Technical Report # 1408

Technical Manual: easyCBM

Daniel Anderson

Julie Alonzo

Gerald Tindal

Dan Farley

P. Shawn Irvin

Cheng-Fei Lai

Jessica L. Saven

Kraig A. Wray

University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

Technical Manual: easyCBM

Edited by

Daniel Anderson, Julie Alonzo, and Gerald Tindal

The following authors, listed alphabetically by last name, contributed to this technical manual. Because their contributions are woven throughout multiple chapters, authors' contributions are noted by analysis rather than chapter.

Author	**Summary Contribution**
Dan Farley	Criterion Validity Evidence
P. Shawn Irvin	Measurement Development & Item Alignment
Cheng-Fei Lai	Construct Validity Evidence
Jessica L. Saven	National Norms
Kraig Wray	Reliability Evidence

Funds for the datasets used to generate this report came from the following federal grants awarded to the UO.

From the U.S. Office of Special Education Programs

- Response to Intervention with Reading Curriculum-Based Measures: Steppingstones of Technology Innovation for Children with Disabilities (H327A090005 funded 2009 – 2011).

From the U.S. Department of Education: Institute of Education Sciences

- Developing Middle School Mathematics Progress Monitoring Measures *(*R324A100026 funded from 2010 - 2014).
- Reliability and Validity Evidence for Progress Measures in Reading (R324A100014 funded from 2010 – 2014).
- Statewide Longitudinal Data Systems (with Oregon Department of Education). Budget $3,717,220 from May 2009 - Sep 2012
- Postdoctoral Fellowships on Progress Monitoring in Reading and Mathematics (R305B080004 funded from 2008 – 2012).
- Assessments Aligned with Grade Level Content Standards and Scaled to Reflect Growth for Students with Disabilities (SWD) (H327A070188 funded from 2007-2011).

In addition, we would like to acknowledge the following members of the research team at Behavioral Research and Teaching for their assistance in compiling this technical manual:

Steffani Mast
Raina Megert
Denise L. Swanson

The good work of the following computer programmers has made easyCBM possible:

Trevor Cords
Aaron Glasgow
Kirt Ulmer

TABLE OF CONTENTS

Chapter 1.	Overview	1
Chapter 2.	Methodology and Results of Norming Study	3
Chapter 3.	Methodological Overview	19
Chapter 4.	Letter Names	30
Chapter 5.	Letter Sounds	36
Chapter 6.	Phoneme Segmenting	43
Chapter 7.	Word Reading Fluency	55
Chapter 8.	Passage Reading Fluency	72
Chapter 9.	Vocabulary	91
Chapter 10.	Multiple Choice Reading Comprehension	100
Chapter 11.	CCSS Reading	122
Chapter 12.	NCTM Math Measures	133
Chapter 13.	CCSS Math Measures	155
Chapter 14.	Spanish Measures	169

Chapter 1

Overview

Since the easyCBM© learning system was first published in 2006, over $8 million of federal funding (both from the Office of Special Education Programs and more recently from the Institute of Education Sciences) has been used to develop, study, and refine the assessments available on the system. This Technical Manual summarizes the ongoing research that is the foundation of the easyCBM© assessments in reading, mathematics, and Spanish literacy.

Two different versions of the easyCBM© system exist: easyCBM Lite©, which is intended for individual teacher use and District easyCBM©, which is intended for systems-level implementation at the school or district level. At the time this Technical Manual was published, over 425,000 educators from across the United States, as well as a few international locations, had easyCBM© accounts, and almost 4 million students had taken over 26 million easyCBM© assessments.

To take advantage of this user-base, we developed new norms for easyCBM during the 2013-2014 school year to better represent reading and mathematics performance across the nation. Previous norms were based on all students taking the measures at any given grade and benchmark season (fall, winter, and spring). Therefore, the scores reflected all students who took the measures (at that time period) rather than the demographics of specific regions and student characteristics (e.g., race-ethnicity and gender). In conducting an analysis of the user base, these students were predominately from the west and white. Therefore, to provide a more comparable base of students in the U.S., new norms were developed using a stratified random sample of students that reflected the children attending schools proportionate to their demographics (region and race-ethnicity-gender). We used the most recent Common Core Data published by the

National Center on Education Statistics[1] to first determine the counts and percentages and then sampled accordingly using the following process.

With such widespread use, analyzing the measures' reliability and evidence of the validity of using them as screening instruments and progress monitoring measures, is essential. To this end, researchers at the University of Oregon's Behavioral Research and Teaching, where the easyCBM© system was developed and continues to be refined, regularly publish the results of studies documenting every step in the assessment development and refinement process (please see http://www.brtprojects.org/publications/technical-reports to access the complete array of technical reports that have been published to date). However, with so much research being conducted and written up, the sheer volume of publications can be overwhelming.

This Technical Manual addresses the need for a single document that summarizes the results of all previously published technical reports, highlighting the key findings from each. For ease of access, the Technical Manual is organized by chapter, with Chapter 2 providing information on the analytic procedures referenced in subsequent chapters, and each measure discussed in its own chapter, beginning with the Reading measures, then moving on to the Mathematics measures, and finally ending with the Spanish literacy measures.

[1] U.S. Department of Education, National Center for Education Statistics, Common Core of Data (CCD), "Public Elementary/Secondary School Universe Survey", 2010-11, Version 2a; and "Local Education Agency Universe Survey", 2010-11, Version 2a; and "State Nonfiscal Survey of Public Elementary/Secondary Education", 2010-11, Version 1a.

Chapter 2: Methodology and Results of Norming Study

Norms were developed in 2013-2014 with a fixed group of 500 students to represent each region and race-ethnicity-gender. This sample size was determined to be optimal in capturing the largest and most stable group of students with scores on the various measures. When fewer students were present, we weighted the population to reflect 500 (as described below).

Process for Establishing Norms

1. We first assigned each state in the CCD database to a region using the following associations (ignoring U.S. territories).

State	Region	State	Region	State	Region	State	Region
AR	MW	CT	NE	AL	SE	AK	W
IL	MW	DE	NE	FL	SE	AZ	W
IN	MW	ME	NE	GA	SE	CA	W
IA	MW	MA	NE	KY	SE	CO	W
KS	MW	NH	NE	LA	SE	HI	W
MI	MW	NJ	NE	MD	SE	ID	W
MN	MW	NY	NE	MS	SE	MT	W
MO	MW	PA	NE	NC	SE	NV	W
NE	MW	RI	NE	SC	SE	NM	W
ND	MW	VT	NE	TN	SE	OR	W
OH	MW			VA	SE	UT	W
OK	MW			WV	SE	WA	W
SD	MW					WY	W
TX	MW						
WI	MW						

2. We then counted the number of students in each state by their grade, race-ethnicity, and gender classification using the following codes from CCD.

Grade	**Race-Ethnicity**	**Gender**
KG=kindergarten	AM= American Indian/Alaska Native	M=Male / F=Female
1-8=Grades 1…8	AS= Asian/Hawaiian Native/Pacific Islander or Asian	
	HI=Hispanic	
	BL=Black	
	WH=White	
	HP= Hawaiian Native/Pacific Islander	
	TR= Two or more races	

3. We calculated the number and percentage of students in each region for each grade, race-ethnicity, and gender. For example, in the following table, we report for each region, the number and percentage of Kindergarten (K) American Indian/Alaska Native (AM) male or female (F) students. Note that the percentage for male and female within each grade and race-ethnicity is the total within gender divided by the total of both genders.

