aimsweb

Technical Manual

ALWAYS LEARNING

PEARSON

Pearson Executive Office 5601 Green Valley Drive Bloomington, MN 55437
800.627.7271 www.PsychCorp.com

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Warning: No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system without permission in writing from the copyright owner.

Pearson, aimsweb, and the **aimsweb logo** are trademarks in the U.S. and/or other countries of Pearson Education, Inc., or its affiliate(s).

Produced in the United States of America.

Product Number 338825

1 2 3 4 5 6 7 8 9 10 11 12 A B C D E

PEARSON

Contents

Introduction 1

Reading Curriculum-Based Measurement (R–CBM) 3

Development of Passages 3

Readability 4

Equivalence of Probe Difficulties 4

Reliability and *SEM* of Probe Scores 7

Validity of Probe Scores 11

Classification Accuracy of Probe Scores 14

Reading Maze 15

Development 15

Reliability and *SEM* of Probe Scores 15

Validity of Probe Scores 16

Mathematics Concepts & Applications (M–CAP) 17

Content Validity 17

Item Development 19

Pilot Studies 19

National Field Test 20

Equivalence of Probes 20

Alternate-Form Reliability and *SEM* of Probe Scores 22

Interrater Reliability 22

Criterion Validity of Probe Scores 23

Classification Accuracy of Probe Scores 24

Reliability of the Rate of Improvement (ROI) 24

Validity of the Rate of Improvement (ROI) 25

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Mathematics Computation (M–COMP) 26

Development 26

Equivalence of Probes 28

Reliability and *SEM of Probe Scores* 29

Validity of Probe Scores 29

Reliability of the Rate of Improvement (ROI) 30

Spelling 31

Reliability of Probe Scores 31

Written Expression 32

Reliability of Probe Scores 32

Validity of Probe Scores 33

Test of Early Literacy (TEL) 35

Rationale 35

Reliability and *SEM* of Probe Scores 35

Validity of Probe Scores 36

Test of Early Numeracy (TEN) 39

Reliability and *SEM* of Probe Scores 39

Validity of Probe Scores 40

References 43

Tables

Table 1 Average Words Read Correctly (WRC) per Minute, by Grade in the R–CBM Field Test 4

Table 2 Correlation of R–CBM Passage Grade Level with Readability Indicators 4

Table 3 Average Words Read Correctly per Minute, and Lexile Measure, of Each R–CBM Probe 5

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 3 Average Words Read Correctly per Minute, and Lexile Measure, of Each R–CBM Probe (*continued)* 6

Table 4 Demographic Characteristics of the R–CBM Alternate-Form Reliability Sample 7

Table 5 Average Alternate-Form Reliability of Single R–CBM Probes, by Grade and Season 8

Table 6 Reliability of the Median of Three R–CBM Probe Scores (Simulation Study) 8

Table 7 Between-Season Stability of R–CBM Screening Scores, by Grade 9

Table 8 Split-Half and Interrater Reliabilities of R–CBM Probe Scores, by Grade 9

Table 9 *SEM* of R–CBM Scores, by Grade 10

Table 10 Criterion Validity of R–CBM Screening Scores, From the **aims**web Database 11

Table 11 Criterion Validity of R–CBM Screening Scores, From Published Studies 12

Table 12 Correlations of R–CBM Screening Scores with Lexile Student-Ability Measures, by Grade 14

Table 13 Classification Accuracy of R–CBM Screening Scores, by Grade 14

Table 14 Between-Season Stability of Maze Scores, by Grade 15

Table 15 Median Correlation of Maze Scores with State-Test Scores, by Grade 16

Table 16 Correlations of Maze Scores with Lexile Student-Ability Measures, by Grade 16

Table 17 NCTM Curriculum Focal Points, by Grade 17

Table 18 Domains Measured by M–CAP, by Grade 18

Table 19 Demographic Characteristics of the M–CAP National Field Test Sample 20

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 20 Average M–CAP Probe Raw Scores in the Field Test 21

Table 21 Average Alternate-Form Reliability of M–CAP Probes, by Grade 22

Table 22 Interrater Reliability of M–CAP Scores, by Grade 22

Table 23 Criterion Validity of M–CAP Scores, From the **aims**web Database 23

Table 24 Classification Accuracy of M–CAP Scores 24

Table 25 Reliability and *SEM* of the ROI of M–CAP Scores, by Grade 24

Table 26 Criterion Validity of the ROI of M–CAP Scores, Controlling for Initial Level 25

Table 27 Content Measured by M–COMP, by Grade 26

Table 28 Demographic Characteristics of the M–COMP Field Test Sample 27

Table 29 Average M–COMP Probe Raw Scores in the Field Test 28

Table 30 Average Alternate-Form Reliability of M–COMP, by Grade 29

Table 31 Criterion Validity of M–COMP Scores With Respect to G–MADE Scores 29

Table 32 Reliability and *SEM* of the ROI of M–COMP Scores, by Grade 30

Table 33 Reliability of Spelling Scores 31

Table 34 Alternate-Form Reliability of Written Expression Scores, by Grade 32

Table 35 Criterion Validity of Written Expression Scores (Jewell & Malecki, 2005) 33

Table 36 Criterion Validity of Written Expression Scores (Gansle et al., 2002) 33

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 37 Criterion Validity of Written Expression Scores (Espin et al., 2000) 34

Table 38 TEL Measures in Relation to the Skills Identified by the National Reading Panel 35

Table 39 Reliability of TEL Scores (Elliott et al., 2001) 35

Table 40 Retest and Alternate-Form Reliability of TEL Scores 36

Table 41 Criterion Validity of TEL Scores 37

Table 42 Criterion Validity of TEL Scores (Elliott et al., 2001) 38

Table 43 Reliability and Stability of TEN Scores (Clarke & Shinn, 2004) 39

Table 44 Reliability and Stability of TEN Scores (Martinez et al., 2009) 39

Table 45 Alternate-Form Reliability of TEN Scores (Baglici et al., 2010) 40

Table 46 Criterion Validity of TEN Scores (Clarke & Shinn, 2004) 40

Table 47 Criterion Validity of TEN Scores (Clarke et al., 2008) 41

Table 48 Criterion Validity of TEN Scores, With Respect to SAT10 (Martinez et al., 2009) 42

Table 49 Criterion Validity of TEN Scores (Baglici et al., 2010) 42

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Introduction

This manual is a compilation of information regarding the technical properties of the **aims**web measures, including findings obtained during development as well as the results of studies carried out after publication. In addition, information about non-**aims**web instruments that closely parallel **aims**web measures is cited when appropriate. Note that this is a dynamic document: new evidence will be added as it is collected.

The technical properties that are important for a given test depend on how the test will be used and interpreted. In the case of assessment systems such as **aims**web that employ general outcome measures (GOMs) for universal screening and progress monitoring, highly important technical properties include the following.

- **Equivalence** of probe difficulty, within each grade level. A student's expected score should be the same regardless of which grade-level probe is administered. This feature is important when progress monitoring so that the trend of scores across time will not be unduly influenced by variations in probe difficulty.
- **Reliability** of probe scores. Reliability refers to the consistency, or repeatability, of scores. It may also be thought of as the degree to which scores are free of measurement error. Two types of reliability are particularly relevant to the **aims**web measures:
  - **Alternate-form reliability:** The agreement between scores on alternate forms (probes) administered relatively close together in time. This type of reliability indicates how free the score is from changes due to day-to-day fluctuations and from differences in the specific content of the probes. (Note that if a measure has only one probe, such as Oral Counting, then **test–retest reliability** indicates consistency over time.)
  - **Interrater reliability:** The agreement among scores calculated by independent raters. This type of reliability is especially relevant when scoring requires judgment.

  Internal-consistency reliability (such as split-half or coefficient alpha) is generally not suitable for **aims**web measures because they assess speed as well as accuracy, and most students do not reach the end of the probe within the time limit.
- **Validity** of probe scores. Validity refers to the accuracy of inferences made from scores, such as, "This student is better at math calculation than about 40% of his or her peers," or, "This student has a high probability of passing the state reading test." The inferences most often made from **aims**web scores are based on assumptions about the probes' content and on the relationship of probe scores with external criteria.
  - **Content validity:** The degree to which the test score measures the designated knowledge/skill domain. Evidence of content validity is typically obtained by comparing item content with a curriculum or with how well the tasks address the purported purpose of the measure.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

- **Criterion validity:** The relationship between test scores and a criterion, such as success in an educational program or scores on other tests. Two subtypes of criterion validity are often differentiated according to the amount of time between the **aims**web administration and the occurrence of the criterion:
    - **Concurrent validity** is the correlation of probe scores with criterion data (e.g., another test) collected at the same time.
    - **Predictive validity** is the correlation of probe scores with a future criterion.
- **Classification accuracy:** An alternative means of expressing criterion validity that is appropriate when there is interest in predicting a dichotomous criterion (e.g., passing or not passing a state test). A cut score on the predictor test (i.e., the **aims**web measure) is chosen such that those who score at or above the cut score are considered likely to pass the criterion, while those who score below the cut score are likely to fail. A classification-accuracy analysis indicates how frequently these expectations prove correct, and the results are reported in a variety of statistics, including:
    - **Sensitivity,** which refers to the proportion of those students who actually fail the criterion who score below the **aims**web cut score.
    - **Specificity,** which refers to the proportion of those students who actually pass the criterion who score above the **aims**web cut score.
    - **Area under the curve,** a global indicator of the degree to which the **aims**web measure correctly predicts the outcome on the criterion.

In addition to these technical characteristics of scores obtained at a single point in time, there are technical properties related to a student's rate of improvement (ROI, or slope) during progress monitoring. The **standard error of measurement (*SEM*) of the ROI** is important to consider when making a decision about whether a student's progress is sufficient to reach the established goal. This *SEM* is a function of the number and variability of the student's scores and their time span. For example, scores that closely follow a straight line across a wide span of time produce a small ROI *SEM*. The average size of the ROI *SEM* relative to the total amount of variability of the ROIs in a population of students indicates the **reliability of the ROI.** Furthermore, the **criterion validity of the ROI** can be evaluated by investigating whether students who show a faster rate of progress on an **aims**web measure are likely to score higher on a criterion (after controlling for the initial score level).

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Reading Curriculum-Based Measurement (R–CBM)

Development of Passages

The brief stories that make up the Standard Reading Assessment Passages (RAPs) (Howe & Shinn, 2002) for R–CBM and Reading Maze were written by educators with a close knowledge of the kinds of writing typically encountered by students at different grade levels. These nine teachers and seven paraprofessionals received onsite training, which included sample passages and the following specific guidelines:

- Passage length, based on word count (250 words for Grades 1 and 2, 300 words for Grade 3, and 350 words for Grades 4 through 8)
- Grade-appropriate number of syllables and sentences per 100 words, based on the Fry readability formula

aimsweb staff reviewed each passage and either returned it to the author (with edits) for revision or rejected it due to inappropriate subject matter. This process yielded a total of 33 passages at Grade 1 and 50 passages at each of Grades 2 through 8.

These passages were field-tested in February and March 2001 with 24 students per grade from Grades 1 through 8 at a suburban/rural school district in the Midwest. The students represented a range of ability levels, according to their performance on standardized reading tests (one-third each at the 75th, 50th, and 25th percentiles). Each student read all of the passages at his or her grade level, spread across five sessions of 10 passages each (six to seven passages per session for Grade 1). The student read each passage aloud for 1 minute, and the examiner recorded the number of words read correctly (WRC) and the number of errors.

Passages were evaluated according to difficulty (average WRC) and alternate-form reliability (average correlation with the other passages at that grade level). In addition, the readability of each passage was assessed by calculating its Lexile measure (Stenner, 1997). Generally, passages whose average WRC differed from the overall average WRC by more than 1 *SEM*, whose Lexile measure was outside the grade range, or whose correlation with other passages was below .70 were rejected. The pool of retained passages included 23 at Grade 1 and 33 at each of Grades 2 through 7.

