precise (generally below .40) for individual students (technical manual, p. 58).
Are criterion measures used to provide evidence of predictive validity?	1	Criterion measures are used to provide evidence of concurrent validity but not of predictive validity.
Is the rationale for choosing these measures provided?	3	Criterion measures are other validated assessments used in the states in which the concurrent validity studies were undertaken (technical manual, p. 52).
Is the distribution of scores on the criterion measure adequate?	3	Criterion measures in concurrent validity studies span multiple grade levels and student achievement.
Is the overall predictive accuracy of the assessment adequate?	1	The overall levels of relationship with the criterion measures are adequate, but they do not assess predictive validity.
Are predictions for individuals whose scores are close to cutpoints of interest accurate?	3	The nature of the computer-adaptive tests allows for equally precise measures across the ability continuum.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

Table 4 indicates that although the MAP scores are sufficiently precise overall and are at the cutpoints of interest, and criterion measures with adequate distributions across grade levels were used in the research studies, these studies did not provide evidence of predictive validity. Rather, the criterion measures are used to provide evidence of concurrent validity. The concurrent relationships are adequate, but they do not provide the type of evidence necessary to support predictive judgments.

Renaissance Learning's STAR Math and Reading assessments

For both STAR Reading and Math assessments reports titled "Understanding Reliability and Validity" provided a wealth of statistical information on reliability and correlations with other outcome measures in the same domain (Renaissance Learning, 2000, 2001b). While evidence is found correlating STAR assessments with a multitude of other measures of mathematics and reading, none of these estimates are of predictive validity. Most are identified as concurrent validity studies, while the rest are labeled "other external validity data" in the technical reports. These data show relationships between the STAR tests and state tests given prior to, rather than subsequent to, the administration of the STAR assessments. Although the documentation provides evidence of relationships between the STAR assessment and many assessments, including three used as state

TABLE 5
Renaissance Learning's STAR: assessment description and use

Item	Description
Publisher	Renaissance Learning
What is measured	STAR Math / STAR Reading
Target groups	All students in grades 1–12
Scoring method	Computer scored using item response theory (IRT)
Type	Computerized adaptive
Mid-Atlantic Region jurisdictions where used	Delaware, Maryland, and New Jersey
Intended use	"First, it provides educators with quick and accurate estimates of students' instructional math levels relative to national norms. Second, it provides the means for tracking growth in a consistent manner over long time periods for all students" (Star Math technical manual, p. 2).

TABLE 6
Renaissance Learning's STAR: predictive validity

Criterion	Score[a]	Comments
Is the assessment score precise enough to use the assessment as a basis for decisions concerning individual students?	3	Adaptive test score standard errors are sufficiently small to use as a predictive measure of future performance.
Are criterion measures used to provide evidence of predictive validity?	1	Numerous criterion studies were found. For Math, however, there were only two studies for the Mid-Atlantic Region (Delaware and Pennsylvania), and neither provided evidence of predictive validity. The Delaware study had a low correlation coefficient (.27).
Is the rationale for choosing these measures provided?	3	Rational for assessments used is clear.
Is the distribution of scores on the criterion measure adequate?	3	Criterion scores span a wide grade range, with large samples.
Is the overall predictive accuracy of the assessment adequate?	1	Criterion relationships vary across grade and outcome, but there is evidence that in some circumstances the coefficients are quite large. The average coefficients (mid-.60s) are modest for Math and higher for Reading (.70–.90). However, these are coefficients of concurrent validity, not predictive validity.
Are predictions for individuals whose scores are close to cutpoints of interest accurate?	3	Because of the computer-adaptive nature of the assessment, scores across the ability continuum can be estimated with sufficient precision.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

assessments in the Mid-Atlantic Region (Pennsylvania State System of Assessment, SAT–9, and Delaware Student Testing Program), these reports provided no evidence of predictive validity for the STAR assessments and the assessments used in the Mid-Atlantic Region.

As with the MAP test, the evidence suggests that the STAR tests provide sufficiently precise scores all along the score continuum and that several studies offer correlations with criterion measures that are well distributed across grades and student ability levels. These correlations are generally stronger in reading than in math. However, while these studies provide evidence of concurrent relationships between the STAR tests and state test measures, they do not provide the kind of validity evidence that would support predictive judgments regarding the STAR test and state tests in the Mid-Atlantic Region.

Study Island's Study Island Math and Reading assessments

The documentation for Study Island's Study Island assessments was limited to the administrator's handbook and some brief research reports (Study Island 2006a, 2006b) on the Study Island web site (www.studyisland.com). Only one report contained information pertaining to the Mid-Atlantic Region, a study comparing proficiency rates on the Pennsylvania State System of Assessment (PSSA) between Pennsylvania schools using Study Island and those not using Study Island. However, since analyses were not conducted relating scores from the Study Island assessments to the PSSA scores, there was no evidence of predictive validity for the Study Island assessments. Nor was evidence of predictive validity found for Study Island assessments and the state assessments used by any of the other Mid-Atlantic Region jurisdictions.