Region	**Count/%**	**K-AM-M**	**K-AM-F**
MW	Count	9,133	8,563
	Percentage	51.61%	48.39%
NE	Count	1,121	1,055
	Percentage	51.52%	48.48%
SE	Count	2,208	2,070
	Percentage	51.61%	48.39%
W	Count	8,656	8,515
	Percentage	50.41%	49.59%

4. Within each region and based on the total count of students at each grade level (across all race-ethnicity and gender groups), we then calculated the number of students needed (to be proportionate) within each race-ethnicity and gender category assuming that 500 students within each category needed to be present. For example, in the table below, the total number of Kindergarten students in the Midwest (MW) is just over 1.24 million with approximately .01 percentage of them American Indian/Alaska Native (AM) from the total above (9,133 males and 8,563 females). If we assume the need for 500 students in the MW region, the number of students we must sample in Kindergarten who are AM and male is 4 and the number of students to sample who are AM and female is 3.

Region	Count	Percent	K-AM-M	K-AM-F
MW	1,241,095	0.014258377	4	3
NE	574,221	1.74149E-06	1	1
SE	954,967	0.002278613	1	1
W	905,870	1.10391E-06	5	5
TOTAL	3,676,153	0.001163717	11	10

5. Finally, we calculated the number of students who took each measure in easyCBM (eliminating students in the data file with missing scores for each measure).

   Regional Random Sample. In our example, because we are determining the norms for Kindergarten students, we applied this step to letter names, letter sounds, and phoneme segmentation for literacy, as well as the grade appropriate measures in mathematics. Therefore, we determined the following numbers of students with **letter names (LN)** scores (under the 'have' column). This step informed us of the need to randomly sample (RS) by region 38% of the students in the MW, 96% of the students in the NE, 7% of the students in the SE, and 4% of the students in the west.

Region	Need	Have	RS @500
1. MW	500	1,330	0.38
2. NE	500	521	0.96
3. SE	500	6,734	0.07
4. W	500	11,171	0.04

Race-Ethnicity-Gender Random Sample. Continuing in this example using Kindergarten students, we then sampled again using these student demographic categories. In the example below, we determined the following numbers of students with **letter names (LN)** scores (under the 'have' row). This step informed us of the need to randomly sample (RS) 22% of the white-male student population, 23% of the students who were white-female, 38% of the students in the non-white-male population and, and 39% of the students who were non-white female.

LN	WhiteM	WhiteF	NWhiteM	NWhiteF
Need	500	500	500	500
Have	2284	2144	1309	1270
RS	0.22	0.23	0.38	0.39

Note that this step sometimes resulted in having fewer students (in the region but more typically in the gender-race-ethnicity category) than had taken the measure for that season. Therefore, we weighted the sample to equal 500 students. For example, if only 370 students took the LN measure (and we needed 500), we weighted the sample by 1.35. For a few measures (that had been only recently introduced), the number of students taking it was below half of the number needed (<250) in which case we did not compute norms.

Data Display Tables

In summary, we followed the steps above to calculate norms for every measure at every grade level and time period (season). We calculated two sets of norms for district and teacher use, respectively: (a) stratified by region and race-ethnicity- gender and (b) reported for each score value (again for each measure, grade level, and time period).

District use norms are presented using two formats: (a) a short form that displays all the grades and seasons within a measure type, reporting the following percentile ranks (PR): 10^{th}, 20^{th}, 50^{th}, 75^{th}, and 90^{th} and (a) long form within each measure, season, and grade that provides specific region and race-ethnicity-gender information, both means and standard deviations as well as PRs at the 5^{th}, 10^{th}, 15^{th}, 20^{th}, 25^{th}, 30^{th}, 50^{th}, 75^{th}, and 90^{th} percentile ranks.

Teacher use norms provide Individual Percentile Ranks (IPRs) so they can determine the exact PR for any given score value. We calculated these PRs by first sampling a unique and random sample of 500 students from every region who took a measure (at every grade and time period) and assembled a file with the total group (usually 2,000 students unless the region had fewer than 500). We then calculated a percentile rank for each student and pivoted the table to reflect the number of students and the PR at each score value. We counted the students at each score value and interpolated the PR when no score value was present. These tables have been incorporated into the classroom rosters that teachers can access within easyCBM and have been converted to a PDF for printing.

Important Disclaimers

We combined race with ethnicity, as the count of students with sufficient data on both variables was insufficient for disaggregation. We included non-white Hispanic as non-white only if actively marked with another race (other than white).

The two sets of norms (district use and teacher use) may differ slightly as they have been based on different random samples of students from each region. This slight variation provides greater generalization to the outcomes and should remind teachers and administrators that all measurement contains some minor amount of error and score values are best represented in confidence bands (the range of scores within which we can be confident a specific score is present). Note also that the IPRs are sampled proportionately by region not by race-ethnicity-gender.

All literacy norms are based on score values present in 2012-2013. All math norms are based on score values present in 2013-2014.

Letter Names

Grade K			
Percentile	**Fall**	**Winter**	**Spring**
10th	3	*	*
25th	11	*	*
50th	24	*	*
75th	34	*	*
90th	45	*	*

*Not normed in this cycle

Letter Sounds

	Grade K			Grade 1		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	0	6	19	15	25	30
25th	1	14	27	24	32	37
50th	6	26	35	31	41	45
75th	13	34	44	37	50	52
90th	24	39	52	44	58	62

Phoneme Segmenting

	Grade K			Grade 1		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	0	6	21	15	*	*
25th	0	16	33	27	*	*
50th	6	31	43	37	*	*
75th	15	41	51	46	*	*
90th	29	50	59	54	*	*

*Not normed in this cycle

Word Reading Fluency

	Grade K			Grade 1			Grade 2		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	*	1	3	3	10	18	13	18	32
25th	*	2	7	8	16	30	24	35	48
50th	*	3	13	15	28	49	41	53	65
75th	*	7	22	31	49	70	58	68	80
90th	*	13	41	54	69	84	72	80	92

*Not normed in this cycle

Passage Reading Fluency

	Grade 1			Grade 2			Grade 3		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	*	10	20	26	32	45	47	72	64
25th	*	16	37	41	57	73	68	92	89
50th	*	32	60	64	83	102	87	117	116
75th	*	69	95	89	108	129	112	150	144
90th	*	107	124	116	128	156	138	172	174

*Not normed in this cycle

Passage Reading Fluency (continued)

	Grade 4			Grade 5			Grade 6		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	69	85	87	90	97	115	92	94	102
25th	87	112	112	120	123	137	114	128	132
50th	107	138	138	145	150	166	239	156	165
75th	132	159	167	169	176	193	164	183	198
90th	156	181	193	193	204	212	189	209	222

	Grade 7			Grade 8		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	110	109	103	118	107	109
25th	127	137	135	140	127	130
50th	150	166	163	167	157	160
75th	174	193	190	194	186	184
90th	200	219	213	219	211	207