The Grade 8 passages were not retained in their initial form because the means and standard deviations (*SD*) of their WRCs were lower than those for the Grade 7 passages and because most of them had average alternate-form correlations below .70. These passages were revised and subjected to a second field test involving 183 students randomly drawn from 14 schools in five states. Each participating student read 10 randomly-selected passages; each passage was administered to an average of 36 students (range = 19 to 55 students). The data from this second field test indicated that even though average WRCs were still lower than at Grade 7, the revised passages had good alternate-form reliability; therefore, 33 passages were retained.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

At this point, three passages at each grade level were selected for use in universal screening (benchmarking) on the basis of having WRC means and *SD*s that were close to each grade's average, high correlations with other passages, and appropriate Lexile measures. The remaining passages were designated for use in progress monitoring.

Table 1 reports the average mean and *SD* of WRC scores at each grade in the field-test samples (23 probes at Grade 1, and 33 probes at each of Grades 2 through 8). Note that there is a developmental trend, with average WRCs increasing from Grade 1 to Grade 7, before dropping at Grade 8.

Table 1

Average Words Read Correctly (WRC) per Minute, by Grade in the R–CBM Field Test

WRC	Grade level							
	1	2	3	4	5	6	7	8
Mean	35.7	89.8	107.6	121.5	132.1	141.8	154.2	137.3
SD	19.2	23.6	28.1	25.3	29.1	25.1	25.0	42.0

Readability

The passages were evaluated using a variety of readability formulas in addition to Lexiles. Table 2 displays the correlations between grade placement of passages (1 through 8) and their readability levels. These correlations are very high (averaging .95 for screening passages and .88 for progress-monitoring passages), indicating that passage readability progresses in a consistent way from grade to grade.

Table 2

Correlation of R–CBM Passage Grade Level with Readability Indicators

Readability indicator	Passage grade level	
	Screening probes	Progress monitoring probes
Lexile	.97	.90
Fry	.94	.91
Flesch	.96	.88
Powers	.92	.89
Spache	.97	.81
SMOG	.93	.84
Median	.95	.88

Equivalence of Probe Difficulties

An important consideration when creating the **aims**web R–CBM materials was ensuring that all the probes at a given grade level were equivalent and produced similar results. Table 3 presents the WRC field-test mean and *SD* and the Lexile measure for each passage. Note that the probe numbers listed are those used in the published version of **aims**web. The passages in the table were sorted by mean WRC to show the range of difficulty within each grade's passages and to demonstrate that the probes selected for universal screening (Probes 1 through 3, bolded) are of average difficulty for their grade.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 3

Average Words Read Correctly per Minute, and Lexile Measure, of Each R–CBM Probe

Grade 1				Grade 2				Grade 3				Grade 4			
Probe	Mean	*SD*	Lexile	Probe	Mean	*SD*	Lexile	Probe	Mean	*SD*	Lexile	Probe	Mean	*SD*	Lexile
4	42.1	22.3	360	31	99.9	24.2	220	17	113.0	27.2	540	29	129.5	23.1	400
19	40.6	20.3	190	7	99.6	25.5	370	19	112.8	28.2	610	8	129.4	26.6	750
12	40.5	23.5	360	9	99.0	22.8	340	28	112.4	24.0	470	6	129.0	32.2	610
16	39.6	18.1	300	18	98.9	20.9	470	26	112.0	29.7	620	32	128.6	27.4	690
22	38.6	22.2	250	20	98.0	26.8	400	6	112.0	29.8	500	11	128.4	29.4	560
23	38.5	21.7	400	4	97.9	24.3	340	7	111.8	27.6	520	30	127.8	27.3	760
10	37.9	19.7	190	30	97.1	23.9	300	27	110.5	28.2	710	26	127.0	20.8	560
14	37.5	19.2	200	27	95.4	21.8	380	9	110.4	26.3	590	14	126.8	18.8	680
1	37.0	18.6	240	11	94.6	21.9	440	31	110.2	23.2	640	22	125.4	25.9	600
18	36.8	18.1	260	8	94.5	21.9	510	32	109.7	27.4	600	20	125.1	29.8	630
17	36.3	17.7	290	28	93.6	24.3	310	20	109.6	23.4	640	33	125.1	29.9	600
3	35.9	18.8	250	15	93.0	23.9	490	3	109.1	28.2	570	19	124.9	24.4	700
2	35.8	18.7	210	10	91.7	16.9	520	14	109.0	31.0	430	17	123.2	28.4	610
6	35.0	14.5	390	24	91.5	23.2	450	12	108.7	31.4	660	10	123.0	21.9	490
11	35.0	22.7	220	23	90.5	29.3	460	2	107.2	31.7	460	3	122.8	24.5	670
20	34.1	15.2	300	29	90.3	22.1	260	1	107.1	29.4	630	21	122.8	29.5	480
15	34.0	20.8	280	1	89.6	20.8	420	23	106.8	24.1	690	24	122.5	19.9	620
8	32.7	17.2	310	14	89.6	30.3	300	13	106.8	28.2	590	2	121.8	27.2	650
9	32.2	20.3	260	3	89.2	23.6	470	33	106.1	30.8	440	1	121.5	20.1	770
21	30.9	15.9	410	2	89.1	19.7	440	8	106.0	26.7	570	23	121.0	25.4	670
13	30.5	19.4	410	33	87.5	24.3	340	15	105.7	32.0	520	4	119.6	28.7	500
5	30.1	16.6	370	12	85.8	21.8	540	11	105.5	29.9	570	28	119.2	24.3	610
7	29.0	19.5	200	13	85.2	23.5	450	25	105.5	28.7	520	12	118.3	22.9	530
				6	84.3	24.5	500	16	105.4	25.1	600	13	117.7	18.9	550
				16	83.7	24.9	320	30	105.4	29.7	630	5	117.4	25.2	580
				19	83.3	24.8	420	10	105.1	30.9	580	7	117.2	22.6	720
				26	82.6	24.8	290	24	105.0	33.6	600	31	116.3	26.5	720
				22	82.5	24.7	490	21	104.6	30.0	580	9	116.2	22.1	580
				32	82.0	22.2	500	18	104.5	27.4	500	25	115.3	20.4	690
				21	81.7	29.8	400	5	103.8	27.7	520	15	115.0	22.7	580
				5	81.6	19.9	530	22	103.8	26.0	620	27	112.1	28.1	590
				17	80.3	23.9	440	29	103.1	23.2	660	18	111.6	29.6	620
				25	80.0	22.6	370	4	102.9	27.6	640	16	109.1	28.8	750
Mean	35.7	19.2	289		89.8	23.6	408		107.6	28.1	576		121.5	25.3	622

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 3

Average Words Read Correctly per Minute, and Lexile Measure, of Each R–CBM Probe (*continued*)

Grade 5				Grade 6				Grade 7				Grade 8			
Probe	Mean	*SD*	Lexile	Probe	Mean	*SD*	Lexile	Probe	Mean	*SD*	Lexile	Probe	Mean	*SD*	Lexile
9	141.7	29.0	790	10	153.6	30.3	700	30	165.0	24.5	650	27	161.9	44.2	880
22	141.5	30.1	780	29	152.9	25.6	560	31	163.4	23.9	880	25	160.5	49.1	820
11	141.5	32.7	710	9	151.4	22.0	610	8	163.3	26.8	990	6	158.3	42.0	910
13	141.3	35.4	560	4	150.4	24.3	570	20	163.3	20.1	910	8	157.7	43.8	1090
27	140.2	28.7	690	19	148.9	28.4	940	16	162.7	23.9	900	10	156.4	36.8	800
14	140.0	28.7	1010	31	147.3	24.7	710	28	162.6	30.7	830	16	156.2	40.7	990
10	138.3	31.6	650	25	146.8	23.6	520	19	162.6	26.8	810	24	156.0	48.2	930
33	137.4	29.6	570	18	146.6	23.7	810	13	162.0	29.5	780	29	155.6	47.3	900
28	137.1	30.1	720	26	146.0	28.9	810	32	161.3	20.9	870	14	152.6	44.3	860
3	136.3	29.3	770	17	145.7	28.7	880	25	159.9	24.5	860	21	150.5	42.8	1100
26	135.5	27.7	820	32	145.4	22.9	590	7	159.7	26.7	970	1	149.5	42.0	1040
23	134.8	28.3	440	23	144.0	27.6	730	24	159.3	29.4	950	18	147.8	41.1	940
31	133.8	32.2	610	2	143.8	23.6	840	29	158.0	29.4	720	5	147.5	39.1	850
30	133.7	27.0	650	7	143.4	28.3	1000	15	157.4	25.0	1010	2	146.7	47.3	1070
6	133.6	30.7	580	12	142.7	26.5	920	6	156.9	21.0	950	3	145.5	35.7	1010
18	133.6	24.0	650	6	142.2	22.6	740	26	156.7	22.8	880	31	145.4	33.5	1170
1	133.0	29.1	810	5	141.8	22.2	630	10	154.9	26.4	910	23	145.3	35.6	1130
2	131.8	29.7	780	24	141.7	20.2	840	18	154.7	24.1	760	15	145.3	47.2	920
24	131.5	32.2	630	1	141.1	27.5	1050	1	154.6	27.6	1060	12	145.1	48.6	980
15	130.3	29.2	560	8	141.1	23.1	740	3	154.0	22.0	1000	22	144.6	36.4	1100
21	129.5	34.8	830	14	140.6	26.5	790	2	153.7	21.4	970	9	143.8	40.5	1090
5	128.2	20.8	780	33	140.6	27.8	690	27	153.1	22.1	900	11	143.6	37.7	980
8	128.2	29.9	930	15	140.1	17.1	1030	33	151.8	20.7	910	13	143.6	48.1	1000
12	128.0	29.0	870	30	139.8	28.2	810	5	151.1	29.2	730	26	143.4	42.4	1010
7	126.7	29.8	740	11	139.0	29.4	780	22	149.2	23.5	870	32	143.1	49.5	1140
19	125.8	27.2	760	3	138.0	26.5	930	21	149.0	22.4	870	7	141.3	39.4	1030
17	125.6	27.4	790	21	136.7	23.1	720	23	148.6	24.0	840	28	141.0	39.9	990
25	125.0	27.1	610	20	134.4	25.0	850	4	145.9	28.1	670	33	140.3	41.5	970
4	123.8	28.1	660	27	134.0	25.7	740	17	143.3	20.9	860	30	140.2	49.5	1020
32	123.7	32.2		22	132.8	19.5	680	11	142.5	23.9	990	19	139.8	40.6	1090
16	123.0	24.1	750	28	132.1	24.4	830	14	140.1	26.4	810	17	138.4	40.6	1170
29	122.8	27.2	660	16	131.5	24.9	770	12	136.5	25.6	830	20	137.7	39.4	1080
20	121.6	27.6	880	13	124.6	24.9	880	9	134.2	25.8	780	4	137.5	34.1	960
Mean	132.1	29.1	720		141.8	25.1	778		154.2	25.0	870		147.3	42.1	1001

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Reliability and *SEM* of Probe Scores

Alternate-Form Reliability

The R–CBM screening scores of 1,000 students per grade were randomly drawn from the population of students whose **aims**web R–CBM scores were recorded using browser-based scoring. All students in the sample participated in screening at all three periods of the 2009–2010 school year (fall, winter, and spring). Furthermore, scores were available for each of the three probes included in each screening administration.

To estimate the demographic characteristics of the reliability sample, the demographic makeup of the participating schools (obtained from the National Center for Educational Statistics, U. S. Department of Education) was weighted by the number of students from each school. As shown in Table 4, the sample closely matches the U.S. population by sex and socioeconomic status (free/reduced lunch), has somewhat fewer African American and Hispanic students than the general population, and includes an overrepresentation of schools from the Midwest.