Whereas the documentation reviewed for the MAP and STAR tests provides evidence of test score precision and correlations between these tests and state test scores, documentation for the Study Island assessments lacks any evidence to support concurrent or predictive judgments—there was no evidence of test score precision or predictive validity for this instrument (see table 8).

CTB/McGraw-Hill's TerraNova Math and Reading assessments

CTB/McGraw-Hill's TerraNova assessments had the most comprehensive documentation of the

TABLE 7
Study Island's Study Island: assessment description and use

Item	Description
Publisher	Study Island
What is measured	Math and Reading content standards
Target groups	All K–12 students
Scoring method	Computer scored
Type	Computer delivered
Mid-Atlantic Region jurisdictions where used	Maryland, New Jersey, and Pennsylvania
Intended use	To "help your child master the standards specific to their grade in your state" (administrators handbook, p. 23).

TABLE 8
Study Island's Study Island: predictive validity

Criterion	Score[a]	Comments
Is the assessment score precise enough to use the assessment as a basis for decisions concerning individual students?	1	No evidence of score precision is provided.
Are criterion measures used to provide evidence of predictive validity?	1	No predictive validity evidence is provided.
Is the rationale for choosing these measures provided?	na	
Is the distribution of scores on the criterion measure adequate?	na	
Is the overall predictive accuracy of the assessment adequate?	na	
Are predictions for individuals whose scores are close to cutpoints of interest accurate?	na	

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE 9
CTB/McGraw-Hill's TerraNova: assessment description and use

Item	Description
Publisher	CTB/McGraw-Hill
What is measured	Reading, math, language arts
Target groups	All K–12 students
Scoring method	Item response theory (IRT) models (a triple parameter logistic and a double parameter logistic partial credit)
Type	Nonadaptive
Mid-Atlantic Region jurisdictions where used	Maryland, New Jersey, and Pennsylvania
Intended use	TerraNova consists of three test editions: Survey, Complete Battery, and Multiple Assessment. TerraNova Multiple Assessment contains multiple-choice and constructed-response items providing measures of academic performance in various content areas including reading, language arts, science, social studies, and mathematics. "TerraNova is an assessment system designed to measure concepts, processes, and skills taught throughout the nation" (technical manual, p. 1).

TABLE 10
CTB/McGraw-Hill's TerraNova: predictive validity

Criterion	Score[a]	Comments
Is the assessment score precise enough to use the assessment as a basis for decisions concerning individual students?	3	Adequately small standard errors of measurement reflect sufficient score precision for individual students.
Are criterion measures used to provide evidence of predictive validity?	3	Linking study to Pennsylvania System of School Assessments (PSSA) provides evidence of predictive validity for grades 3–11 in mathematics and reading.
Is the rationale for choosing these measures provided?	3	Linking study documentation provides rationale for using PSSA as outcome.
Is the distribution of scores on the criterion measure adequate?	3	Distribution of PSSA scores shows sufficient variability within and between grade levels.
Is the overall predictive accuracy of the assessment adequate?	3	Linking study provides predictive validity coefficients ranging from .67 to .82.
Are predictions for individuals whose scores are close to cutpoints of interest accurate?	3	Accuracy at the cutpoints is sufficient.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

assessments reviewed for this study. In addition to a robust technical report of more than 600 pages (CTB/McGraw-Hill, 2001b), there was a teachers guide (CTB/McGraw-Hill, 2002b) and a research study linking the TerraNova assessment to the Pennsylvania System of School Assessments (PSSA) test (CTB/McGraw-Hill, 2002a). The technical report exhaustively details the extensive test development, standardization, and validation procedures undertaken to ensure a credible, reliable, and valid assessment instrument. The teachers guide details the assessment development procedure and provides information on assessment content, usage, and score interpretation for teachers. A linking study provided clear and convincing evidence of predictive validity for the TerraNova Reading and Math assessments in predicting student performance on the PSSA for grades 5, 8, and 11 (CTB/McGraw-Hill, 2002a). No predictive validity evidence was found, however,

relating TerraNova assessments to state assessments in Delaware, the District of Columbia, Maryland, or New Jersey.

In contrast to the other measures reviewed in this report, the TerraNova documentation provided support for predictive judgments regarding the use of TerraNova in relation to at least one state test measure in use in the Mid-Atlantic Region. TerraNova possesses good test score precision overall and at the relevant cutpoints. The criterion measure using the predictability or linking study (the Pennsylvania System of School Assessments) has adequate variability both within and across grades (see table 10). Further, the rationale for the use of this assessment is well specified and, more important, the predictive relationships range from adequate (.67) to strong (.82). While this evidence supports the use of TerraNova as a benchmark assessment to predict scores on the Pennsylvania System of School Assessments, comparable evidence to support the use of TerraNova to predict scores on state assessments used in the other Mid-Atlantic Region states is lacking.

Additional research is needed specifically linking these reviewed benchmark assessments with the state assessments currently in use

NEED FOR FURTHER RESEARCH

To provide the Mid-Atlantic Region jurisdictions with adequate information on the predictive validity of the benchmark assessments they are currently using, additional research is needed specifically linking these reviewed benchmark assessments with the state assessments currently in use. Even in the one case where evidence of predictive validity was provided, it is clear that more evidence is needed: "The study presents preliminary evidence of the relationship between the two instruments and does present cause for future investigations based upon the changing nature of the Pennsylvania PSSA" (CTB/McGraw-Hill, 2002a, p. 7).