Vocabulary

	Grade 2			Grade 3			Grade 4		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	4	6	7	9	11	13	10	12	13
25th	5	9	10	13	14	16	13	15	16
50th	9	11	11	16	17	18	16	17	18
75th	11	12	12	18	19	19	19	19	19
90th	12	12	12	19	20	20	20	20	20

	Grade 5			Grade 6			Grade 7		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	11	12	13	13	13	15	11	12	12
25th	14	15	16	16	16	17	15	16	16
50th	17	17	18	18	18	18	17	18	18
75th	18	19	19	19	19	19	19	19	19
90th	19	19	20	20	20	20	20	20	20

Vocabulary (continued)

Grade 8			
Percentile	**Fall**	**Winter**	**Spring**
10th	13	13	13
25th	16	16	17
50th	18	18	18
75th	19	19	19
90th	20	20	20

Multiple Choice Reading Comprehension

	Grade 2			Grade 3			Grade 4		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	3	4	5	6	6	7	6	7	7
25th	5	6	7	8	8	10	8	10	11
50th	7	9	10	11	11	14	12	14	15
75th	9	11	11	14	13	16	15	16	17
90th	10	11	12	15	15	18	17	18	18

Multiple Choice Reading Comprehension (continued)

	Grade 5			Grade 6			Grade 7		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	8	9	9	9	9	9	8	9	7
25th	12	14	13	13	11	12	11	12	10
50th	14	16	15	15	14	15	14	15	12
75th	16	18	16	17	16	17	16	17	14
90th	17	19	17	18	17	18	18	18	16

	Grade 8		
Percentile	**Fall**	**Winter**	**Spring**
10th	9	8	7
25th	12	11	11
50th	15	13	13
75th	17	15	15
90th	18	16	17

CCSS Reading

	Grade 3			Grade 4			Grade 5		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	7	9	10	7	11	4	9	8	9
25th	13	17	19	16	18	18	16	18	17
50th	20	21	23	21	21	23	20	22	21
75th	23	22	24	23	23	24	22	23	23
90th	25	24	25	24	24	25	23	24	24

	Grade 6			Grade 7			Grade 8		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	8	12	11	6	13	11	11	11	11
25th	17	18	18	15	19	18	18	17	18
50th	21	22	21	20	22	22	22	21	21
75th	23	23	22	22	23	24	23	23	23
90th	24	24	23	23	24	25	24	24	24

CCSS Math

	Grade K			Grade 1			Grade 2		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	10	13	15	14	18	20	14	17	18
25th	13	17	19	16	22	25	17	21	23
50th	16	20	22	20	26	29	21	25	28
75th	19	23	25	23	29	31	25	29	31
90th	22	26	27	26	32	33	30	32	33

	Grade 3			Grade 4			Grade 5		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	17	21	20	15	16	17	16	17	18
25th	21	25	29	20	22	25	20	22	24
50th	25	29	34	25	27	31	25	27	29
75th	28	32	36	30	32	35	29	31	34
90th	32	36	38	34	34	38	33	35	37

CCSS Math (continued)

	Grade 6			Grade 7			Grade 8		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	16	14	14	14	14	14	14	16	16
25th	20	20	22	18	19	20	18	21	23
50th	25	26	30	23	26	28	23	28	31
75th	30	32	36	28	33	35	31	34	37
90th	34	37	39	34	38	39	36	39	41

NCTM Math

	Grade K			Grade 1			Grade 2		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	17	24	28	18	23	25	16	18	23
25th	21	30	35	22	28	33	20	24	29
50th	27	36	38	26	34	38	25	33	35
75th	33	40	41	31	39	41	30	39	41
90th	38	43	43	35	42	43	35	42	43

NCTM Math (continued)

	Grade 3			Grade 4			Grade 5		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	22	26	28	21	24	25	21	22	28
25th	27	31	34	27	30	31	26	30	37
50th	31	36	39	32	35	36	31	36	41
75th	35	39	42	38	39	40	37	41	43
90th	39	42	44	41	42	43	41	43	44

	Grade 6			Grade 7			Grade 8		
Percentile	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**	**Fall**	**Winter**	**Spring**
10th	19	20	18	17	17	17	16	17	16
25th	24	27	27	22	23	23	21	23	22
50th	31	32	35	29	30	31	27	30	30
75th	36	38	41	36	37	38	34	37	37
90th	40	42	43	40	42	42	39	42	42

Chapter 3: Methodological Overview

In this chapter, we provide a brief, conceptual and theoretical overview of the methodologies used to examine the technical adequacy evidence for easyCBM©. We use a similar structure to other chapters, starting with measurement development, then moving to reliability, and concluding with validity. Within each section, we discuss major methodologies used, why the specific methodology was chosen over others, and what we hoped to learn by using the methodology.

Measurement Development

During the development of the easyCBM© measures, an array of statistical methodologies were used. Two of these methodologies, Rasch modeling and analysis of variance (ANOVA), were fundamental to the construction of test forms and are discussed below. Generally, Rasch modeling provides information on the difficulty and functioning of each individual item, while ANOVA provides information on differences in terms of mean difficulty between test forms.

Rasch Modeling

Rasch modeling is part of the item response theory (IRT) family of statistical models for test development. Rasch modeling offers certain advantages over classical test statistics, two of which were central to the development of easyCBM©. First, each item can be calibrated onto the same scale, with an *item difficulty* statistic calculated. That is, the difficulty of each item can be directly compared. This information can be used to construct alternate test forms that are equivalently difficult and have similar ranges of item difficulties. Second, the *item fit* to the model expectations can be evaluated. That is, are students responding to the item erratically? For instance, if students of high ability (as indicated by their responses to all other items) are

responding to an easy item incorrectly, and/or students of low ability are correctly responding to a difficult item, then the item would not fit the model expectations well. Evaluating item fit helps to "weed out" poor items. It is important to note, however, that Rasch modeling can only be used for tests that include discrete items. For a more in complete introduction to Rasch modeling, we recommend Snyder and Sheehan (1992) and Bond and Fox (2007).

Analysis of Variance (ANOVA)

Passage reading fluency tests do not have discrete items, unlike other easyCBM© measures. Rasch modeling was therefore impossible. Constructing alternate forms of equivalent difficulty is vital to the validity of monitoring students' growth over time. If changes in test form difficulty occur then changes in students' scores cannot be meaningfully interpreted as changes in students ability, given that the change in score could just as easily have been the result of the test form changing difficulty. A common practice to constructing passages to be equivalently difficult is to "level" the passages with readability formula (e.g., Lexile®, Spache, etc.). However, research generally finds that readability formula alone are not sufficient to ensure passage equivalence (Ardoin, Suldo, Witt, Aldrich, & McDonald, 2005).

In the construction of the easyCBM© Passage Reading Fluency forms, we used the Flesch-Kincaid readability formula as an initial estimate of passage difficulties. We then piloted all test forms to examine how the test forms functioned *empirically*. ANOVAs were conducted to examine whether the mean difficulty of one test form was statistically more or less difficult than the other test forms. The ANOVAs allowed us to test for mean differences in terms of difficulty across multiple test forms at once. Discrepant test forms were then revised to bring them into alignment with the other test forms prior to release for operational use.