Table 4

Demographic Characteristics of the R–CBM Alternate-Form Reliability Sample

Grade	*N*	Sex (%)		Race/ethnicity (%)					Free/ reduced lunch (%)	Region (%)			
		Female	Male	Asian	African American	Hisp.	White	Other		NE	MW	South	West
1	1,000	49	51	5	12	14	68	1	47	15	42	28	16
2	1,000	50	50	4	10	12	74	0	38	5	59	24	13
3	1,000	49	51	4	11	11	72	3	40	7	58	23	12
4	1,000	49	51	5	10	15	68	3	45	10	40	33	18
5	1,000	49	51	3	10	11	75	1	41	14	43	33	11
6	1,000	50	50	2	6	7	76	9	44	7	74	19	0
7	1,000	50	50	3	14	4	76	2	40	0	76	24	0
8	1,000	50	50	3	14	4	73	6	36	0	77	24	0
U.S. Population		49	51	5	16	22	54	2	44	16	22	38	24

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

At each screening period, three interprobe correlations were calculated for each grade (Probe 1 vs. Probe 2, Probe 1 vs. Probe 3, and Probe 2 vs. Probe 3). These three alternate-form correlations were averaged, and the results are shown in Table 5. These single-probe alternate-form reliabilities are remarkably consistent across grade levels, with an average value of .94 at each period.

Table 5

Average Alternate-Form Reliability of Single R–CBM Probes, by Grade and Season

Grade	*N*	Single probe		
		Fall	Winter	Spring
1	1,000	.95	.95	.94
2	1,000	.94	.94	.94
3	1,000	.94	.93	.94
4	1,000	.95	.94	.94
5	1,000	.94	.94	.94
6	1,000	.94	.93	.94
7	1,000	.94	.93	.95
8	1,000	.94	.93	.93
Mean		.94	.94	.94

A screening period score is the median score on the three probes administered at that period. In order to estimate the reliability of this median, a simulation study was conducted in which two sets of three probe scores were randomly generated for each of five simulated samples of 1,000 cases each. The probe scores were generated in such a way as to have a target intercorrelation of .94, matching the actual average intercorrelation shown in Table 6. The resulting average reliability of the median score (.972) is a best estimate of the alternate-form reliability of R–CBM screening scores (i.e., median of three probes).

Table 6

Reliability of the Median of Three R–CBM Probe Scores (Simulation Study)

Simulation	Average intercorrelation		Reliability of the median (actual)
	Probes 1–3	Probes 4–6	
1	.941	.944	.976
2	.938	.938	.970
3	.942	.944	.973
4	.942	.941	.973
5	.937	.933	.968
Average	.940		.972

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Consistency Across Screening Periods (Long-Term Test–Retest Reliability)

Note that the correlation between R–CBM screening scores obtained four months apart (September to January or January to May) is not a direct measure of reliability because true change as well as measurement error affects the correlation. However, these correlations are valuable as lower bounds of reliability. Using the same national sample as in the alternate-form study, correlations were computed between **aims**web R–CBM screening scores (i.e., median of three probe scores) obtained at adjacent screening periods. Results, presented in Table 7, indicate that R–CBM screening scores have long-term test–retest reliabilities that are, at a minimum, in the mid .90s.

Table 7

Between-Season Stability of R–CBM Screening Scores, by Grade

Grade	*N*	Fall–Winter	Winter–Spring
1	1,000		.91
2	1,000	.93	.94
3	1,000	.93	.94
4	1,000	.94	.95
5	1,000	.95	.95
6	1,000	.95	.95
7	1,000	.95	.95
8	1,000	.95	.96
Mean		.94	.94

Split-Half Reliability

In order to evaluate the consistency of performance across time, a special study was conducted in which scores on the first 30 seconds of the 1-minute R–CBM administration were correlated with scores on the second 30 seconds. Note that this is similar to an alternate-form reliability study. Correlations between the half-test scores were adjusted using the Spearman-Brown formula to estimate the reliability of scores for 1-minute time periods.

The study sample was obtained in March and April 2011 at Grades 2, 4, 6, and 8 from five public schools in Minnesota and Texas. Each administration was audio recorded, with the first-half and second-half scores obtained by listening to the recordings. Results, shown in Table 8, were consistent with the alternate-form reliabilities (see Table 5) and demonstrated a reliability of approximately .94 for individual probes.

Table 8

Split-Half and Interrater Reliabilities of R–CBM Probe Scores, by Grade

Grade	*N*	Split-half reliability	Interrater reliability
2	61	.94	.99
4	63	.95	.99
6	71	.93	.99
8	63	.95	.99

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Interrater Reliability

The audio recordings that were collected for the split-half reliability study were also used for an interrater reliability study. At each grade, half of the recordings were scored by two scorers working independently, while the other half were scored by a different pair of scorers working independently. Interrater reliability coefficients were calculated using Shrout and Fleiss (1979) Formula 2,1, which takes into account differences in the level of scores assigned by different raters as well as differences in how they rank-order students. As shown in Table 8, agreement between raters was extremely high.

SEM of Probe Scores

Table 9 presents *SEM*s for single-probe scores and screening scores (i.e., median score on three probes), based on the average of the within-grade alternate-form reliabilities reported in Table 5 and the average *SD* of the number of WRCs per minute in that reliability study sample.

Table 9

SEM of R–CBM Scores, by Grade

Grade	*N*	Single probe		Screening (median of 3)	
		Reliability	*SEM*	Reliability	*SEM*
1	1,000	.95	7.6	.97	6.1
2	1,000	.94	9.0	.97	6.4
3	1,000	.94	9.9	.97	7.2
4	1,000	.94	9.3	.97	6.6
5	1,000	.94	10.3	.97	7.3
6	1,000	.94	10.5	.97	7.6
7	1,000	.94	10.3	.97	7.8
8	1,000	.93	9.4	.96	6.9

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Validity of Probe Scores

Criterion Validity

Table 10 reports correlations of R–CBM screening scores (fall, winter, and spring) with scores on the North Carolina and Illinois reading tests administered at the end of the same school year (note that all data are from the 2009–2010 school year). The correlations were adjusted for range restriction, using the national norm sample as the reference group. These analyses indicate that R–CBM scores correlate approximately .7 with state reading tests in Grades 3 through 5 and in the mid-to-low .60s in Grades 6 through 8.

Table 10

Criterion Validity of R–CBM Screening Scores, From the **aims**web Database

Criterion	Grade (season)	*N*	Correlation	
			Unadjusted	Adjusted
North Carolina End of Grade Test	3 (fall)	1,087	.67	.69
	3 (winter)	1,087	.70	.71
	3 (spring)	1,087	.71	.72
	4 (fall)	1,174	.65	.70
	4 (winter)	1,174	.67	.71
	4 (spring)	1,174	.68	.72
	5 (fall)	1,088	.66	.68
	5 (winter)	1,088	.66	.67
	5 (spring)	1,088	.67	.69
Illinois Standards Achievement Test	6 (fall)	1,326	.64	.64
	6 (winter)	1,326	.66	.65
	6 (spring)	1,326	.65	.64
	7 (fall)	1,328	.65	.63
	7 (winter)	1,328	.65	.63
	7 (spring)	1,328	.64	.62
	8 (fall)	911	.62	.60
	8 (winter)	911	.62	.60
	8 (spring)	911	.62	.60

Table 11 presents the results from **aims**web criterion-validity studies that have been reported in the literature. Note that the interval between R–CBM administration and the criterion was as great as two-and-a-half years in several of the analyses. Details of these studies include:

- Keller-Margulis, Shapiro, and Hintze (2008) conducted a longitudinal study of about 200 elementary-school students in a Pennsylvania school district. R–CBM scores at Grades 1 through 4 were correlated with scores from standardized group reading tests administered one to two-and-a-half years later.
- Shapiro, Keller, Lutz, Santoro, and Hintze (2006) collected data at Grades 3 through 5 from an urban/suburban school district in eastern Pennsylvania, reporting correlations of R–CBM scores (fall, winter, and spring) with scores on group reading tests administered in the spring of the same grade.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

- Merino and Beckman (2010), using a sample of 233 students in Grades 2 through 4 at a Nebraska elementary school, evaluated the correlation of R–CBM screening scores in the spring with scores on the Measures of Academic Progress (MAP) the following fall.
- Silberglitt and Hintze (2005) conducted a longitudinal study of approximately 2,000 students in five suburban/rural school districts in Minnesota, calculating correlations of R–CBM scores at Grades 1 through 3 with scores on the state reading test administered in the spring of Grade 3.
- Andren (2010) studied the correlation of R–CBM scores from administrations conducted in the fall and winter of Grade 3 with scores from group reading tests given in the fall, winter, and spring of the same grade. The sample consisted of 137 students from a suburban school district in the northeastern United States.

Table 11

Criterion Validity of R–CBM Screening Scores, From Published Studies

Criterion	R–CBM grade (season)	Criterion grade (season)	*N*	*r*	Source
Pennsylvania System of School Assessment	1 (fall)	3 (spring)	~ 200	.53	Keller-Margulis et al. (2008)
	1 (winter)	3 (spring)	~ 200	.60	
	1 (spring)	3 (spring)	~ 200	.69	
	2 (fall)	3 (spring)	~ 200	.69	
	2 (winter)	3 (spring)	~ 200	.71	
	2 (spring)	3 (spring)	~ 200	.71	
TerraNova	2 (fall)	4 (spring)	~ 200	.62	
	2 (winter)	4 (spring)	~ 200	.63	
	2 (spring)	4 (spring)	~ 200	.69	
	3 (fall)	5 (spring)	~ 200	.74	
	3 (winter)	5 (spring)	~ 200	.74	
	3 (spring)	5 (spring)	~ 200	.72	
	4 (fall)	5 (spring)	~ 200	.67	
	4 (winter)	5 (spring)	~ 200	.69	
	4 (spring)	5 (spring)	~ 200	.69	
Pennsylvania System of School Assessment	3 (fall)	3 (spring)	185	.65	Shapiro et al. (2006)
	3 (winter)	3 (spring)	185	.66	
	3 (spring)	3 (spring)	185	.67	
Metropolitan Achievement Test 8: Total Reading	4 (fall)	4 (spring)	213	.72	
	4 (winter)	4 (spring)	213	.71	
	4 (spring)	4 (spring)	213	.70	
Pennsylvania System of School Assessment	5 (fall)	5 (spring)	206	.68	
	5 (winter)	5 (spring)	206	.69	
	5 (spring)	5 (spring)	206	.67	

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 11

Criterion Validity of R–CBM Screening Scores, From Published Studies *(continued)*

Criterion	R–CBM grade (season)	Criterion grade (season)	*N*	*r*	Source
Measures of Academic Progress	2 (spring)	3 (fall)	85	.68	Merino & Beckman (2010)
	3 (spring)	4 (fall)	71	.72	
	4 (spring)	5 (fall)	77	.70	
Minnesota Comprehensive Assessment	1 (winter)	3 (spring)	1,475	.47	Silberglitt & Hintze (2005)
	1 (spring)	3 (spring)	1,549	.57	
	2 (fall)	3 (spring)	1,864	.60	
	2 (winter)	3 (spring)	1,893	.66	
	2 (spring)	3 (spring)	1,935	.67	
	3 (fall)	3 (spring)	2,051	.68	
	3 (winter)	3 (spring)	2,120	.70	
	3 (spring)	3 (spring)	2,126	.71	
Measures of Academic Progress	3 (fall)	3 (fall)	137	.81	Andren (2010)
	3 (fall)	3 (winter)	137	.76	
New England Common Assessment Program	3 (fall)	3 (spring)	137	.71	
Measures of Academic Progress	3 (winter)	3 (winter)	137	.77	
	3 (winter)	3 (spring)	137	.68	

All of these studies yielded similar findings. When the interval between R–CBM and the criterion was one year or less, correlations were approximately .70. For two-year intervals, the correlations were lower, generally in the .50s or low .60s.