Additional research is therefore recommended to develop the type of evidence of predictive validity for each of these benchmark assessments with the state assessments for all grade levels tested across the entire Mid-Atlantic Region. Such evidence is crucial for school districts to make informed decisions about which benchmark assessments correspond to state assessment outcomes so that instructional decisions meant to improve student learning, as measured by state tests, have a reasonable chance of success.

In some jurisdictions such studies are already under way. For example, a study is being conducted in Delaware to document the predictive validity of assessments used in that state. To judge the efficacy of remedial programs that target outcomes identified through high-stakes state assessments, the data provided by benchmark assessments are crucial. While some large school districts may have the psychometric staff resources to document the predictive qualities of the benchmark assessments in use, most districts do not.

APPENDIX A
METHODOLOGY

The benchmark assessments included in this review were identified through careful research and consideration; however, it was not the intent to conduct a comprehensive survey of all benchmark assessments in all districts, but rather to identify a small number of benchmark assessments widely used in the Mid-Atlantic Region. Thus, this report is intended to be illustrative, not exhaustive.

Data collection

Some 40 knowledgeable stakeholders were consulted in the five jurisdictions that constitute the Mid-Atlantic Region: Delaware, the District of Columbia, Maryland, New Jersey, and Pennsylvania. Participants included state coordinators and directors of assessment and accountability, state content area coordinators and directors, district assessment and testing coordinators, district administrators of curriculum and instruction, district superintendents, representatives of assessment publishers, and university faculty.

These consultations yielded a list of more than 20 assessment tools currently in use in the region, along with some information on their range of use. Precise penetration data were not available because of publisher claims of proprietary information and the limits imposed on researchers by U.S. Office of Management and Budget requirements. A future curriculum review study will include questions about benchmark assessment usage to clarify this list.

Three criteria were used to make the final selection of benchmark assessments. First, researchers looked for assessments that were used in more than one Mid-Atlantic Region jurisdiction. In New Jersey the Riverside Publishing Company has created a library of assessments known as "NJ PASS" that have been developed around and correlated against the New Jersey State Standards and so would be relevant only to New Jersey districts. The State of Delaware has contracted for the development of assessments correlated with that state's standards. Assessments that are in use in most of the jurisdictions rather than just one were selected.

Second, given the small number of assessments that a study of this limited scope might review, the decision was made to incorporate assessments of interest to a wide range of schools and districts rather than local assessments of interest to a single district or a small group of districts. Thus district-authored assessments were excluded. Maryland, for example, has a history of rigorous development of local assessments and relies heavily on them for benchmarking. Local assessments might be included in a future, more comprehensive study.

Finally, researchers looked for assessments for which there was evidence, anecdotal or otherwise, of significant levels of use within the region. In some cases, a high level of use was driven by state support. In other cases, as with the Study Island Reading and Math assessments, adoption is driven by teachers and schools.

While not all of the assessments selected are widely used in every state, each has significant penetration within the region, as reported through the consultations with stakeholders. Short of a large-scale survey study, actual numbers are difficult to derive as some of the publishers of these assessments consider that information to be proprietary. For the illustrative purposes of this report, the less formal identification process employed is considered adequate.

This process yielded five assessments in both reading and mathematics: Northwest Evaluation Association's Measures of Academic Progress (MAP) Math and Reading assessments; Renaissance Learning's STAR Math and STAR Reading assessments; Study Island's Study Island Math and Reading assessments; TerraNova Math and Reading assessments; and Success For All's 4Sight Math and Reading assessment. The 4Sight assessments were reviewed for this report but were subsequently dropped from the analysis since the purpose of

the assessments, according to the publisher, is not to predict a future score on state assessments but rather "to provide a formative evaluation of student progress that predicts how a group of students would perform if the [state assessment] were given on the same day." As a result, it was argued that concurrent, rather than predictive, validity evidence was a more appropriate form of validity evidence in evaluating this assessment.

For the review of the benchmark assessments summarized here, direct measures (rather than self-report questionnaires) of technical adequacy and predictive validity were collected from December 2006 through February 2007. National standards call for a technical manual to be made available by the publisher so that any user can obtain information about the norms, reliability, and validity of the instrument as well as other relevant topics (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 70). Extensive efforts were made to obtain scoring manuals, technical reports, and predictive validity evidence associated with each benchmark assessment. Because assessment publishers vary in the amount and quality of information they provide in test manuals, this review encompasses a wide range of published and unpublished measures obtained from the four assessment developers.

Two sources of data were used to collect information: the publishers' web sites and the publishers themselves. Each publisher's web site was searched for test manuals, norm tables, bulletins, and other materials. Technical and administrative information was found online for STAR Math and Reading, MAP Math and Reading, and TerraNova Math and Reading. None of the assessment publishers provided access to a comprehensive technical manual on their web sites, however.