Reliability

Reliability is a prerequisite to validity, and refers to the consistency of test scores, or the test items, in measuring the skill of interest. There are three types of reliability that are relevant to easyCBM©: internal consistency, alternate form, and test-retest. An overview of these reliability types, as well as the methods used to address them, is discussed below. We also dedicate a section to Generalizability theory analyses, given that the technique can be used to address multiple areas of reliability.

Internal Consistency

Internal consistency refers to the consistency of the test items, or test features (i.e., passage), in measuring the same trait. Perhaps the most commonly used measure of internal consistency is Cronbach's alpha, also called coefficient alpha. Cronbach's alpha requires the test have discrete items and the formula is based on the correlations among the items. Cronbach's alpha ranges from 0-1.0, with higher values indicating more reliable measurement. Although no "rules" exist for interpreting Cronbach's alpha, general rules of thumb suggest measures should have a value of at least 0.8 for acceptable internal consistency.

Alternate Form Reliability

Theoretically, if students' ability did not change, then they should receive the same score on each easyCBM® alternate test form, given that the forms were constructed to be equivalently difficult. Alternate form reliability analyses empirically test this theory by examining the correlations between multiple alternate test forms administered to students on the same day. Of course, in reality we not expect students to receive the *exact* same score across test forms (given difference in motivation, attentiveness to the test items, etc.), but the scores should be reasonably consistent (i.e., reliable) so that we can have confidence that the scores represent students' true

level of ability (and are not dependent on the test form or the time the student took the test). The order of test form administration is generally counter-balanced across students in alternate form reliability studies so test exhaustion effects do not impact the correlations between forms.

Alternate form reliability is generally evaluated with standard Pearson's bivariate correlations, which range from -1.0 to 1.0, but in practice (with alternate form reliability analyses) values should never be negative. Similar to Cronbach's alpha, values above 0.8 are generally considered strong for alternate form reliability. Pearson's correlation (*r*) assumes, however, that the distribution of the variable approximates a normal distribution. If the distribution of students' scores is skewed, nonparametric statistics may be required (e.g., Spearman's Rho), which are interpreted similarly to Pearson's *r*.

Test-retest Reliability

Test-retest reliability is evaluated with similar statistics to alternate form reliability, and has a similar theoretical purpose. The primary difference is that rather than evaluating the correlation between students' scores on alternate test forms, the correlation between students' scores on the *same test* is evaluated at two points in time. If students' ability were held constant, they should receive the same score on one specific test form each time they took it (assuming practice effects did not exist). To ensure students' ability does not change (at least dramatically) the tests must be administered within a small time frame, typically one week apart. Test-retest correlations are interpreted similarly to Cronbach's alpha and alternate test form.

Generalizability Theory

Generalizability theory evaluates the variance associated with the test, the individual student, and various "facets" of the measurement process (such as the test items, form, or the day the test was administered). Decision studies generally follow generalizability studies, and

provide an indication of how one could alter the measurement process to obtain a sufficiently reliable estimate of students' abilities (e.g., administer three tests instead of one, increase the number of items in the test, etc.). Decision studies will not be discussed in this technical report, but can be found in the full reports (see the corresponding measure for the specific report).

Generalizability theory allows for variance in students' scores to be parsed into student and facet factors. For instance, if low test-retest reliability was found, was it due to the time between testing occasions? The test form itself? The students? An interaction among these facets? Generalizability studies allow us to delve deeper into the measurement process to answer these questions. After a generalizability theory analysis has been conducted, a G-coefficient can be produced, which summarizes the reliability or internal consistency of the measure. It is interpreted analogously to Cronbach's alpha and represents the ratio of variance in scores attributable to students relative to the total variance (total variance including variance from each facet under consideration, plus error variance). Like Rasch modeling, generalizability is a complex topic, and a full description of the method is beyond the scope of this document. For more extended reading, we recommend Brennan (1992, 2001), and Shavelson and Webb (1991).

Validity Evidence

Test validity is a complex and nuanced topic, with various researchers providing slightly different interpretations of the topic over the years. According to Messick (1995), "Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores" (p. 741). This definition implies that no test can be valid; rather, it is the use and interpretation of the test scores that are or are not valid. All easyCBM© measures were developed for use within an RTI framework, and, as such, were investigated for their validity

within this context. No single study can validate a measure for an intended use. Rather, it is the accumulation of evidence over time that result in our confidence in the measure being more or less valid for an intended use. Validity is not an all or nothing property, but is a continuum along which specific measures sit for different uses.

In what follows, we discuss the primary methodologies used to investigate the validity of the use of easyCBM© within an RTI framework. We begin by discussing criterion validity studies, which explore the degree to which easyCBM© scores correlate with other measures with known validity for measuring the underlying skill. We have conducted both concurrent (at the same time) and predictive (criterion administered at a later point in time) criterion studies. We conclude by discussing construct validity evidence, which was garnered primarily through latent factor analyses.

Criterion Validity

Criterion validity studies explore the relation between a focal measure (e.g., easyCBM©) and a criterion measure (e.g., state test). The criterion measure must be chosen carfully, as it represents the "gold standard" in the analysis. Criterion measures must have documented reliability and validity evidence for measuring the underlying skill targeted by the focal measure. If the focal measure has a strong relation with the criterion measure, then our confidence in the validity of the focal measure to measure the underlying skill is increased. Focal measures generally have a different purpose than the criterion measure in the study (e.g., formative versus summative) and we would therefore not expect the relation to be perfect. However, if the relation between the measures is not sufficiently high, then we may have little confidence that the measures are tapping the same underlying skill. In this case, the validity of the focal measure for measuring the underlying skill would be questionable.

Methodological Overview

Criterion validity is typically divided into predictive and concurrent subcomponents. Predictive validity studies explore how well a test taken at one point in time will predict performance on a criterion measure taken at a later date. For instance, many easyCBM© predictive validity studies examine how the fall and winter benchmark tests predict performance on the state test taken during the spring. Predictive validity is most important when the test is being used as a screener for some future potential issue (e.g., risk for future low achievement). Concurrent validity studies explore the relation between two measures taken within the same time frame (e.g., one week). Similar to alternate form/test-retest reliability, the students' ability is thought to stay constant between assessment occassions. The relation between the measures then more accurately represents the degree to which the measures tap the same underlying skill. Generally, measures used within RTI should be relatively short and easy to administer. Yet, they should maintain strong predictive and concurrent validity, which can be established by examining the relation between the measure used within RTI and more comprehensive summative measures designed to measure the same skill.

Criterion validity studies typicaly use linear regression to examine the relation between the focal and criterion measures. Regression slopes are then calculated, from which the percent of variance in the criterion measure accounted for by the focal measure can be calculated. The percent of variance accounted for provides some indication of how accurate the regression slope is in predicting students' scores, with higher variance accounted for resulting in more accurate predictions. Regression models can also include a host of additional "control" variables, such as student demographics, to help provide a clearer picture of the "true" relation between the measures. Along with regression coefficients, criterion valdity studies typically include estimation of the raw correlation between the measures, with higher correlations indicating a

stronger relation between the measures. It is important to note that regression and correlation are not entirely separate concepts, as regression models are based on correlation structures.