During the May 2011 **aims**web screening period, the R–CBM measure was aligned with the Lexile scale through a linking study (MetaMetrics, 2011). A total of 5,444 students from school districts in several states took R–CBM and a specially-constructed Lexile calibration test. The test, a group-administered, 40-minute, 35-item assessment of reading comprehension, asked students to read a brief passage and then select (from four options) the word that best completed a sentence related to the content of the passage. The internal-consistency (alpha) reliability of the Lexile test at each grade ranged from .90 to .92. Table 12 reports correlations of R–CBM with the Lexile test, which ranged from .59 to .73 (median = .66).

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 12

Correlations of R–CBM Screening Scores with Lexile Student-Ability Measures, by Grade

Grade	*N*	R–CBM raw score		Lexile measure		*r*
		Mean	*SD*	Mean	*SD*	
1	724	64.9	39.6	140	287	.64
2	760	98.5	43.1	413	253	.72
3	659	108.7	41.5	562	248	.73
4	699	133.0	41.6	794	287	.71
5	712	143.0	43.7	887	253	.68
6	804	159.9	41.5	1016	295	.65
7	608	153.0	38.7	1012	298	.59
8	478	150.0	37.3	1094	267	.65

Classification Accuracy of Probe Scores

Table 13 summarizes the results of predictive classification-accuracy studies in which the criterion was passing the end-of-year state reading assessment. The samples included those reported in Table 10 (correlations with North Carolina and Illinois reading tests) and those in the Keller-Margulis et al. (2008) study (see Table 11). The high values obtained for area under the curve and sensitivity/specificity indicate that R–CBM screening scores are reasonably strong predictors of state test success.

Table 13

Classification Accuracy of R–CBM Screening Scores, by Grade

R–CBM grade (season)	Criterion grade	*N*	Area under the curve	Sensitivity	Specificity
1 (spring)	3	~ 200	.88	.72	.90
2 (spring)	4	~ 200	.94	.79	.91
3 (fall)	3	1,105	.85	.77	.76
3 (winter)	3	1,248	.86	.77	.75
3 (spring)	5	~ 200	.88	.77	.81
4 (fall)	4	1,193	.83	.78	.74
4 (winter)	4	1,332	.86	.78	.77
5 (fall)	5	1,105	.83	.75	.74
5 (winter)	5	1,171	.84	.79	.73
6 (fall)	6	1,393	.83	.77	.74
6 (winter)	6	1,599	.83	.78	.73
7 (fall)	7	1,444	.84	.76	.75
7 (winter)	7	1,656	.84	.77	.76
8 (fall)	8	1,276	.86	.80	.79
8 (winter)	8	1,223	.84	.79	.74

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Reading Maze

Development

aimsweb Reading Maze uses the same passages as those used for R–CBM. (Note that the R–CBM Development section of this manual describes the procedures and analyses supporting the grade-appropriateness and equivalence of these passages.) These passages were adapted for Maze in a manner similar to that recommended by Fuchs and Fuchs (1992): Starting with the second sentence of a given passage, every seventh word was replaced by a set of three words (in brackets). Of these three words, one was the original word, another was a word from the passage that was of the same part of speech but did not make sense or preserve meaning, and the third was a word from the passage that was of a different part of speech and did not make sense in context.

Reliability and *SEM* of Probe Scores

Screening scores in the **aims**web database from the 2009–2010 school year were analyzed to determine the alternate-form stability of Reading Maze raw scores over a four-month interval (fall to winter, Table 14). Because of the long time period between administrations, these are conservative estimates of Maze alternate-form reliability.

Table 14

Between-Season Stability of Maze Scores, by Grade

Grade	*N*	Reliability	*SEM*
2	13,420	.68	3.2
3	21,747	.70	3.8
4	24,881	.74	3.8
5	25,418	.78	3.9
6	11,690	.78	4.5
7	7,201	.76	4.5
8	6,095	.75	4.8

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Validity of Probe Scores

Maze scores from universal screening in Grades 3 through 8 were correlated with scores on state end-of-year reading tests based on 2009–2010 data in the **aims**web database. Table 15 reports the median of the within-state correlations at each grade, using winter screening scores. (Correlations based on fall and spring screening scores were similar.)

Table 15

Median Correlation of Maze Scores with State-Test Scores, by Grade

Grade	Total *N*	Number of states	Median correlation
3	9,625	11	.59
4	8.41	10	.59
5	6,913	9	.58
6	4,984	4	.57
7	4,607	5	.51
8	2,876	3	.55

Marcotte and Hintze (2009) reported a correlation of .67 between Maze scores and the total score on the *Group Reading Assessment and Diagnostic Evaluation* (GRADE), administered concurrently, for 111 Grade 4 students from two schools in western Massachusetts. In the same sample, Maze correlated .72 with **aims**web R–CBM.

Furthermore, during the May 2011 **aims**web screening period, the Maze measure was aligned with the Lexile scale via a linking study (MetaMetrics, 2011). A total of 5,316 students from school districts in several states took Maze and a specially-constructed Lexile calibration test—a group-administered, 40-minute, 35-item assessment of reading comprehension in which the student read a brief passage and then selected (from four options) the word that best completed a sentence related to the content of the passage. The internal-consistency (alpha) reliability of the Lexile test at each grade ranged from .90 to .92. Table 16 reports correlations of Maze with the Lexile test, which ranged from .47 to .63 (median = .57) and were consistent in level across the grade range.

Table 16

Correlations of Maze Scores with Lexile Student-Ability Measures, by Grade

Grade	*N*	MAZE raw score		Lexile measure		*r*
		Mean	*SD*	Mean	*SD*	
1	693	11.3	8.3	153	284	.57
2	751	19.1	10.4	415	252	.47
3	693	17.4	9.2	462	246	.58
4	693	20.3	8.8	795	287	.58
5	687	22.5	8.2	887	254	.54
6	827	26.4	10.1	1012	295	.62
7	512	26.6	9.9	983	287	.55
8	460	23.7	8.6	1097	269	.63

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Mathematics Concepts & Applications (M–CAP)

Content Validity

Since the late 1980s, the National Council of Teachers of Mathematics (NCTM) has shaped the national perception regarding what is important for students to learn in mathematics by developing standards designed to ensure that students have the ability to use mathematics to solve real-life problems. First published in 1989 and revised in 2000, the NCTM *Principles and Standards for School Mathematics* set the stage for a continuous discussion among educators regarding *what* should be taught in mathematics and *when* it should be taught. In its 2006 publication, *Curriculum Focal Points for Prekindergarten through Grade 8 Mathematics: A Quest for Coherence,* the NCTM provided further clarity regarding mathematics expectations, as shown in the excerpt in Table 17.

Table 17

NCTM Curriculum Focal Points, by Grade

Grade	M–CAP	Focal point
2	X	Number and Operations: Base Ten, Place Value
		Number and Operations and Algebra; Basic Facts
	X	Measurement: Linear Measurement and Length
3		Number and Operations and Algebra: Understanding of Facts
	X	Number and Operations: Fractions
	X	Geometry: Two-Dimensional Shapes
4		Number and Operations and Algebra: Facts and Fluency with Whole Numbers
	X	Number and Operations: Decimals, Fractions and Decimals
	X	Measurement: Two-Dimensional Shapes
5	X	Number and Operations and Algebra: Fluency with Whole Numbers
	X	Number and Operations: Decimals, Fractions and Decimals
	X	Geometry and Measurement: Three-Dimensional Shapes
6		Number and Operations: Fluency with Multiplication/Division of Fractions and Decimals
	X	Number and Operations: Ratios and Rate
	X	Algebra: Expressions and Equations
7	X	Number and Operations and Algebra and Geometry: Proportionality and Similarity
	X	Measurement and Geometry and Algebra
	X	Number and Operations and Algebra: Rational Numbers and Linear Equations
8	X	Algebra: Linear Equations
	X	Geometry and Measurement: Space, Figures, Angles
	X	Data Analysis and Number and Operations and Algebra: Data Sets

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

According to the National Research Council (NRC) report *Adding It Up* (Mathematics Learning Study Committee & National Research Council, 2001 p. 5), an understanding of number concepts and operations—what the authors term *mathematics proficiencies*—is the most crucial of the domains that make up the K through 8 curriculum. The authors state:

> "Mathematical proficiency, as we see it, has five strands:
>
> - conceptual understanding—comprehension of mathematical concepts, operations, and relations
> - procedural fluency—skill in carrying out procedures flexibly, accurately, efficiently, and appropriately
> - strategic competence—ability to formulate, represent, and solve mathematical problems
> - adaptive reasoning—capacity for logical thought, reflection, explanation, and justification
> - productive disposition—habitual inclination to see mathematics as sensible, useful, and worthwhile, coupled with a belief in diligence and one's own efficacy
>
> The most important observation we make about these five strands is that they are interwoven and interdependent."

The Mathematics Concepts & Applications (M–CAP) probes were designed with this in mind, while also incorporating the mathematical domains identified by the NCTM 2006 focal points. Furthermore, M–CAP was designed to also reflect the NRC's recommendations regarding instruction, with a focus on ensuring students' problem-solving, logical reasoning, and application of analytical skills to problems (Mathematics Learning Study Committee & National Research Council, 2001).

Table 18 describes the domains measured by M–CAP. The content blueprint of the *Stanford Achievement Test, Tenth Edition* (SAT10; Harcourt Educational Measurement, 2002) was used as a guideline for the proportion of items from each domain at each grade level.

Table 18

Domains Measured by M–CAP, by Grade

Domain	Grade						
	2	3	4	5	6	7	8
Number Sense	√	√	√	√	√	√	√
Operations	√	√	√	√	√	√	√
Patterns & Relationships	√	√	√	√	√	√	√
Measurement	√	√	√	√	√	√	√
Geometry	√	√	√	√	√	√	√
Data & Probability	√	√	√	√			
Algebra				√	√	√	√
Probability					√	√	√
Data & Statistics					√	√	√

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Item Development

Experienced item writers with expertise in mathematics curriculum created approximately 11,200 items in accordance with grade-level and domain-specific criteria. After internal review and editing, the items were assembled into an "anchor" probe at each grade, according to the domain blueprint previously referenced. To maximize the amount of data collected from at-risk learners, items were placed in order of increasing estimated difficulty.

A group of mathematics teachers and other content experts from across the United States reviewed the items and anchor probes with regard to the following questions:

- Is the item representative of content in your curriculum?
- Is the item grade-appropriate?
- Is the item at the appropriate difficulty level?
- Is the vocabulary used in the item grade-appropriate?
- Are all of the mathematical symbols in the item developmentally appropriate and consistent with your district's approach?
- Is the art appropriate, and is the item easy to understand?
- Does the item need more specific directions?
- Does the potential for bias exist? (i.e., should the wording be revised?)

The reviewers' feedback was then used to revise the items and probes.

Pilot Studies

Three pilot studies were conducted to evaluate the items and finalize probe design prior to the field test.

Pilot Study 1: The anchor probes were individually administered, with a liberal time limit, to 56 students to evaluate item-completion times and the clarity of general instructions and item wording. Individual items that took a disproportionate amount of time to complete were identified and eliminated. In addition, students' questions provided insight into how the administration directions were functioning.

Pilot Study 2: Two probes, the anchor and a "cloned" probe, were group-administered to 140 students to determine the administration time required to obtain adequate alternate-form reliability. To this end, students marked their progress at each minute from 6 through 12 minutes. Alternate-form correlations at various time limits indicated that optimal administration time was 8 minutes for Grades 2 through 6 and 10 minutes for Grades 7 and 8.

Pilot Study 3: The anchor probes were group-administered to another 140 students, using no time limit, to collect additional item statistics, including item-total correlations and item difficulties. This information was then used to revise the composition of the anchor probes.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

National Field Test

At each grade, 44 clones of the anchor probe were constructed, consisting of items parallel to the anchor-probe items. Each clone had the same sequence of item types as the anchor probe. All probes were administered to a national field test sample of 6,550 students in the spring of 2009. The demographic characteristics of the sample are shown in Table 19.