Since detailed technical information was not available online, and because unpublished reports often contain critical information, publishers were contacted directly. All provided some additional materials. Table A1 provides details on the availability of test manuals and other relevant research used in this review, including what information was available online and what information was available only by request in hard copy.

Rating

A rating guide, based on accepted standards in the testing profession and sound professional recommendations in the research literature (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education,1999; Rudner, 1994), was developed for reviewing the documentation for each benchmark assessment. First, the lead

TABLE A1
Availability of assessment information

Type of information	Measures of Academic Progress (MAP)	STAR	Study Island	TerraNova
Available online (on test developers' web site)				
Technical manual				
Test manual (users guide)	✓	✓		✓
Predictive validity research and relevant psychometric information				
Hard copy materials provided on request				
Technical manual	✓	✓		✓
Test manual (users guide)			✓	
Predictive validity research and relevant psychometric information		✓		

author, a trained psychometrician, rated each element on the rating guide based on a review of information collected for each assessment. Each element was rated either 3, indicating yes or true; 2, indicating somewhat true; 1, indicating no or not true; or na, indicating that the element was not applicable. Each profile was submitted to the assessment publisher or developer for review by its psychometric staff. The publishers were asked to confirm or contest the initial ratings and invited to submit additional information that might better inform the evaluation of that assessment. This second phase resulted in modifications to fewer than 10 percent of the ratings, mostly due to the acquisition of additional documentation providing specific evidence previously unavailable.

APPENDIX B
GLOSSARY

Benchmark assessment. A benchmark assessment is a formative test of performance, usually with multiple equated forms, that is administered at multiple times over a school year. In addition to formative functions, benchmark assessments allow educators to monitor the progress of students against state standards and to predict performance on state exams.

Computerized adaptive tests. A computer-based, sequential form of individual testing in which successive items in the test are chosen based primarily on the psychometric properties and content of the items and the test taker's response to previous items.

Concurrent validity. The relationship of one measure to another that assesses the same attribute and is administered at approximately the same time. See *criterion validity.*

Construct validity. A term used to indicate the degree to which the test scores can be interpreted as indicating the test taker's standing on the theoretical variable to be measured by the test.

Content validity. A term used in the 1974 *Standards* to refer to an aspect of validity that was "required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent" (American Psychological Association, 1974, p. 28). In the 1985 *Standards* the term was changed to *content-related evidence,* emphasizing that it referred to one type of evidence within a unitary conception of validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985). In the current *Standards*, this type of evidence is characterized as "evidence based on test content" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

Correlation. The tendency for certain values or levels of one variable to occur with particular values or levels of another variable.

Correlation coefficient. A measure of association between two variables that can range from –1.00 (perfect negative relationship) to 0 (no relationship) to +1.00 (perfect positive relationship).

Criterion. A standard or measure on which a judgment may be based.

Criterion validity. The ability of a measure to predict performance on a second measure of the same outcome, computed as a correlation. If both measures are administered at approximately the same time, this is described as *concurrent validity.* If the second measure is taken after the first, the ability is described as *predictive validity.*

Curriculum-based measure. A set of measures tied to the curriculum and used to assess student progress and to identify students who may need additional or specific instruction.

Decision consistency. The extent to which an assessment, if administered multiple times to the same respondent, would classify the respondent in the same way. For example, an instrument has strong decision consistency if students classified as proficient on one administration of the assessment would be highly likely to be classified as proficient on a second administration.

Formative assessment. An assessment designed to provide information to guide instruction.

Internal consistency. The extent to which the items in an assessment that are intended to measure the same outcome or construct do so consistently.

Internal consistency coefficient. An index of the reliability of test scores derived from the statistical interrelationships of responses among item responses or scores on separate parts of a test.

Item response. The correct or incorrect answer to a question designed to elicit the presence or absence of some trait.

Item response function (IRF). An equation or the plot of an equation that indicates the probability of an item response for different levels of the overall performance.

Item response theory (IRT). Test analysis procedures that assume a mathematical model for the probability that a test taker will respond correctly to a specific test question, given the test taker's overall performance and the characteristics of the test questions.

Item scaling. A mathematical process through which test items are located on a measurement scale reflecting the construct the items purport to measure.

Norms. The distribution of test scores of some specified group. For example, this could be a national sample of all fourth graders, a national sample of all fourth-grade males, or all fourth graders in some local district.

Outcome. The presence or absence of an educationally desirable trait.

Predictive accuracy. The extent to which a test accurately predicts a given outcome, such as designation into a given category on another assessment.

Predictive validity. The ability of one assessment tool to predict future performance either in some activity (a job, for example) or on another assessment of the same construct.

Rasch model. One of a family of mathematical formulas, or item response models, describing the relationship between the probability of correctly responding to an assessment item and an individual's level of the trait being measured by the assessment item.

Reliability. The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group.

Reliability coefficient. A coefficient of correlation between two administrations of a test or among items within a test. The conditions of administration may involve variations in test forms, raters, or scorers or the passage of time. These and other changes in conditions give rise to varying descriptions of the coefficient, such as parallel form reliability, rater reliability, and test-retest reliability.