Regression models and standard correlations (Pearson's *r*) assume the distribution of test scores on each measure approximates a normal curve. For some measures, particularly those in early reading (Grade K), this assumption may be too restrictive. Nonparametric statistics make fewer assumptions about the distributions of variables, but can generally be interpreted analogously to the standard, parametric statistics. In the case of non-normal distributions, we used Spearman's rho (r_s) rather than Pearson's *r* to evaluate criterion validity evidence. Spearman's rho is, essentially, the nonparametric parallel to Pearson's *r* and can be interpreted identically.

Construct validity

Construct validity is perhaps the most complicated form of validity to assess, as one must evaluate the degree to which the test measures the underlying construct (or skill) it purports to measure. There are various methods that can be used to evaluate construct validity, but perhaps the most common are through expert judgment, and statistical models known as latent factor analyses. The construct validity of easyCBM© has been evaluated largely through the latter, although alignment studies can provide some evidence of construct validity as well.

Factor Analysis

Factor analyses can be either exploratory, where the number of latent factors are extracted based on the correlations among the items, or confirmatory, where the number of latent factors is specified *a priori* as well as which items load on which factors. That is, if a math test were intended to measure geometry and algebra, a two factor confirmatory model may be specified. However, if there was not a specified structure behind the test, then an exploratory

model may be specified to explore how many factors are present in the test. After the factors are extracted, items within each factor are examined and each factor is provided a label. For instance, with the math test example suppose three factors were extracted. The items within each factor could then be evaluated, and a label could be assigned to the factor that adequately represents all the items within that factor, such as *Number and Operations, Algebra,* and *Ratios and Proportions*.

The construct validity of easyCBM© has been evaluated primarily through confirmatory factor analyses. Confirmatory factor analyses essentially evaluate the degree to which the correlations among the items/measures "go along" with the hypothesized structure. If the observed correlation structure closely matches the hypothesized structure, then the model will fit the data well. However, if the variables correlate in ways outside the hypothesized structure, the model will not fit the data well and little to no evidence will exist to support the hypothesized model. The easyCBM© tests were constructed with a hypothesized structure in mind for all tets *a priori*. The fit of the data to the model was then evaluated. Items or tests that did not fit the hypothesized model were then revised so the data fit the model adequately.

Rasch modeling, discussed earlier in this chapter, is also a restricted form of factor analysis. Rasch models assume all items measure a single latent factor. However, Rasch modeling is a restricted form of factor analysis because each item is assumed to contribute to the latent factor equally (an assumption standard factor analyses do not make). Yet, the model results can still be used as evidence for construct validity by evaluating each item's fit to the model. Items that fit the model well show good evidence of measuring the latent factor, while items that do not fit the model well show poor evidence. The item fit for all items within a test can then be

plotted to get an overall picture of the degree to which items within the test measure the same latent factor (e.g., see Anderson et al., 2010).

References

Anderson, D., Lai, C. F., Nese, J. F. T., Park, B. J., Sáez, L., Jamgochian, E. M., et al. (2010). Technical adequacy of the easyCBM primary-level mathematics measures (grades k-2), 2009-2010 version (technical report 1006). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Ardoin, S. P., Suldo, S. M., Witt, J., Aldrich, S., & McDonald, E. (2005). Accuracy of readability estimates' predictions of CBM performance. *School Psychology Quarterly, 20*, 1-22. doi: 10.1521/scpq.20.1.1.64193

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (Second ed.). New York: Routledge.

Brennan, R. L. (1992). Generalizability Theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34. doi: 10.1111/j.1745-3992.1992.tb00260.x

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer.

Messick, S. (1995). Validity of Psychological Assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi: 10.1037/0003-066X.50.9.741

Shavelson, R. J., & Webb, N. M. (1991). *Generalizabity Theory: A Primer*. Los Angeles, CA: Sage.

Snyder, S., & Sheehan, R. (1992). Research methods - the Rasch measurement model: An introduction. *Journal of Early Intervention, 166*, 87-95. doi: 10.1177/105381519201600108

Chapter 4: Letter Names

The technical evidence gathered for the easyCBM© Letter Names (LN) tests to date suggests the measure is functioning largely as intended. The developmental process used an advanced statistical process for scaling items (Rasch modeling), which provided information on the difficulty and functioning of each LN item (Alonzo & Tindal, 2007). Alonzo and Tindal (2009) and Wray, Lai, Saez, Alonzo, and Tindal (2014) found the alternate form reliability to be quite high, indicating students' scores were stable across test forms, while Alonzo and Tindal also found the test-retest reliability to be quite high, indicating stability across testing occasions. Lai, Alonzo, and Tindal (2013) found the LN measure to have a high relation with a criterion measure assessing the same skill, while Lai, Nese, Jamgochian, Alonzo, and Tindal (2010) used latent factor analyses to show that students' reading ability was a strong predictor of their LN scores, further increasing the validity evidence for the measures.

In what follows, we summarize the studies cited above examining the technical adequacy evidence for the easyCBM© LN measures. We begin with the developmental process used, including creating alternate test forms of equivalent difficulty. We then summarize reliability evidence, including alternate form and test-retest reliability evidence. Finally, we conclude by discussing validity evidence, including the relation between LN and a relevant criterion measures, along with the degree to which LN loads on a latent "Reading" factor. For a conceptual discussion of the theoretical purpose for any analysis conducted, please see Chapter 3. Note that this document is intended to provide a summary of evidence, not to explain in detail the different studies from which the evidence was gleaned. For a complete account of each study, the full reports are available at www.brtprojects.org/publications/technical-reports.

Test Development and Alignment

Item piloting was conducted with a sample of students located in the Pacific Northwest during the 2006-2007 school year. Data were collected on each letter, with a range of 297 to 1,036 student responses per letter. All items were scaled with a Rasch model. Linking items (letters) had more responses than others because they were repeated across all test forms and were responded to by more students during piloting. Full scaling results, including the difficulty and fit of each item to the model expectations, are reported by Alonzo and Tindal (2007). Results from the Rasch analyses helped ensure test forms had adequate range (from easy to difficult items) to sufficiently classify students into risk categories, along with an adequate number of items on the lower tail of the distribution to detect small changes for students being progress monitored over time. The Rasch analyses also allowed us to construct the forms so the average item difficulty was essentially equivalent across forms.

Reliability

Alternate form and test-retest reliability for the easyCBM© LN measures was investigated by Alonzo and Tindal (2009). Alternate form reliability was also investigated by Wray et al. (2014), with a sample of convenience from 222 students in the Pacific Northwest.

Alternate Form Reliability. Alonzo and Tindal (2009) administered LN forms 1, 3, and 5 to Grade 1 students on each of two separate occasions, spaced one week apart in the spring of 2008. The correlation between students' scores on each form ranged from .82 to .89, indicating a very strong relationship. In other words, students' scores were quite stable regardless of the specific test form administered. Wray et al. (2014) investigated the alternate form reliability for forms 8-17 in Kindergarten and Grade 1. For Kindergarten, correlations ranged from .61-.90, while Grade 1 correlations ranged from .66-.90. Correlations were reported for measures taken at

the same time point, and across time points collected approximately bi-monthly. Measures taken closer in time generally had higher correlations than those taken further apart. Alternate form reliability coefficients for measures taken on the same day ranged from .87 to .90 for Grade K and .85 to .90 for Grade 1.