Table 19

Demographic Characteristics of the M–CAP National Field Test Sample

		Sex (%)		Race/ethnicity (%)					Median family income (%)			Region (%)			
Grade	***N***	Female	Male	Asian	African American	Hisp.	White	Other	Low	Middle	High	NE	MW	South	West
2	1,064	52	48	2	10	33	53	2	51	24	25	8	29	55	9
3	965	52	48	1	9	22	68	0	40	35	26	7	24	58	10
4	1,026	49	51	1	10	21	65	3	61	11	28	12	36	46	6
5	867	50	50	1	9	22	64	4	65	19	16	1	36	54	10
6	858	48	52	2	5	32	58	3	55	25	19	12	29	50	9
7	912	50	50	8	9	28	53	2	42	19	39	5	30	51	14
8	858	52	48	6	8	23	57	6	33	31	36	5	32	47	16
U.S. Population		49	51	5	16	22	54	2	—	—	—	16	22	38	24

First, the 45 probes at each grade level were grouped into 11 sets of five: the anchor probe (always administered first) and four additional probes (in counterbalanced sequence). Then, each student took one set of five probes. Lastly, each grade's final 33 probes were chosen on the basis of a) having a mean raw score close to the average for all probes and b) having a high mean correlation with the other probes in the set.

Equivalence of Probes

Table 20 reports the M–CAP raw score means and *SD*s obtained during the spring field test. At each grade, the probes are displayed in descending order of the mean score to show the range of values. Note that the mean scores are fairly consistent, with a within-grade *SD* of about two raw score points.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 20

Average M–CAP Probe Raw Scores in the Field Test

	Grade 2		Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
	26.4	10.3	22.5	9.3	30.4	8.7	19.3	11.0	25.5	9.3	25.4	11.0	20.2	9.9
	26.0	10.1	22.2	8.6	29.6	9.2	18.6	9.9	24.2	8.8	24.2	10.2	20.0	10.4
	25.8	11.2	22.1	9.2	29.6	9.5	18.6	11.0	23.3	9.3	24.0	10.5	19.5	9.5
	25.6	9.8	22.1	8.5	29.0	8.7	18.1	9.8	23.1	9.8	23.5	10.4	19.3	8.9
	25.4	10.1	21.9	9.0	28.9	9.0	18.1	10.2	22.8	9.5	23.4	9.8	19.2	8.4
	25.2	10.8	21.7	9.3	28.5	9.5	18.0	10.6	22.8	9.8	23.0	10.0	19.1	9.5
	24.9	10.2	21.5	9.5	28.0	8.0	17.8	9.9	22.6	9.1	23.0	10.8	19.0	9.8
	24.4	9.7	21.5	9.0	27.6	10.3	17.7	9.7	22.5	9.1	22.9	10.5	18.8	8.7
	24.4	10.5	21.4	9.9	27.3	10.5	17.0	9.9	22.4	10.3	22.8	7.2	18.8	9.6
	24.0	9.8	21.4	9.3	27.2	8.1	16.8	9.1	22.3	9.0	22.2	7.8	18.7	10.3
	23.9	10.2	21.3	9.1	27.2	10.2	16.7	9.4	22.1	9.9	21.8	8.6	17.7	9.1
	23.8	9.6	21.2	8.2	27.0	9.0	16.4	9.1	22.0	9.5	21.8	7.5	16.9	9.3
	22.9	11.1	21.2	7.8	26.5	8.8	16.0	7.1	22.0	10.2	21.7	8.7	16.7	9.4
	22.9	9.6	20.7	7.7	26.5	10.6	15.9	8.3	21.7	10.3	21.3	7.6	16.4	8.6
	22.9	10.1	20.7	8.4	26.3	10.4	15.8	8.4	21.3	10.1	21.3	6.6	16.4	8.9
	22.8	11.7	20.0	8.3	25.8	9.1	15.7	8.7	21.2	9.2	20.6	9.2	16.3	9.1
	22.4	9.0	19.9	7.1	25.6	9.2	15.5	7.3	21.1	10.0	20.6	9.6	16.3	8.3
	22.2	9.9	19.8	8.9	25.6	9.8	15.5	8.0	20.8	10.1	20.4	11.0	16.0	8.3
	22.2	9.3	19.0	9.2	25.5	7.9	15.4	7.4	20.2	10.6	19.9	11.7	15.7	8.8
	22.1	10.0	18.8	7.6	25.3	9.7	15.2	8.1	20.0	9.4	19.9	11.3	15.6	8.2
	21.9	9.3	18.6	7.2	24.7	8.7	15.2	7.4	19.9	10.2	19.8	11.6	15.4	7.4
	21.8	11.0	18.6	7.0	24.5	8.5	15.2	9.2	19.3	9.5	19.4	9.5	15.4	8.8
	21.7	9.5	18.5	8.8	24.3	10.3	15.1	8.7	19.3	10.5	19.2	10.3	15.0	7.1
	21.6	10.5	18.5	7.7	24.2	9.2	15.0	8.9	19.2	9.5	19.0	9.9	15.0	7.1
	21.5	9.9	17.9	7.4	23.6	10.0	14.9	7.4	19.1	9.6	18.7	11.0	14.9	8.3
	21.5	10.6	17.4	7.4	23.1	9.2	14.8	6.8	18.7	8.7	18.6	10.3	14.6	7.6
	21.4	9.0	17.2	7.0	23.1	9.5	14.7	7.9	18.7	10.3	18.5	10.0	14.5	7.3
	21.4	10.2	16.8	6.9	22.9	10.4	14.6	7.7	18.5	9.0	18.0	13.4	14.2	7.9
	21.3	10.8	16.3	7.1	22.9	9.5	14.5	8.9	18.2	9.1	17.6	13.9	14.1	8.3
	21.2	9.7	16.1	7.2	22.8	10.5	14.4	7.2	18.1	8.9	16.7	9.3	13.6	6.6
Mean	23.2	10.1	19.9	8.3	26.1	9.4	16.2	8.8	21.1	9.6	21.0	10.0	16.8	8.6
SD	1.7	0.7	1.9	0.9	2.2	0.8	1.4	1.2	1.9	0.5	2.2	1.7	2.0	1.0

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Alternate-Form Reliability and *SEM* of Probe Scores

Table 21 shows the alternate-form reliability (average correlation between probes) based on the field-test sample, as well as the corresponding *SEM*, at each grade level. The median reliability is .86.

Table 21

Average Alternate-Form Reliability of M–CAP Probes, by Grade

Grade	*N*	Reliability	*SEM*
2	1,064	.86	4.6
3	965	.81	4.5
4	1,026	.80	5.1
5	867	.84	4.3
6	858	.86	4.4
7	912	.88	4.2
8	858	.86	4.0

Interrater Reliability

Sixty cases were selected at random from the field test sample and independently scored by two individuals who were provided with the training materials from the M–CAP Administration and Scoring Guide. Using Shrout and Fleiss's (1979) Formula 2,1—which takes into account differences in both level and rank-ordering of cases—interrater reliability is extremely high, as shown in Table 22.

Table 22

Interrater Reliability of M–CAP Scores, by Grade

Grade	*N*	Reliability
2	60	.99
3	60	.99
4	60	.99
5	60	.99
6	60	.99
7	59	.99
8	60	.97

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Criterion Validity of Probe Scores

Table 23 reports correlations of M–CAP screening scores (fall, winter, and spring) with scores on the North Carolina and Illinois end-of-year state mathematics tests, using data from **aims**web users for the 2009–2010 school year. Note that the correlations have been adjusted for range restriction, using the national norm sample as the reference group. These analyses indicated that M–CAP scores correlate at a consistent level (.60s and low .70s) with state math tests across Grades 3 through 8.

Table 23

Criterion Validity of M–CAP Scores, From the **aims**web Database

Criterion	M–CAP grade (season)	*N*	Correlation	
			Unadjusted	Adjusted
North Carolina End of Grade Test	3 (fall)	553	.60	.63
	3 (winter)	733	.64	.67
	3 (spring)	736	.64	.64
Illinois Standards Achievement Test	4 (fall)	700	.60	.67
	4 (winter)	699	.56	.60
	4 (spring)	665	.58	.64
	5 (fall)	752	.60	.57
	5 (winter)	736	.63	.60
	5 (spring)	746	.65	.60
	6 (fall)	631	.78	.76
	6 (winter)	855	.74	.75
	6 (spring)	959	.78	.78
	7 (fall)	723	.74	.61
	7 (winter)	942	.76	.66
	7 (spring)	930	.80	.71
	8 (fall)	640	.74	.69
	8 (winter)	783	.73	.71
	8 (spring)	784	.76	.73

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Classification Accuracy of Probe Scores

Table 24 summarizes the results of predictive classification-accuracy studies in which the criterion was passing the end-of-year state mathematics assessment for North Carolina (Grade 3) or Illinois (Grades 4 through 8). The high values obtained for area under the curve and sensitivity/specificity indicate that M–CAP screening scores are reasonably strong predictors of state test success.

Table 24

Classification Accuracy of M–CAP Scores

M–CAP grade (season)	Criterion grade	*N*	Area under the curve	Sensitivity	Specificity
3 (fall)	3	553	.81	.80	.68
3 (winter)	3	733	.84	.82	.72
4 (fall)	4	700	.84	.77	.75
4 (winter)	4	699	.83	.75	.74
5 (fall)	5	752	.82	.78	.70
5 (winter)	5	736	.83	.76	.75
6 (fall)	6	631	.91	.85	.82
6 (winter)	6	855	.85	.80	.76
7 (fall)	7	723	.88	.82	.78
7 (winter)	7	942	.89	.83	.77
8 (fall)	8	640	.88	.83	.79
8 (winter)	8	783	.87	.83	.76

Reliability of the Rate of Improvement (ROI)

Table 25 reports the reliability and *SEM* of the ROI per week in M–CAP raw scores of students from the 2010–2011 **aims**web database whose progress had been monitored at least 10 times during the school year. The sample's average frequency of monitoring was about once every two weeks. The reliability coefficient reported is a split-half measure based on the correlation between ROIs calculated from odd- and even-numbered probe administrations, adjusted using the Spearman-Brown formula.

Table 25

Reliability and *SEM* of the ROI of M–CAP Scores, by Grade

Grade	*N*	Reliability	*SEM*	Average duration (months)	Average number of administrations
2	6,632	.78	.15	6.9	15.2
3	7,739	.78	.14	7.0	15.5
4	7,553	.79	.16	7.0	15.6
5	7,047	.77	.12	7.1	15.3
6	4,212	.76	.13	7.2	15.4
7	2,906	.75	.14	7.2	15.3
8	3,083	.78	.13	7.0	14.8

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Validity of the Rate of Improvement (ROI)

Table 26 reports the results from two studies of the validity of the M–CAP ROI as a predictor of the Ohio or Texas state mathematics achievement test scores, controlling for the initial M–CAP score level. The samples consisted of students from the 2010–2011 **aims**web database who were progress monitored and who had taken M–CAP at least 10 times during the school year. The partial correlation indicates that, for students with similar fall M–CAP scores, those with higher ROIs tend to have higher math achievement scores on the spring state test.

Table 26

Criterion Validity of the ROI of M–CAP Scores, Controlling for Initial Level

Grade	Criterion	*N*	Partial correlation of ROI with criterion	Average duration (months)	Average number of administrations
3	Ohio Achievement Test	58	.42	5.6	13.2
4	Texas Assessment of Knowledge and Skills	40	.68	5.5	15.4

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Mathematics Computation (M–COMP)

Development

A content blueprint was constructed for Mathematics Computation (M–COMP) on the basis of guidance from internal and external experts in math curriculum (see Table 27). Experienced mathematics item writers created items in accordance with the blueprint, and these items were reviewed and edited by **aims**web staff.