Standard errors of measurement. The standard deviation of an individual's observed scores from repeated administrations of a test (or parallel forms of a test) under identical conditions. Because such data cannot generally be collected, the standard error of measurement is usually estimated from group data.

Test documents. Publications such as test manuals, technical manuals, users guides, specimen sets, and directions for test administrators and scorers that provide information for evaluating the appropriateness and technical adequacy of a test for its intended purpose.

Test norming. The process of establishing normative responses to a test instrument by administering the test to a specified sample of respondents, generally representative of a given population.

Test score precision. The level of test score accuracy, or absence of error, at a given test score value.

Validity. The degree to which evidence and theory support specific interpretations of test scores entailed by proposed uses of a test.

Variation in test scores. The degree to which individual responses to a particular test vary across individuals or administrations.

APPENDIX C
DETAILED FINDINGS OF BENCHMARK ASSESSMENT ANALYSIS

Findings for Northwest Evaluation Association's Measures of Academic Progress (MAP) Math and Reading assessments

TABLE C1

Northwest Evaluation Association's Measures of Academic Progress: reliability coefficients

Reliability coefficient[a]	Coefficient value	Interpretation
✓ Internal consistency	.92–.95	These values reflect strong internal consistency.
✓ Test-retest with same form	.79 –.94	Almost all coefficients above .80. Exceptions in grade 2.
✓ Test-retest with equivalent forms	.89–.96	Marginal reliabilities calculated from norm sample (technical manual, p. 55).
✓ Item/test information (IRT scaling)		Uses Rasch model.
✓ Standard errors of measurement (SEM)	2.5–3.5 in Rochester Institute of Technology (RIT) scales (or .25–.35 in logit values)	These values reflect adequate measurement precision. Scores typically range from 150–300 on the RIT scale.
Decision consistency		

Note: See appendix B for definitions of terms.

a. Checkmarks indicate reliability information that is relevant to the types of interpretations being made with scores from this instrument.

TABLE C2

Northwest Evaluation Association's Measures of Academic Progress: predictive validity

Criterion	Score[a]	Comments
Is the assessment score precise enough to use the assessment as a basis for decisions concerning individual students?	3	Estimated score precision based on SEM values suggests the scores are sufficiently precise for individual students (technical manual, p. 58).
Are criterion measures used to provide evidence of predictive validity?	1	Criterion measures are used to provide evidence of concurrent validity but not of predictive validity.
Is the rationale for choosing these measures provided?	3	Criterion measures are other validated assessments used in the states in which the concurrent validity studies were undertaken (technical manual, p. 52).
Is the distribution of scores on the criterion measure adequate?	3	Criterion measures in concurrent validity studies span multiple grade levels and student achievement.
Is the overall predictive accuracy of the assessment adequate?	1	The overall levels of relationship with the criterion measures are adequate, but they do not indicate evidence of predictive validity.
Are predictions for individuals whose scores are close to cutpoints of interest accurate?	3	The nature of the computer-adaptive tests allows for equally precise measures across the ability continuum.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C3

Northwest Evaluation Association's Measures of Academic Progress: content/construct validity

Criterion	Score[a]	Comments
Is there a clear statement of the universe of skills represented by the assessment?	2	No clear statement of universe of skills in reviewed documents, but there are vague statements about curriculum coverage with brief examples in technical manual. No listing of learning objectives is provided in reviewed documentation.
Was sufficient research conducted to determine desired assessment content and evaluate content?	2	Not clear from reviewed documentation. However, the content alignment guidelines detail a process that likely ensures appropriate content coverage.
Is sufficient evidence of construct validity provided for the assessment?	3	Concurrent validity estimates are provided in the technical manual and the process for defining the test content is found in the content alignment guidelines.
Is adequate criterion validity evidence provided?	3	Criterion validity evidence in the form of concurrent validity is provided for a number of criteria.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C4

Northwest Evaluation Association's Measures of Academic Progress: administration of the assessment

Criterion	Score[a]	Comments
Are the administration procedures clear and easy to understand?	3	Procedures are clearly explained in the technical manual (pp. 36–39).
Do the administration procedures replicate the conditions under which the assessment was validated and normed?	3	Norm and validation samples were obtained using the same administration procedure outlined in the technical manual.
Are the administration procedures standardized?	3	Administration procedures are standardized in the technical manual.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C5

Northwest Evaluation Association's Measures of Academic Progress: reporting

Criterion	Score[a]	Comments
Are materials and resources available to aid in interpreting assessment results?	3	Materials to aid in interpreting results are in the technical manual (pp. 44–45) and available on the publisher's web site.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

Findings for Renaissance Learning's STAR Math and Reading assessments

TABLE C6

Renaissance Learning's STAR: reliability coefficients

Reliability coefficient[a]	Coefficient value	Interpretation
✓ Internal consistency	.77–.88 Math .89–.93 Reading	Strong internal consistency.
✓ Test-retest with same form	.81–.87 Math .82–.91 Reading	Strong stability.
✓ Test-retest with equivalent forms	.72–.79 Math .82–.89 Reading	Strong equivalence across forms.
✓ Item/test information (IRT scaling)		Both tests use the Rasch model.
✓ Standard errors of measurement (SEM)	Average 40 (Math scale) Average 74 (Reading scale)	Math scale ranges from 1 to 1,400 through linear transformation of the Rasch scale. Reading scale ranges from 1 to 1,400, but the Rasch scale is transformed through a conversion table. For reading this equates to roughly a .49 in an IRT scale, or classical reliability of approximately .75.
Decision consistency		

Note: See appendix B for definitions of terms.

a. Checkmarks indicate reliability information that is relevant to the types of interpretations being made with scores from this instrument.