Test-Retest Reliability. Alonzo and Tindal (2009) also investigated the test-retest reliability of forms 1, 3, and 5 in Grade 1. They found correlations ranging from .79 to .82, indicating a strong relation.

Validity Evidence

Criterion Validity

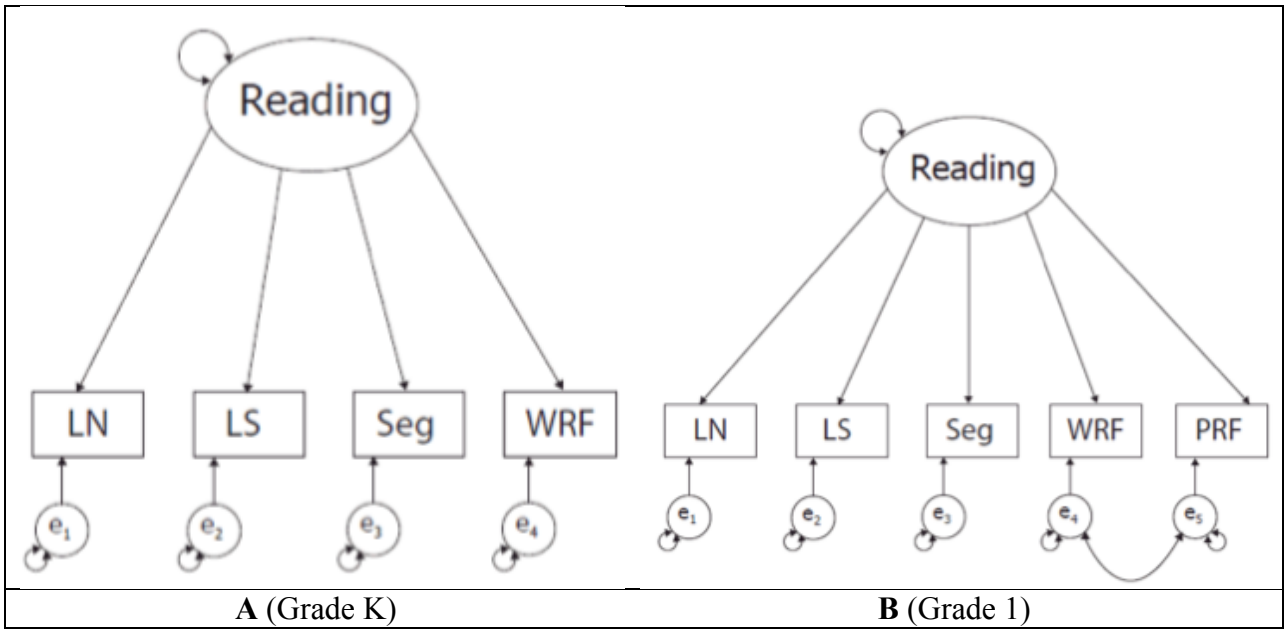
Lai et al. (2013) explored the relation between the easyCBM LN benchmarks and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Letter Naming Fluency (LNF) measure. The study included a large sample of Grade K and 1 students in the Pacific Northwest. The authors explored the relation between the measures with Spearman's rho, a nonparametric correlation (see Chapter 2). The authors found the relation between easyCBM© LN and DIBELS PSF to be quite high, at .86 for Grade K and .80 for Grade 1, indicating the measures are likely measuring the same underlying skills.

Wray et al. (2014) investigated the criterion validity of the easyCBM© early literacy measures by using multiple regression models. The letter names, letter sounds, and phoneme segmenting measures were used to predict performance on the Stanford Achievement Test, 10th edition (SAT-10). Across time points, the measures combined to account for 35-40% of the variance in Kindergarten SAT-10 Sounds and Letters performance, and 48-58% of the variance in SAT-10 Word Reading performance. For Grade 1, the measures combined to account for 14-32% of the variance in SAT-10 Word Study Skills performance, and 49-56% of the variance in

SAT-10 Word Reading Fluency performance. Perhaps unsurprisingly, the relations were strongest for measures that targeted constructs similar to those targeted by easyCBM©. See the full report for semi-partial correlations for the LN measure (i.e., the proportion of variance among the easyCBM© measures uniquely accounted for by LN).

Construct Validity

To gather construct validity evidence for the easyCBM© LN measures, Lai et al. (2010) conducted a series of confirmatory factor analyses. In Grade K, it was hypothesized that each of the early literacy measures (letter names, letter sounds, word reading fluency, and phoneme segmenting) each measured a portion of a unified, latent (i.e., unobservable), "Reading" trait (see Figure 4.1a). The same model was hypothesized in Grade 1, but with a passage reading fluency measure also included (see Figure 4.2b).



A (Grade K) **B** (Grade 1)

Figure 4.1. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM early literacy measure contributes to a single, latent (unobservable), "Reading" trait. LN = Letter Names, LS = Letter Sounds, Seg = Phoneme Segmenting, WRF = Word Reading Fluency, PRF = Passage Reading Fluency.

Lai et al. (2010) conducted a separate confirmatory factor analysis for each season within each grade. The data were then evaluated for their fit to the hypothesized, one-factor model. Overall, the model fit statistics suggested the data had fair to good fit to the model for all seasons across both grades (for a full description of the model fit statistics and model fit criteria, see Lai et al. 2010, pages 8-9 and results in Tables 33-36). The LN measures had consistent and strong "loadings" on the latent reading factor (.80s-.90s for grade Kindergarten and .90s for grade one), suggesting that students reading ability is a strong predictor of their performance on the easyCBM LN measure.

References

Alonzo, J., & Tindal, G. (2007). The development of early literacy measures for use in a progress monitoring assessment system: Letter names, letter sounds and phoneme segmenting (technical report 39). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009). Alternate form and test-tetest reliability of easyCBM reading measures (technical report 0906). Eugene, OR: Behavioral Research and Teaching, University of Oregon. .

Lai, C. F., Alonzo, J., & Tindal, G. (2013). easyCBM reading criterion related validity evidence: Grades k-1 (technical report 1309). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Nese, J. F. T., Jamgochian, E. M., Alonzo, J., & Tindal, G. (2010). Technical adequacy of the easyCBM primary-level reading measures (grades k-1), 2009-2010 version (technical report 1003). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Wray, K., Lai, C., F., Saez, L., Alonzo, J., & Tindal, G. (2014). *easyCBM Beginning Reading Measures: Grades K-1 Alternate Form Reliability and Criterion Validity With the SAT-10* (Technical Report No. 1403). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Chapter 5: Letter Sounds

The technical evidence gathered for the easyCBM© LS tests to date suggests the measure is functioning largely as intended. The developmental process used an advanced statistical process for scaling items (Rasch modeling), which provided information on the difficulty and functioning of each LS item (Alonzo & Tindal, 2007). Alonzo and Tindal (2009), Anderson, Park, Lai, Alonzo, and Tindal (2012), and Wray, Lai, Saez, Alonzo, and Tindal (2014) found the reliability of the measures to be quite high across test forms. Finally, Lai, Nese, Jamgochian, Alonzo, and Tindal (2010), Lai, Alonzo, and Tindal (2013) and Wray et al. (2014) found the measures to have modest to strong relations with external criteria.