Table 27

Content Measured by M–COMP, by Grade

Domain	Grade							
	1	2	3	4	5	6	7	8
Column Addition	√	√	√					
Basic Facts	√	√	√	√	√	√		
Complex Computation	√	√	√	√	√	√	√	
Decimals				√	√	√	√	√
Fractions				√	√	√	√	√
Conversions					√	√	√	√
Percentages					√	√	√	√
Integers						√	√	√
Expressions						√		
Reducing						√	√	
Equations							√	√
Exponents							√	√

The revised items were sent to external reviewers, who were asked to ensure that:

- no errors existed in the item structure (stem, format, etc.),
- the item could be solved by students at the intended grade level, and
- the keyed answer was correct.

One preliminary anchor probe was then developed for each grade level (1 through 8) to serve as the template for all probes at the corresponding grade level. To enhance the sensitivity of the instrument by maximizing the number of items completed by less-able students, items were arranged in approximate order of estimated difficulty while maintaining item-to-item variation in format and domain.

The preliminary anchor probes underwent a series of pilot studies to evaluate item difficulty and validity, adequacy of scale floors and ceilings, administration times, and scoring criteria.

Pilot Study 1: Sixteen students were tested to ensure that all directions for administration and individual items were clear and grade-level appropriate, that no issues existed with item progression, and that all items were functioning properly. Qualitative and quantitative feedback led to minor adjustments in the anchor probes.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Pilot Study 2: Two clones of the anchor probe were created at each grade. All three probes (anchor plus two clones) were administered to each of 337 students. During each probe administration, students made marks to indicate their progress at each of several time points, so that the effects of administration time on performance could be studied. Alternate-form correlations were calculated for each administration time point to determine the length of time necessary to obtain satisfactory reliability. Based on this information, a time limit of 8 minutes was chosen.

Pilot Study 3: The anchor probes were administered to 444 students in an untimed group administration to obtain more extensive data about item difficulty and validity. Split-half and coefficient alpha reliabilities were calculated. Finally, the anchor probes were modified in light of this information.

Once the final anchor probes had been designed, an additional 44 probes were constructed at each grade level that mirrored the content and placement of items in the anchor probe. A national field test of all 45 probes per grade was conducted in May 2010 with a sample of 7,703 students. Table 28 reports the demographic characteristics of this field test sample.

Table 28

Demographic Characteristics of the M–COMP Field Test Sample

		Sex (%)		Race/ethnicity (%)					Median family income (%)			Region (%)			
Grade	***N***	Female	Male	Asian	African American	Hisp.	White	Other	Low	Middle	High	NE	MW	South	West
1	919	53	47	2	10	26	58	4	50	16	33	22	25	46	7
2	976	56	44	2	9	25	61	3	53	12	35	17	28	49	6
3	971	56	44	2	9	28	60	1	39	28	33	20	18	55	8
4	913	56	44	1	10	28	59	2	36	40	25	21	11	62	6
5	1,048	54	46	1	9	22	65	3	35	41	24	17	14	63	6
6	981	51	49	1	9	16	71	3	24	33	43	10	38	45	7
7	944	53	47	1	4	25	66	4	31	39	30	9	29	46	15
8	948	57	43	2	6	20	68	4	23	38	40	10	38	37	16
U.S. Population		49	51	5	16	22	54	2	—	—	—	16	22	38	24

At each grade level, the 45 probes were grouped into 22 sets, each consisting of the anchor probe (administered first) and four other probes in counterbalanced order. In addition, each five-probe set included either an off-grade-level M–COMP probe (for vertical scaling) or an M–CBM or M–CBM2 probe (for norms equating).

Finally, the field test data were used to select each grade's final set of 33 probes based on each probe's average correlation with other probes and how close its mean raw score was to the average mean.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Equivalence of Probes

Table 29 reports the means and *SD*s of raw scores (obtained during field testing) for the final set of probes at each grade. To show the range of values, the probes are arranged in descending order of the mean score at each grade level. Note that the mean scores are fairly consistent (the *SD* of the means is about two points).

Table 29

Average M–COMP Probe Raw Scores in the Field Test

	Grade 1		Grade 2		Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
	40.1	9.7	41.5	7.8	54.0	12.5	54.7	12.8	37.9	20.9	35.7	16.4	41.9	17.5	37.5	21.0
	39.9	10.5	40.9	8.7	53.7	12.6	54.6	15.1	37.6	17.0	35.5	14.8	41.0	18.7	36.9	16.4
	39.9	11.0	40.8	8.2	53.6	12.1	54.4	16.0	37.4	21.1	35.5	14.5	40.9	17.4	36.8	20.9
	37.9	9.7	40.1	8.4	53.5	12.5	54.1	13.1	36.8	15.3	35.4	19.7	40.6	17.9	36.5	15.5
	37.8	10.1	39.9	8.6	53.4	12.4	53.7	12.9	36.7	15.5	35.2	18.4	40.1	19.9	36.2	19.3
	37.7	10.0	39.8	8.8	53.4	13.6	53.6	10.3	36.4	16.8	35.1	12.7	37.6	17.8	34.7	15.3
	37.7	10.2	39.7	10.0	53.3	11.6	53.5	11.6	35.8	18.1	34.8	18.1	37.5	17.1	33.9	20.7
	37.6	10.2	39.7	9.9	53.3	12.2	52.9	12.5	35.4	16.9	34.7	16.1	36.9	17.7	33.5	16.8
	37.4	9.0	39.6	9.3	53.0	12.7	52.9	11.8	35.3	15.9	34.4	14.3	36.7	14.7	33.4	18.2
	36.7	11.4	39.4	8.5	52.8	14.8	52.7	15.4	34.7	17.7	34.3	13.0	36.7	15.6	32.9	18.0
	36.6	11.8	39.3	9.0	52.6	13.1	52.6	13.9	34.3	15.6	34.0	13.8	36.1	16.9	32.8	17.4
	36.6	10.4	39.3	8.7	52.6	13.4	52.4	12.5	34.1	17.3	34.0	13.0	35.6	16.6	31.9	18.9
	36.6	9.7	38.8	10.3	52.4	13.2	52.2	12.6	34.1	18.2	33.6	12.6	35.4	14.4	31.5	17.9
	36.4	11.7	38.4	10.5	52.3	13.1	51.8	14.1	34.0	18.0	33.6	15.7	35.3	15.8	31.3	15.8
	36.2	10.5	38.4	8.9	52.1	14.2	51.8	12.6	33.7	18.1	33.6	13.8	35.2	16.0	31.2	15.8
	36.2	12.5	38.3	11.2	51.4	14.5	51.7	15.3	33.2	17.2	33.6	14.1	35.1	13.6	31.2	14.7
	36.0	10.9	38.3	9.9	51.3	12.8	51.6	14.2	33.0	17.6	33.3	14.8	34.8	17.5	30.6	17.4
	35.9	10.7	38.2	9.1	51.2	13.0	51.5	15.8	33.0	16.3	33.3	15.0	34.4	14.8	30.0	13.3
	35.5	11.2	38.1	10.4	51.1	11.9	51.3	14.4	32.9	15.9	33.2	15.2	34.4	15.1	29.6	14.9
	35.5	10.8	37.8	10.6	51.0	14.5	50.9	15.2	32.9	16.5	32.1	15.4	34.2	16.0	28.8	13.2
	35.4	11.5	37.5	9.3	51.0	13.8	50.9	14.4	32.9	17.0	31.9	15.2	34.1	15.6	28.6	12.8
	35.2	12.1	37.4	8.4	50.8	13.2	50.4	14.1	32.7	17.6	31.8	15.2	34.0	16.4	28.5	12.8
	35.0	10.4	37.0	11.0	50.3	16.6	50.3	15.5	32.7	16.5	31.4	16.5	34.0	17.0	28.1	14.7
	34.7	11.0	36.8	11.5	50.3	13.9	50.0	16.2	32.6	17.9	31.4	13.1	33.9	15.7	27.9	12.3
	34.3	10.6	36.3	11.4	50.0	15.2	49.8	14.7	32.2	16.5	30.6	13.9	33.7	14.2	27.9	14.9
	34.2	11.0	36.2	9.5	49.8	16.1	49.5	16.5	32.1	16.0	30.5	12.4	33.6	15.9	27.8	12.1
	34.2	11.3	36.1	10.4	49.7	13.5	48.7	13.7	32.0	15.3	29.1	13.8	33.5	16.5	27.7	12.7
	34.1	10.4	35.9	10.7	48.8	13.4	48.4	13.7	31.8	18.9	29.0	13.5	33.3	16.4	27.7	13.8
	33.8	10.3	35.9	10.3	48.7	16.1	48.1	14.8	31.5	17.9	28.8	15.0	33.3	15.1	27.4	19.3
	31.8	12.0	35.6	8.5	47.9	16.1	47.8	15.8	30.6	18.1	28.8	13.3	33.0	15.7	26.8	19.3
	31.7	11.6	35.0	9.6	47.1	16.2	47.0	14.5	30.5	15.0	28.5	14.5	33.0	14.7	26.7	11.9
	31.1	11.7	34.6	9.3	46.9	14.7	46.9	14.6	30.3	16.1	28.2	12.7	32.9	15.1	26.4	13.3
	30.9	12.0	34.1	9.2	45.3	16.4	46.6	15.5	30.1	17.1	28.1	13.4	32.6	16.4	26.2	13.6
Mean	35.8	10.8	38.0	9.6	51.2	13.8	51.2	14.1	33.7	17.1	32.5	14.7	35.6	16.2	30.9	15.9
SD	2.3	0.8	1.9	1.0	2.2	1.4	2.3	1.5	2.2	1.4	2.5	1.7	2.6	1.4	3.5	2.8

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Reliability and *SEM of Probe Scores*

Table 30 shows the alternate-form reliability (average correlation between probes) in the field test sample, as well as the corresponding *SEM*, at each grade. The median reliability is .88.

Table 30

Average Alternate-Form Reliability of M–COMP, by Grade

Grade	*N*	Reliability	*SEM*
1	919	.86	4.8
2	976	.82	4.8
3	971	.89	5.8
4	916	.85	6.6
5	1,048	.89	6.7
6	981	.89	5.9
7	944	.90	6.1
8	948	.88	6.5

Validity of Probe Scores

Criterion Validity of Probe Scores

During field testing, a correlation study with the *Group Mathematics Assessment and Diagnostic Evaluation* (G–MADE) was conducted at Grades 1, 3, and 8. M–COMP and G–MADE were given in counterbalanced sequence. As shown in Table 31, M–COMP correlated highly with the G–MADE total score.

Table 31

Criterion Validity of M–COMP Scores With Respect to G–MADE Scores

Grade	*N*	M–COMP		G–MADE		*r*
		Mean	*SD*	Mean	*SD*	
1	98	36.0	12.2	17.9	5.6	.84
3	98	45.5	14.2	15.9	5.7	.73
8	54	28.4	13.6	13.4	3.8	.76

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Reliability of the Rate of Improvement (ROI)

Table 32 reports the reliability and *SEM* of the ROI per week in M–COMP raw scores of students whose progress had been monitored at least 10 times during the 2010–2011 school year. The sample's average frequency of monitoring was two to three administrations per month. The reliability coefficient reported is a split-half measure based on the correlation between ROIs calculated from odd- and even-numbered probe administrations, adjusted using the Spearman-Brown formula.

Table 32

Reliability and *SEM* of the ROI of M–COMP Scores, by Grade

Grade	*N*	Reliability	*SEM*	Average duration (months)	Average number of administrations
1	3,285	.79	.25	6.4	15.6
2	6,289	.75	.23	6.8	15.9
3	6,687	.75	.29	7.0	15.8
4	6,756	.77	.32	6.8	15.7
5	6,183	.82	.27	6.9	15.8
6	3,833	.77	.23	7.0	15.5
7	2,534	.76	.23	6.9	15.7
8	2,510	.74	.20	6.9	15.4

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Spelling

Reliability of Probe Scores

Using samples of 30 students each in Grades 4 and 5 at a Minnesota elementary school, Tindal, Germann, and Deno (1983) studied the alternate-form and retest reliability of spelling measures constructed in the same manner as the **aims**web probes. As shown in Table 33, they found similar levels of reliability for two scoring methods: Words Spelled Correctly (WSC) and Correct Letter Sequences (CLS).