TABLE C7

Renaissance Learning's STAR: content/construct validity

Criterion	Score[a]	Comments
Is there a clear statement of the universe of skills represented by the assessment?	3	Content domain well specified for both Math and Reading.
Was sufficient research conducted to determine desired assessment content and evaluate content?	3	Content specifications well documented for both Math and Reading.
Is sufficient evidence of construct validity provided for the assessment?	3	Construct validity evidence provided for both Math and Reading assessments in technical manuals.
Is adequate criterion validity evidence provided?	3	Criterion validity estimates are provided for 28 tests in Math (276 correlations) and 26 tests in Reading (223 correlations).

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C8

Renaissance Learning's STAR: appropriate samples for assessment validation and norming

Criterion	Score[a]	Comments
Is the purpose of the assessment clearly stated?	3	The purpose for both Math and Reading assessments is clearly stated in the documentation.
Is a description of the framework for the assessment clearly stated?	3	The framework for the Math assessment is well delineated in the technical manual, along with a list of the objectives covered. The STAR Reading assessment technical manual states, "After an exhaustive search, the point of reference for developing STAR Reading items that best matched appropriate word-level placement information was found to be the 1995 updated vocabulary lists that are based on the Educational Development Laboratory's (EDL) A Revised Core Vocabulary (1969)" (p. 11).
Is there evidence that the assessment adequately addresses the knowledge, skills, abilities, behavior, and values associated with the intended outcome?	3	Yes, a list of objectives is provided in the technical manual for STAR Math. For STAR Reading the measures are restricted to vocabulary and words in the context of an authentic text passage.
Were appropriate samples used in pilot testing?	3	Sufficient samples were used in assessment development processes (calibration stages).
Were appropriate samples used in validation?	3	Sufficient samples were used in validation.
Were appropriate samples used in norming?	3	Norm samples are described in detail in the technical manual.
If normative date is provided, was the norm sample collected within the last five years?	2	A norm sample was collected in spring 2002 for Math, but in 1999 for Reading.
Are the procedures associated with the gathering of the normative data sufficiently well described so that procedures can be properly evaluated?	3	The norming procedure is well described in the technical manuals for both assessments.
Is there sufficient variation in assessment scores?	3	Scores are sufficiently variable across grades as indicated by scale score standard deviations in the technical manuals from calibration or norm samples.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C9

Renaissance Learning's STAR: administration of the assessment

Criterion	Score[a]	Comments
Are the administration procedures clear and easy to understand?	3	Procedures are defined in the technical manual (pp. 7–8).
Do the administration procedures replicate the conditions under which the assessment was validated and normed?	3	Procedures appear consistent with procedures used in norm sample.
Are the administration procedures standardized?	3	Administration procedures are standardized (technical manuals, pp. 7–8).

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C10
Renaissance Learning's STAR: reporting

Criterion	Score[a]	Comments
Are materials and resources available to aid in interpreting assessment results?	3	Information on assessment score interpretation provided in technical documentation.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

Findings for Study Island's Study Island Math and Reading assessments

TABLE C11
Study Island's Study Island: reliability coefficients

Reliability coefficient[a]	Coefficient value	Interpretation
Internal consistency	None	
Test-retest with same form	None	
Test-retest with equivalent forms	None	
Item/test information (IRT scaling)	None	
Standard errors of measurement	None	
Decision consistency	None	

Note: See appendix B for definitions of terms.

a. Checkmarks indicate reliability information that is relevant to the types of interpretations being made with scores from this instrument.

TABLE C12
Study Island's Study Island: content/construct validity

Criterion	Score[a]	Comments
Is there a clear statement of the universe of skills represented by the assessment?	3	Statement indicates entirety of state standards.
Was sufficient research conducted to determine desired assessment content and evaluate content?	1	None provided.
Is sufficient evidence of construct validity provided for the assessment?	1	None provided.
Is adequate criterion validity evidence provided?	1	No criterion validity evidence is provided.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C13

Study Island's Study Island: appropriate samples for assessment validation and norming

Criterion	Score[a]	Comments
Is the purpose of the assessment clearly stated?	3	Yes, in the administrators handbook.
Is a description of the framework for the assessment clearly stated?	3	Framework relates to state standards.
Is there evidence that the assessment adequately addresses the knowledge, skills, abilities, behavior, and values associated with the intended outcome?	1	No validation evidence is provided.
Were appropriate samples used in pilot testing?	na	No pilot testing information provided.
Were appropriate samples used in validation?	na	No validation information provided.
Were appropriate samples used in norming?	na	No normative information provided.
If normative date is provided, was the norm sample collected within the last five years?	na	None provided.
Are the procedures associated with the gathering of the normative data sufficiently well described so that procedures can be properly evaluated?	na	No normative information provided.
Is there sufficient variation in assessment scores?	na	None provided.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C14