In what follows, we summarize the studies cited above examining the technical adequacy evidence for the easyCBM© LS measures. We begin with the developmental process used, including creating alternate test forms of equivalent difficulty. We then summarize reliability evidence, including alternate form, test-retest, and generalizability theory analyses. Finally, we conclude by discussing validity evidence, including the relation between PS and relevant criterion measures, and the degree to which PS loads a latent "Reading" factor. For a conceptual discussion of the theoretical purpose for any analysis conducted, please see Chapter 3. Note that this document is intended to provide a summary of evidence, not to explain in detail the different studies from which the evidence was gleaned. For a complete account of each study, the full reports are available at www.brtprojects.org/publications/technical-reports.

Test Development and Alignment

Item piloting was conducted with a sample of students in the Pacific Northwest during the 2006-2007 school year. All items were scaled with a Rasch model (see Chapter 2). Data were collected on each letter (sound), with a range of 554 to 1,801 student responses per letter.

Linking items (letters) having relatively more responses than others because they were repeated across all test forms and were responded to by more students during piloting. After scaling piloted items, six letters (B, C, d, j, p, and Qu) were excluded from all LS test forms because they did fit the model expectations well. The difficulty and fit of each item in all LS test forms are reported by Alonzo and Tindal (2007). Results from our Rasch analyses helped us ensure that test forms had adequate range (from easy to difficult items) to sufficiently classify all students, along with an adequate number of items at the lower end of the distribution to detect small changes in performance of students whose letter sound naming skills are progress monitored over time. The Rasch analyses also allowed us to construct the forms so the average item difficulty was essentially equivalent across forms.

Reliability

Alternate form reliability for the easyCBM© LS measures was investigated by Alonzo and Tindal (2009), Anderson et al. (2012), and Wray et al. (2014), while Alonzo and Tindal and Anderson et al. also evaluated test-retest reliability. Generalizability and decision studies were also conducted by Anderson et al. to explore how various facets of the measurement process impacted the reliability of the LS measures.

Alternate Form Reliability. Alonzo and Tindal (2009) administered LS forms 1, 3, and 5 to Grade 1 students on each of two separate occasions, spaced one week apart in the spring of 2008. The correlation between students' scores on each form ranged from .76 to .88, indicating a strong relationship. In other words, students' scores were quite stable regardless of the specific test form administered. Anderson et al. (2012) investigated the alternate form reliability of the Grade 1 LS measures using a similar approach, but with test forms 11, 12, 13, 14, 15, and 16. Overall, the correlations between forms ranged from .82 to .89, indicating the relation among the

measures was similar to that found by Alonzo and Tindal. Wray et al. (2014) investigated alternate form reliability from tests taken on the same day, and across five time points collected approximately bi-monthly, for test forms 8-17 for Grade K and 8, 10, 12,15 and 17 for Grade 1. For Grade K, tests taken the same day, correlations ranged from .88 to .92. Across time points, correlations ranged from .53 to .92, with the correlations generally decreasing as the time between assessments increased. For Grade 1, correlations ranged from .49 to .83, although no forms were collected on the same day. Across studies, the correlations suggest students' scores are quite stable across forms. Please see the full reports for complete correlation tables for each study.

Test-Retest Reliability. Test-retest reliability was also investigated by Alonzo and Tindal (2009) and Anderson et al. (2012) with the same forms used to investigate alternate form reliability (administrations were spaced one week apart). Alonzo and Tindal found the test-retest correlations to range from .64 to .68, indicating a moderately strong relation. Anderson et al. found the test-retest correlations had a median ranged from .77 to .87, also indicating a strong relation.

Generalizability Study. In 2012, Anderson et al. conducted a generalizability study primarily to examine how specific facets of the measurement process related to the reliability. Across test forms, Anderson et al. found the G-coefficients ranged from .87 to .95.

Validity

Criterion Validity

Criterion validity evidence for the easyCBM letter sounds (LS) measures stems primarily from two studies. Lai et al. (2010) evaluated the relation between the easyCBM LS benchmarks and the Stanford-10 (Stanford-10), while Lai et al. (2013) explored the relation between the

easyCBM LS benchmarks and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Initial Sound Fluency (ISF) measure in Grade K and the relation between the easyCBM LS benchmarks and the DIBELS Nonsense Word Fluency (NWF) measure in Grade 1. The Lai et al. (2010) study included a sample of approximately 2,000 Grade K students in the Pacific Northwest, while Lai et al. (2013) used a sample of just over 200 students. Lai et al. 2010 explored both the predictive and concurrent validity of LS, while Lai et al. 2013 explored only concurrent validity.

Predictive validity. Lai et al. (2010) used the fall and winter LS measures in Grade K and Grade 1 to predict students' spring performance on the Stanford-10. All models were run with other easyCBM© early reading measures, including phoneme segmenting and word reading fluency. In Grade K, the fall measure *uniquely* accounted for approximately 10% of the variance in Stanford -10 reading scores, while the winter measure uniquely accounted for 3%. In Grade 1, the fall measure uniquely accounted for approximately 3% of the variance in Stanford-10 reading scores, while the winter measure uniquely accounted for 7% of the variance in the Stanford -10 reading scores.

Wray et al. (2014) investigated the criterion validity of the easyCBM© early literacy measures by using multiple regression models. The letter names, letter sounds, and phoneme segmenting measures were used to predict performance on the Stanford Achievement Test, 10th edition (SAT-10). Across time points, the measures combined to account for 35-40% of the variance in Kindergarten SAT-10 Sounds and Letters performance, and 48-58% of the variance in SAT-10 Word Reading performance. For Grade 1, the measures combined to account for 14-32% of the variance in SAT-10 Word Study Skills performance, and 49-56% of the variance in SAT-10 Word Reading Fluency performance. Perhaps unsurprisingly, the relations were