Table 33

Reliability of Spelling Scores

Grade	*N*	Type	Score	Reliability
4	30	Alternate Form	Words Spelled Correctly	.82
			Correct Letter Sequences	.82
5	30	Retest	Words Spelled Correctly	.94
			Correct Letter Sequences	.93

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Written Expression

Reliability of Probe Scores

McMaster and Campbell (2007) administered two 3-minute narrative-writing probes on each of two occasions (November and May) to students in Grades 3, 5, and 7 at a school in a large midwestern city. The resulting alternate-form reliabilities are reported in Table 34. (Values that were not statistically significant at the .001 level were not reported.)

Table 34

Alternate-Form Reliability of Written Expression Scores, by Grade

Grade	Score	November		May	
		N	Reliability	*N*	Reliability
3	Total Words Written	25	.73	21	.70
	Words Spelled Correctly		.78		.76
	Correct Writing Sequences		.86		
5	Total Words Written	43	.60	32	.70
	Words Spelled Correctly		.54		.78
	Correct Writing Sequences		.58		
7	Total Words Written	55	.70	41	.76
	Words Spelled Correctly		.69		.80
	Correct Writing Sequences		.69		.57

Gansle, Noell, VanDerHeyden, Naquin, and Slider (2002) administered two 3-minute story-writing probes on consecutive days to all 179 students in Grades 3 and 4 of a suburban school in the Southeast. Interscorer agreements (proportion agreement) were .96 (Total Words Written), .95 (Words Spelled Correctly), and .86 (Correct Writing Sequences). Alternate-form reliabilities were .62 (Total Words Written), .53 (Words Spelled Correctly), and .46 (Correct Writing Sequences).

Espin et al. (2000) administered two 3-minute story-writing probes in January to 87 students in Grades 7 and 8 of an urban middle school in the Midwest. Alternate-form reliabilities were .73 (Total Words Written), .72 (Words Spelled Correctly), and .76 (Correct Writing Sequences).

Espin et al. (2008) administered two narrative-writing probes in the fall of Grade 10 to 183 students at an urban midwestern high school. Although students were allowed to write for 10 minutes, their scores after 3, 5, and 7 minutes were also calculated. The alternate-form reliabilities for the 3-minute scores were .64 (Total Words Written), .64 (Words Spelled Correctly), and .66 (Correct Writing Sequences).

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Validity of Probe Scores

Jewell and Malecki (2005) administered a 3-minute narrative-writing probe to students in Grades 2 (*N* = 87), 4 (*N* = 59), and 6 (*N* = 57) in a rural Illinois school district. The students also took the *Stanford Achievement Test, Ninth Edition* (SAT9; Harcourt Brace Educational Measurement, 1997) in the same month as the writing probe. Correlations of writing-probe scores with SAT9 Language scores are shown in Table 35.

Table 35

Criterion Validity of Written Expression Scores (Jewell & Malecki, 2005)

Grade	*N*	Correlation with SAT9 Language score		
		Total Words Written	Words Spelled Correctly	Correct Writing Sequences
2	87	.24	.38	.57
4	59	.22	.29	.46
6	57	–.14	-.05	.23

Gansle, Noell, VanDerHeyden, Naquin, and Slider (2002) administered two 3-minute story-writing probes in February to all 179 students in Grades 3 and 4 of a suburban school in the Southeast. Probe scores were then correlated with teacher rankings of students' writing ability (collected at the same time) and either the Iowa Tests of Basic Skills, Language Score (Grade 3) or the Louisiana Educational Assessment Program Write Competently and Conventions of Language subtests (Grade 4), both administered in March. Correlations of writing probe scores with the criterion variables are reported in Table 36.

Table 36

Criterion Validity of Written Expression Scores (Gansle et al., 2002)

Grade	*N*	Criterion	Score	*r*
3 & 4	177	Teacher ranking of writing ability	Total Words Written	.08
			Words Spelled Correctly	.21
			Correct Writing Sequences	.36
3	75	ITBS Total Language Score	Total Words Written	.15
			Words Spelled Correctly	.24
			Correct Writing Sequences	.43
4	96	LA Educational Assessment Program, Write Competently subtest	Total Words Written	.28
			Words Spelled Correctly	.29
			Correct Writing Sequences	.28
		LA Educational Assessment Program, Conventions of Language subtest	Total Words Written	.16
			Words Spelled Correctly	.26
			Correct Writing Sequences	.41

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Espin et al. (2000) administered two 3-minute story-writing probes in January to 87 students in Grades 7 and 8 of an urban middle school in the Midwest. The criteria were teachers' June ratings of students' writing ability and, for the Grade 8 students, the district writing test. Validity coefficients are shown in Table 37.

Table 37

Criterion Validity of Written Expression Scores (Espin et al., 2000)

Criterion	Grade	*N*	Score	*r*
Teacher rating	7 & 8	87	Total Words Written	.46
			Words Spelled Correctly	.48
			Correct Writing Sequences	.59
District writing test	8	37	Total Words Written	.46
			Words Spelled Correctly	.48
			Correct Writing Sequences	.61

Espin et al. (2008) administered two narrative-writing probes in the fall of Grade 10 to 183 students at an urban Minnesota high school and correlated the average scores on the two probes with the students' results on the state writing test administered the following January. (Note that this correlation is higher than the correlation that would be obtained between a single probe and the criterion.) The state test was an untimed expository essay that was scored using the focused holistic method. Resulting validity coefficients were .26 for Total Words Written, .29 for Words Spelled Correctly, and .44 for Correct Writing Sequences.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Test of Early Literacy (TEL)

Rationale

A synthesis of the scientific research on reading by the National Reading Panel (2000) indicated some critical reading skills that should be assessed in kindergarten and early in Grade 1, including Phonemic Awareness and elements of Phonics such as letter names, letter sounds, and the ability to read non-real (nonsense) words. The **aims**web Test of Early Literacy (TEL) assesses these skills, as shown in Table 38. Letter Naming Fluency has been frequently identified as the best single indicator of potential reading failure, and Letter Sound Fluency has also been found to be a good predictor (Elliott, Lee, & Tollefson, 2001; Hintze, Ryan, & Stoner, 2003).

Table 38

TEL Measures in Relation to the Skills Identified by the National Reading Panel

Measure	Task	National Reading Panel area
Letter Naming Fluency	Say the names of visually presented letters for 1 minute	Phonics
Letter Sound Fluency	Say the sounds of visually presented letters for 1 minute	Phonics
Phonemic Segmentation Fluency	Identify the phonemes in orally presented words for 1 minute	Phonemic Awareness
Nonsense Word Fluency	Say the sounds of visually presented non-real words for 1 minute	Phonics

Reliability and *SEM* of Probe Scores

Elliott, Lee, and Tollefson (2001) administered measures designed identically to **aims**web Letter Naming Fluency, Letter Sound Fluency, and Phonemic Segmentation Fluency to 75 children at three schools in a midwestern city. Testing was conducted during the spring of kindergarten, in three testing sessions each two weeks apart. Students took alternate forms of each measure in the first and second sessions, and then took the same form in the third session as in the second session. In addition, 50 cases of each measure were independently scored by two scorers from audio recordings. Table 39 reports retest reliability, alternate-form reliability, and interscorer agreement for each measure.

Table 39

Reliability of TEL Scores (Elliott et al., 2001)

Type of reliability	Letter Naming Fluency	Letter Sound Fluency	Phonemic Segmentation Fluency
Retest	.90	.83	.85
Alternate-form	.80	.82	.84
Interscorer agreement	.94	.82	.87

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 40 reports the retest and alternate-form stability of scores for each of the four TEL measures over a four-month interval, based on **aims**web user data from 2007–2008. Because of the long interval between administrations, these are lower-bound estimates of reliability.

Table 40

Retest and Alternate-Form Reliability of TEL Scores

Grade	Interval	Type	*N*	Reliability	*SEM*
Letter Naming Fluency					
K	Fall–winter	Retest	1,463	.81	5.4
K	Fall–winter	Alternate-form	665	.73	7.4
K	Winter–spring	Alternate-form	674	.82	6.9
Letter Sound Fluency					
K	Winter–spring	Retest	1,222	.82	4.0
Phonemic Segmentation Fluency					
K	Winter–spring	Alternate-form	674	.59	10.9
1	Fall–winter	Alternate-form	942	.63	9.5
1	Winter–spring	Alternate-form	925	.61	9.5
Nonsense Word Fluency					
K	Winter–spring	Alternate-form	674	.71	10.2
1	Fall–winter	Alternate-form	942	.74	12.8
1	Winter–spring	Alternate-form	925	.78	14.7

Validity of Probe Scores

Table 41 reports the correlations of TEL measure scores with several criteria, based on **aims**web user data from the 2007–2008 and 2009–2010 school years. For many of these correlations, a two- or three-year interval existed between the TEL measure and the criterion.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 41

Criterion Validity of TEL Scores

Criterion	TEL grade (season)	Criterion grade (season)	*N*	*r*
Letter Naming Fluency				
Pennsylvania System of School Assessment	1 (fall)	1 (spring)	437	.47
Illinois Standards Achievement Test	K (fall, winter, spring)	3 (spring)	206–295	.51–.55
Minnesota Comprehensive Assessment	K (fall, winter, spring)	3 (spring)	75–76	.53–.60
Minnesota Comprehensive Assessment	1 (fall)	3 (spring)	75	.50
R–CBM	1 (fall)	1 (spring)	48	.76
R–CBM	1 (fall)	2 (fall, spring)	44	.72, .73
Letter Sound Fluency				
R–CBM	K (fall)	1 (winter)	28	.61
Illinois Standards Achievement Test	K (winter, spring)	3 (spring)	241, 247	.43, .52
Pennsylvania System of School Assessment	1 (fall)	1 (spring)	435	.33
Phonemic Segmentation Fluency				
Illinois Standards Achievement Test	K (winter, spring)	3 (spring)	289, 293	.33, .40
Minnesota Comprehensive Assessment	K (winter, spring)	3 (spring)	76, 76	.48, .56
Minnesota Comprehensive Assessment	1 (fall, winter, spring)	3 (spring)	130–134	.41–.51
R–CBM	1 (winter)	1 (spring)	46	.76
Nonsense Word Fluency				
Illinois Standards Achievement Test	K (winter, spring)	3 (spring)	174, 295	.49, .61
Minnesota Comprehensive Assessment	K (winter, spring)	3 (spring)	76, 76	.43, .55
R–CBM	K (spring)	1 (winter)	32	.68
R–CBM	1 (fall)	1 (spring)	46	.72
Pennsylvania System of School Assessment	1 (fall, winter, spring)	1 (spring)	434–438	.44–.51
Minnesota Comprehensive Assessment	1 (fall, winter, spring)	3 (spring)	130–134	.42–.53

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Elliott, Lee, and Tollefson (2001) administered measures designed identically to **aims**web Letter Naming Fluency, Letter Sound Fluency, and Phonemic Segmentation Fluency to 75 children at three schools in a midwestern city. Testing was conducted during the spring of kindergarten, in three testing sessions each two weeks apart. Students took alternate forms of each measure in the first and second sessions, and then took the same form in the third session as in the second session. Students also took the *Test of Phonological Awareness* (Torgesen & Bryant, 1994) in the week between the first and second sessions, as well as the *Woodcock-Johnson Psycho-Educational Battery, Revised* (Woodcock & Johnson, 1990) Broad Reading and Reading Skills clusters two weeks after the third session. Furthermore, teachers administered the *Developing Skills Checklist, Pre-Reading Total Score* (CTB Macmillan/McGraw-Hill, 1990) early in the school year and rated each student's level of prereading ability at the end of the year. Table 42 reports the correlations of each student's average score on the three administrations of each CBM measure with each of the criteria. Because averaging the scores from three administrations raises reliability, these correlations are higher than would be obtained for a single CBM measure.