Study Island's Study Island: administration of the assessment

Criterion	Score[a]	Comments
Are the administration procedures clear and easy to understand?	3	Outlined in the administrators handbook (p. 20).
Do the administration procedures replicate the conditions under which the assessment was validated and normed?	na	
Are the administration procedures standardized?	3	Computer delivers assessments in a standardized fashion.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C15

Study Island's Study Island: reporting

Criterion	Score[a]	Comments
Are materials and resources available to aid in interpreting assessment results?	3	The administrators handbook and web site offer interpretative guidance.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

Findings for CTB/McGraw-Hill's TerraNova Math and Reading assessments

TABLE C16

CTB/McGraw-Hill's TerraNova: reliability coefficients

Reliability coefficient[a]	Coefficient value	Interpretation
✓ Internal consistency	.80 – .95	Strong internal consistency.
Test-retest with same form		
✓ Test-retest with equivalent forms	.67–.84	Moderate to strong evidence of stability for various grade levels.
Item/test information (IRT scaling)		
✓ Standard errors of measurement (SEM)	2.8 – 4.5	Variability in SEMs across grade, but standard errors are sufficiently small. These standard errors equate to roughly .25–.33 standard deviation units for the test scale. Scores typically range from 423 to 722 in reading in grade 3 and from 427 to 720 in math in grade 3.
✓ Decision consistency:		Generalizability coefficients exceed .86.

Note: See appendix B for definitions of terms.

a. Checkmarks indicate reliability information that is relevant to the types of interpretations being made with scores from this instrument.

TABLE C17

CTB/McGraw-Hill's TerraNova: content/construct validity

Criterion	Score[a]	Comments
Is there a clear statement of the universe of skills represented by the assessment?	3	The domain tested is derived from careful examination of content of recently published textbook series, instructional programs, and national standards publications (technical manual, p. 17).
Was sufficient research conducted to determine desired assessment content and/or evaluate content?	3	"Comprehensive reviews were conducted of curriculum guides from almost every state, and many districts and dioceses, to determine common educational goals" (technical manual, p. 17).
Is sufficient evidence of construct validity provided for the assessment?	3	Construct validity evidence provided in technical manual (pp. 32–58).
Is adequate criterion validity evidence provided?	3	Linking study to Pennsylvania State System of Assessment provides evidence of predictive validity for grades 3–11.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C18

CTB/McGraw-Hill's TerraNova: appropriate samples for test validation and norming

Criterion	Score[a]	Comments
Is the purpose of the assessment clearly stated?	3	Purpose clearly stated in the technical manual (p.1).
Is a description of the framework for the assessment clearly stated?	3	The framework is described in the development process in the technical manual (p. 18).
Is there evidence that the assessment adequately addresses the knowledge, skills, abilities, behavior, and values associated with the intended outcome?	3	Construct validity evidence provided in the technical manual (pp. 32–58).

(CONTINUED)

TABLE C18 (CONTINUED)
CTB/McGraw-Hill's TerraNova: appropriate samples for test validation and norming

Criterion	Score[a]	Comments
Were appropriate samples used in pilot testing?	3	"The test design entailed a target N of at least 400 students in the standard sample and 150 students in each of the African-American and Hispanic samples for each level and content area. More than 57,000 students were involved in the TerraNova tryout study" (technical manual, p. 26).
Were appropriate samples used in validation?	3	Standardization sample was appropriate and is described in the technical manual (pp. 63–66).
Were appropriate samples used in norming?	3	Standardization sample was appropriate and is described in the technical manual (pp. 63–66). More than 275,000 students participated in the standardization sample.
If normative date is provided, was the norm sample collected within the last five years?	3	According to the technical manual, TerraNova national standardization occurred in the spring and fall of 1996 (p. 61). According to the technical quality report, standardization samples were revised in 1999 and 2000, and further documentation from the vendor indicates that the norms were updated again in 2005.
Are the procedures associated with the gathering of the normative data sufficiently well described so that the procedures can be properly evaluated?	3	National standardization study detailed in technical manual (pp. 61–90).
Is there sufficient variation in assessment scores?	3	TerraNova assessment scores from the national sample reflect score variability across and within grades.

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C19
CTB/McGraw-Hill's TerraNova: administration of the assessment

Criterion	Score[a]	Comments
Are the administration procedures clear and easy to understand?	3	The teachers guide details the test administration procedure in an understandable manner (pp. 3–4)
Do the administration procedures replicate the conditions under which the assessment was validated and normed?	3	Standardization sample was drawn from users, thus the conditions of assessment use and standardized sample use are comparable.
Are the administration procedures standardized?	3	Procedures are standardized and detailed in the teachers guide. (pp. 3–4)

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

TABLE C20
CTB/McGraw-Hill's TerraNova: reporting

Criterion	Score[a]	Comments
Are materials and resources available to aid in interpreting assessment results?	3	An "Information system" was developed to "ensure optimal application of the precise data provided by TerraNova" (Technical quality report, p. 29).

a. 3 is yes, true; 2 is somewhat true; 1 is not true; and na is not applicable.

NOTES

1. This example comes from Enhancing Education Through Technology site visit reports (Metiri Group, 2007).

2. The Study Island assessments reviewed for this report are those associated with their state test preparation software. The authors did not review the assessments recently developed for Pennsylvania and other states that Study Island specifically refers to as "Benchmark Assessments."

3. In addition, the technical manual (Northwest Evaluation Association, 2003) provides concurrent validity evidence for the tests through correlation analysis with a number of criterion outcome measures. These include: Arizona Instrument to Measure Standards, Colorado Student Assessment Program, Illinois Standards Achievement Test, Indiana Statewide Testing for Educational Progress-Plus, Iowa Test of Basic Skills, Minnesota Comprehensive Assessment and Basic Skills Test, Nevada Criterion Referenced Assessment, Palmetto Achievement Challenge Tests, Stanford Achievement Test, Texas Assessment of Knowledge and Skills, Washington Assessment of Student Learning, and the Wyoming Comprehensive Assessment System.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: Authors.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Joint Committee). (1999). *Standards for educational and psychological testing.* Washington, DC: Authors.

American Psychological Association. (1974). *Standards for educational and psychological tests.* Washington, DC: Author.

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*(2), 234–248.

CTB/McGraw-Hill. (2001a). *TerraNova Technical bulletin* (2nd Ed.). Monterey, CA: Author.

CTB/McGraw-Hill. (2001b). *TerraNova Technical report.* Monterey, CA: Author.

CTB/McGraw-Hill. (2002a). A *linking study between the TerraNova assessment series and Pennsylvania PSSA Assessment for reading and mathematics.* Monterey, CA: Author.

CTB/McGraw-Hill. (2002b). *Teacher's guide to TerraNova* (2nd Ed.). Monterey, CA: Author.

Education Market Research. (2005a, February). *The complete K-12 newsletter: The reading market.* Rockaway Park, NY. Retrieved March 2007 from http://www.ed-market.com/r_c_archives/display_article.php?article_id=76.

Education Market Research. (2005b, June). *The complete K-12 newsletter: The mathematics market.* Rockaway Park, NY. Retrieved March 10, 2007 from http://www.ed-market.com/r_c_archives/display_article.php?article_id=83

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*, 257–288.

McGlinchey, M. T., & Hixson, M.D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*(2), 193–203.

Metiri Group. (2007). *Evaluation of the Pennsylvania Enhancing Education Through Technology Program: 2004–2006*. Culver City, CA.: Author.

Northwest Evaluation Association. (2003). *Technical manual for the NWEA Measures of Academic Progress and Achievement Level Tests.* Lake Oswego, OR: Author.

Northwest Evaluation Association. (2004). *Reliability and validity estimates: NWEA Achievement Level Tests and Measures of Academic Progress.* Lake Oswego, OR: Author.

Northwest Evaluation Association. (2005, September). *RIT scale norms for use with Achievement Level Tests and Measures of Academic Progress.* Lake Oswego, OR: Author.

Northwest Evaluation Association. (2006). *Teacher handbook: Measures of Academic Progress (MAP).* Lake Oswego, OR: Author. Available online at http://www.nwea.org/assets/documentLibrary/Step%201%20-%20MAP%20Administration%20Teacher%20Handbook.pdf

Northwest Evaluation Association. (n.d.). *NWEA Measures of Academic Progress (MAP) content alignment guidelines and processes and item development processes.* Lake Oswego, OR: Author.

Olson, L. (2005, December 30). Benchmark assessments offer regular checkups on student achievement. *Education Week 22*(37): 13–14.

Renaissance Learning. (2000). *STAR Reading: Understanding reliability and validity.* Wisconsin Rapids, WI:

Renaissance Learning, Inc. Retrieved from http://research.renlearn.com/research/pdfs/133.pdf

Renaissance Learning. (2001a). *STAR Math technical manual.* Wisconsin Rapids, WI: Author.

Renaissance Learning. (2001b). *STAR Math: Understanding reliability and validity.* Wisconsin Rapids, WI: Author. Retrieved from http://research.renlearn.com/research/pdfs/135.pdf

Renaissance Learning. (2002). *STAR Reading technical manual.* Wisconsin Rapids, WI: Author.

Renaissance Learning. (2003). *The research foundation for Reading Renaissance goal-setting practices.* Madison, WI: Author. Retrieved from http://research.renlearn.com/research/pdfs/162.pdf

Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation, 4*(2). Retrieved February 14, 2007 from http://PAREonline.net/getvn.asp?v=4&n=2.

Schilling, S. G., Carlisle, J. F., Scott, S. E., & Zeng, J. (2007). Are fluency measures accurate predictors of reading achievement? *The Elementary School Journal, 107,* 429–448.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407–419.

Study Island. (2006a). *Study Island administrator handbook.* Dallas, TX: Author.

Study Island. (2006b). *Evidence of success: Pennsylvania System of School Assessments: Solid research equals solid results research.* Retrieved from http://www.studyisland.com/salesheets

VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review, 30*(3), 363–382.