Table 42

Criterion Validity of TEL Scores (Elliott et al., 2001)

Criterion	Letter Naming Fluency	Letter Sound Fluency	Phonemic Segmentation Fluency
Woodcock-Johnson Revised, Broad Reading	.63	.58	.44
Woodcock-Johnson Revised, Reading Skills	.75	.72	.60
Test of Phonological Awareness	.50	.68	.52
Teacher rating	.63	.62	.53
Developmental Skills Checklist	.67	.69	.54

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Reliability and *SEM* of Probe Scores

Clarke and Shinn (2004) administered TEN to 52 students in Grade 1 at two schools in the Pacific Northwest. Two forms of each TEN measure were given in the fall and winter, and one form was administered in the spring. Alternate-form reliabilities and test–retest stabilities (fall to winter and fall to spring) are reported in Table 43.

Table 43

Reliability and Stability of TEN Scores (Clarke & Shinn, 2004)

Type of reliability	Season or interval	Oral Counting	Number Identification	Quantity Discrimination	Missing Number
Alternate-form	Fall	NA	.89	.93	.83
	Winter	NA	.93	.92	.78
Retest stabilitiy	Fall–winter	.80	.85	.85	.79
	Fall–spring	.78	.76	.86	.81

Martinez, Missall, Graney, Aricak, and Clarke (2009) administered the TEN measures to 23 students at a rural midwestern school in the fall and spring of kindergarten. (Oral Counting was not administered in the spring.) In addition, a third administration was conducted two weeks after the spring administration. Table 44 reports alternate-form reliability (spring–spring) and stability (fall–spring) coefficients.

Table 44

Reliability and Stability of TEN Scores (Martinez et al., 2009)

Measure	Type of reliability	Season or interval	Reliability
Number Identification	Retest stability	Fall–spring	.92
	Alternate-form	Spring	.91
Quantity Discrimination	Retest stability	Fall–spring	.80
	Alternate-form	Spring	.77
Missing Number	Retest stability	Fall–spring	.89
	Alternate-form	Spring	.79

Baglici, Codding, and Tryon (2010) administered two forms of each TEN measure in the winter and spring of kindergarten to 61 students in a suburban school near New York City. Table 45 reports the alternate-form reliabilities for the kindergarten sample.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 45

Alternate-Form Reliability of TEN Scores (Baglici et al., 2010)

Measure	Season	Reliability
Number Identification	Fall	.84
	Winter	.71
Quantity Discrimination	Fall	.91
	Winter	.89
Missing Number	Fall	.86
	Winter	.81

Validity of Probe Scores

Clarke and Shinn (2004) administered TEN to 52 students in Grade 1 at two schools in the Pacific Northwest. Two forms of each TEN measure were given in the fall and winter, and one form was administered in the spring. In addition, students took the Applied Problems subtest of the *Woodcock-Johnson Psycho-Educational Battery, Revised* (Woodcock & Johnson, 1990) in the fall and spring, the *Number Knowledge Test* (Okamoto & Case, 1996) in the fall, and **aims**web Mathematics-Curriculum Based Measurement (median of three probes) in the winter and spring. Results of this study are presented in Table 46.

Table 46

Criterion Validity of TEN Scores (Clarke & Shinn, 2004)

TEN measure	TEN season	WJ-R Applied Problems		Number Knowledge Test	M–CBM	
		Fall	Spring	Fall	Winter	Spring
Oral Counting	Fall	.64	.72	.70	.56	.56
	Winter		.68		.49	.46
	Spring		.60			.50
Number Identification	Fall	.65	.72	.70	.68	.60
	Winter		.68		.66	.58
	Spring		.63			.60
Quantity Discrimination	Fall	.71	.79	.80	.76	.70
	Winter		.79		.71	.71
	Spring		.79			.75
Missing Number	Fall	.68	.72	.74	.78	.67
	Winter		.71		.75	.72
	Spring		.69			.74

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Clarke, Baker, Smolkowski, and Chard (2008) administered early numeracy measures constructed identically to **aims**web TEN to 55 kindergarten students at schools in the Pacific Northwest. The early numeracy measures and the *Stanford Early School Achievement Test, Fourth Edition* (SESAT-2; Harcourt Brace Educational Measurement, 1996) were administered at both the beginning and the end of the school year. Table 47 reports the correlations between early numeracy probe scores and criterion scores.

Table 47

Criterion Validity of TEN Scores (Clarke et al., 2008)

TEN measure	TEN season	SESAT-2 season	*r*
Oral Counting	Fall	Fall	.59
		Spring	.55
	Spring	Fall	.50
		Spring	.55
Number Identification	Fall	Fall	.53
		Spring	.58
	Spring	Fall	.52
		Spring	.61
Quantity Discrimination	Fall	Fall	.62
		Spring	.60
	Spring	Fall	.53
		Spring	.62
Missing Number	Fall	Fall	.60
		Spring	.57
	Spring	Fall	.58
		Spring	.64

Martinez, Missall, Graney, Aricak, and Clarke (2009) administered three of the TEN measures (i.e., not Oral Counting) and the *Stanford Achievement Test, Tenth Edition* (SAT10; Harcourt Educational Measurement, 2002) in the spring to 58 kindergarten students at a rural midwestern school. Fifty-two of the students had also taken the TEN measures (including Oral Counting) the previous fall. Table 48 presents the correlations of fall and spring TEN measure scores with scores on the SAT10 Mathematics subtest.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Table 48

Criterion Validity of TEN Scores, With Respect to SAT10 (Martinez et al., 2009)

Measure	Season	*r*
Oral Counting	Fall	.45
	Spring	
Number Identification	Fall	.31
	Spring	.44
Quantity Discrimination	Fall	.46
	Spring	.63
Missing Number	Fall	.36
	Spring	.47

In a longitudinal study, Baglici, Codding, and Tryon (2010) administered the TEN measures in the winter and spring to 61 kindergarten and Grade 1 students in a suburban school near New York City. The authors also administered the **aims**web Mathematics-Curriculum Based Assessment in the winter and spring of Grade 1, and collected teacher ratings of students' mathematics abilities at the end of Grade 1 using the *Academic Competence Evaluation Scales—Mathematics* (ACES; DiPerna & Elliott, 2000). Table 49 reports the predictive correlations of kindergarten TEN scores with the Grade 1 criteria.

Table 49

Criterion Validity of TEN Scores (Baglici et al., 2010)

TEN measure	TEN season (Kindergarten)	M–CBM (Grade 1)		Teacher rating (ACES), Grade 1
		Winter	Spring	
Oral Counting	Winter	.36	.35	.39
	Spring	.31	.33	.36
Number Identification	Winter	.41	.41	.46
	Spring	.37	.26	.57
Quantity Discrimination	Winter	.23	.22	.51
	Spring	.03	.02	.40
Missing Number	Winter	.42	.29	.44
	Spring	.52	.47	.58

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

References

Andren, K. J. (2010). *An analysis of the concurrent and predictive validity of curriculum based measures (CBM), the Measures of Academic Progress (MAP), and the New England Common Assessment Program (NECAP) for reading* (Dissertation, University of Southern Maine). Retrieved from http://www.usm.maine.edu/sites/default/files/School%20Psychology/Andren.pdf

Baglici, S. P., Codding, R., & Tryon, G. (2010). Extending the research on the Tests of Early Numeracy: Longitudinal analysis over two school years. *Assessment for Effective Intervention, 35,* 89–102.

Clarke, B., Baker, S., Smolkowski, K., & Chard, D. J. (2008). An analysis of early numeracy curriculum-based measurement. *Remedial and Special Education, 29,* 46–57.

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

CTB Macmillan/McGraw-Hill. (1990). *Developing Skills Checklist.* Riverside, CA: Author.

DiPerna, J. C., & Elliott, S. N. (2000). *Academic Competence Evaluation Scales.* San Antonio, TX: Pearson.

Elliott, J., Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills-Modified. *School Psychology Review, 30,* 33–49.

Espin, C., Shin, J., Deno, S., Skare, S., Robinson, S., & Benner, B. (2000). Identifying indicators of written expression proficiency for middle school students. *Journal of Special Education, 34,* 140–153.

Espin, C., Wallace, T., Campbell, H., Lembke, E. S., Long, J. D., & Ticha, R. (2008). Curriculum-based measurement in writing: Predicting the success of high-school students on state standards tests. *Exceptional Children, 74,* 174–193.

Fuchs, L., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21,* 45–59.

Gansle, K. A., Noell, G. H., VanDerHeyden, A. M., Naquin, G. M., & Slider, N. J. (2002). Moving beyond total words written: The reliability, criterion validity, and time cost of alternative measures for curriculum-based measurement in writing. *School Psychology Review, 31,* 477–497.

Harcourt Brace Educational Measurement. (1996). *Stanford Early School Achievement Test, Fourth Edition.* San Antonio, TX: Author.

Harcourt Brace Educational Measurement. (1997). *Stanford Achievement Test, Ninth Edition.* San Antonio, TX: Author.

Harcourt Educational Measurement. (2002). *Stanford Achievement Test, Tenth Edition.* San Antonio, TX: Author.

Hintze, J. M., Ryan, A. L, & Stoner, G. (2003). Concurrent validity and diagnostic accuracy of the Dynamic Indicators of Basic Early Literacy Skills and the Comprehensive Test of Phonological Processing. *School Psychology Review, 32,* 541–556.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Howe, K. B., & Shinn, M. M. (2002). *Standard reading assessment passages (RAPs) for use in general outcome measurement: A manual describing development and technical features.* San Antonio, TX: Pearson.

Jewell, J., & Malecki, C. K. (2005). The utility of CBM written language indices: An investigation of production-dependent, production-independent, and accurate-production scores. *School Psychology Review, 34,* 27–44.

Keller-Margulis, M. A., Shapiro, E. S., & Hintze, J. M. (2008). Long-term diagnostic accuracy of curriculum-based measures in reading and mathematics. *School Psychology Review, 37,* 374–390.

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology, 47,* 315–335.

Martinez, R. S., Missall, K. N., Graney, S. B., Aricak, O. T., & Clarke, B. (2009). Technical adequacy of early numeracy curriculum-based measurement in kindergarten. *Assessment for Effective Intervention, 34,* 116–125.

Mathematics Learning Study Committee & National Research Council. (2001). *Adding it up: Helping children learn mathematics.* Washington, DC: National Academy Press.

McMaster, K. L., & Campbell, H. (2007). *Technical features of new and existing CBM writing measures within and across grades.* (Technical Report No. 7.) Minneapolis, MN: Research Institute on Progress Monitoring.

Merino, K., & Beckman, T. O. (2010). Using reading curriculum-based measurements as predictors for the Measure of Academic Progress (MAP) standardized test in Nebraska. *International Journal of Psychology: A Biopsychosocial Approach, 6,* 85–98.

MetaMetrics. (2011). *Linking the R–CBM and Maze with the Lexile Framework for Reading.* Durham, NC: Author.

National Council of Teachers of Mathematics. (1989). *Principles and standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Council of Teachers of Mathematics (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence.* Reston, VA: Author.

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U. S. Department of Education.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Okamoto, Y., & Case, R. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. *Monographs of the Society for Research in Child Development, 61,* 27–59.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests. *Journal of Psychoeducational Assessment, 24,* 19–35.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Silberglitt, B., & Hintze, J. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23,* 304–325.

Stenner, A. J. (1997). The Lexile Framework: A map to higher levels of achievement. Durham, NC: MetaMetrics.

Tindal, G., Germann, G., & Deno, S. (1983). *Descriptive research on the Pine County norms: A compilation of findings* (Research Report No. 132). Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities, University of Minnesota.

Torgesen, J. K., & Bryant, B. R. (1994). *Test of Phonological Awareness.* Burlingame, CA: Psychological and Educational Publications.

Woodcock, R. W., & Johnson, M. B. (1990). *Woodcock-Johnson Psycho-Educational Battery, Revised.* Allen, TX: DLM Teaching Resources.

Copyright © 2012 NCS Pearson, Inc. All rights reserved.