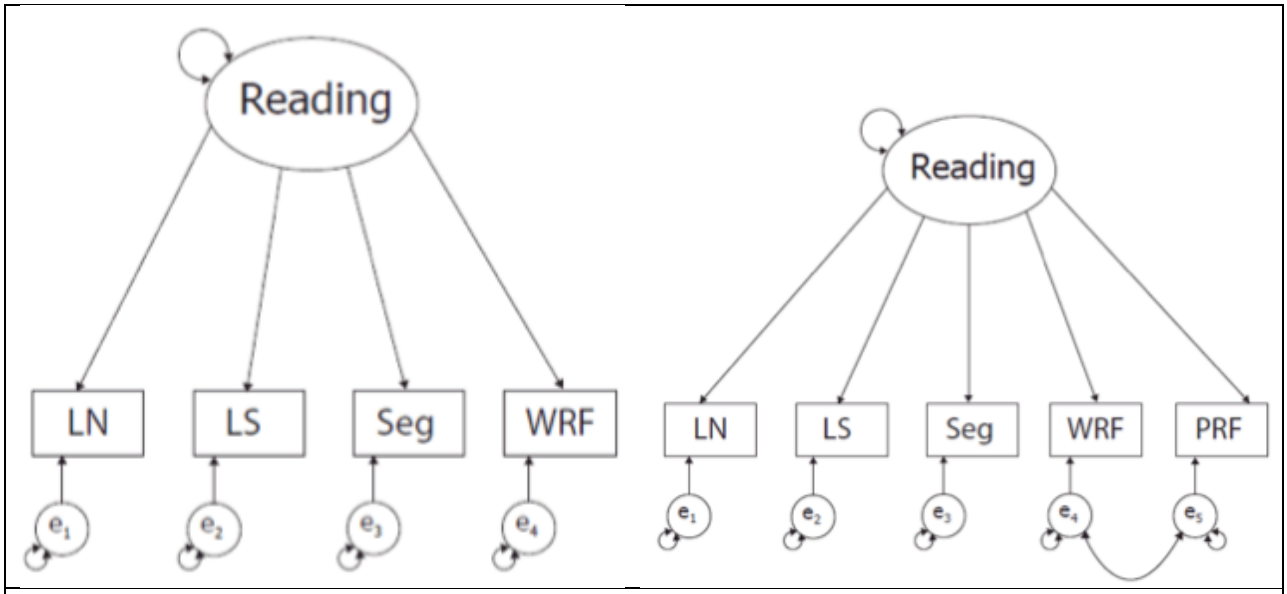
strongest for measures that targeted constructs similar to those targeted by easyCBM©. See the full report for semi-partial correlations for the LS measure (i.e., the proportion of variance among the easyCBM© measures uniquely accounted for by LS).

Concurrent validity. Both Lai et al. (2010) and Lai et al. (2013) investigated the concurrent validity (both tests administered at the same time) of the easyCBM LS measures in Grade K. Lai et al. (2010) found that the spring measure accounted for approximately 10% of the variance in students' overall Stanford-10 reading scores in Grade K. Concurrent validity information was not gathered for Grade 1.

Lai et al. 2013 explored the relation between the easyCBM© LS measures and the criterion measures using a "nonparametric" correlation statistic, Spearman's Rho. Overall, the nonparametric correlation between easyCBM© LS and DIBELS ISF at Grade K was .55, a moderate association. The correlation between easyCBM© LS and DIBELS NWF was also moderate, at .58 for Grade 1. These results provide evidence that the LS measures are assessing similar constructs, yet may reflect some differences in test format.

Construct Validity

To gather construct validity evidence for the easyCBM© PS measures, Lai et al. (2010) conducted a series of confirmatory factor analyses. In Grade K, it was hypothesized that each of the early literacy measures (letter names, letter sounds, word reading fluency, and phoneme segmenting) each measured a portion of a unified, latent (i.e., unobservable), "Reading" trait (see Figure 5.1a). The same model was hypothesized in Grade 1, but with a passage reading fluency measure also included (see Figure 5.1b).



A (Grade K) **B** (Grade 1)

Figure 5.1. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM early literacy measure contributes to a single, latent (unobservable), "Reading" trait. LN = Letter Names, LS = Letter Sounds, Seg = Phoneme Segmenting, WRF = Word Reading Fluency, PRF = Passage Reading Fluency.

Lai et al. (2010) used a selected set of model fit statistics to assess the hypothesized factor structure. The recommended criteria used to judge model fit statistics suggested fair to good model fit for the hypothesized 1-factor model for all seasons across both grades (see for full description of the model fit statistics and model fit criteria in pages 8-9 and results in Tables 33-36 in Technical Report #1003), providing construct validity evidence for the Grades K and 1 LS measures as part of the easyCBM© early literacy measures.

References

Alonzo, J., & Tindal, G. (2007). The development of early literacy measures for use in a progress monitoring assessment system: Letter names, letter sounds and phoneme segmenting (technical report 39). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009). Alternate form and test-tetest reliability of easyCBM reading measures (technical report 0906). Eugene, OR: Behavioral Research and Teaching, University of Oregon. .

Anderson, D., Park, B. J., Lai, C., F., Alonzo, J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 1 (technical report 2016). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2013). easyCBM reading criterion related validity evidence: Grades k-1 (technical report 1309). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Nese, J. F. T., Jamgochian, E. M., Alonzo, J., & Tindal, G. (2010). Technical adequacy of the easyCBM primary-level reading measures (grades k-1), 2009-2010 version (technical report 1003). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Wray, K., Lai, C., F., Saez, L., Alonzo, J., & Tindal, G. (2014). *easyCBM Beginning Reading Measures: Grades K-1 Alternate Form Reliability and Criterion Validity With the SAT-10* (Technical Report No. 1403). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Chapter 6: Phoneme Segmenting

The technical adequacy evidence gathered for the easyCBM© Phoneme Segmenting (PS) measures to date suggests they are functioning largely as intended. An advanced statistical technique for scaling items (Rasch modeling) was used during the developmental process, providing information on the difficulty and functioning of each PS item. These results then guided test form creation to help ensure all 20 test forms were essentially equivalent in terms of difficulty and the distribution of easy and difficult items within each test form. Sáez, Irvin, Alonzo, and Tindal (2012) showed that PS items were reasonably well aligned with the CCSS, with misaligned items generally being more rigorous than called for by the corresponding standard. Alternate-form, test-retest, and generalizability theory studies conducted by Alonzo and Tindal (2009) and Anderson, Park, Lai, Alonzo, and Tindal (2012) suggested the measures were generally quite consistent both within and across test forms. Alternate form reliability was also shown to be quite high by Wray, Lai, Saez, Alonzo, and Tindal (2014). Finally, Lai, Nese, Jamgochian, Alonzo, and Tindal (2010), Lai, Alonzo, and Tindal (2013), and Wray et al. (2014) showed that easyCBM© PS measures had a strong relation with various criterion measures, both predictively and concurrently, and that students' reading ability moderately predicted students PS performance. These studies suggest the easyCBM© PS measures have a high degree of validity for measuring students' phonemic segmentation skills within a response to intervention framework. In sum, these studies suggest the technical adequacy evidence for the easyCBM© PS measures is quite strong.

In what follows, we summarize the studies cited above examining the technical adequacy evidence for the easyCBM© PS measures. We begin with the developmental process used, including creating alternate test forms of equivalent difficulty, as well as the alignment between

the measures and the Common Core State Standards (CCSS). We then summarize reliability evidence, including alternate form, test-retest, and generalizability theory analyses. Finally, we conclude by discussing validity evidence, including the relation between PS and relevant criterion measures, and the degree to which PS loads a latent "Reading" factor. For a conceptual discussion of the theoretical purpose for any analysis conducted, please see Chapter 3. Note that this document is intended to provide a summary of evidence, not to explain in detail the different studies from which the evidence was gleaned. For a complete account of each study, the full reports are available at www.brtprojects.org/publications/technical-reports.

Measure Development

Item piloting was conducted with a sample of students located in the Pacific Northwest during the 2006-2007 school year. Data were collected on each item (phoneme), with a range of 110 to 2,067 student responses per word, with linking items having significantly more responses because they were repeated across all forms in the calibration study. All items were scaled with a Rasch model. The difficulty and model fit of each item in all PS test forms are reported by Alonzo and Tindal (2007). Results from the Rasch analyses helped ensure test forms had adequate range (from easy to difficult items) to sufficiently classify students into risk categories, along with an adequate number of items on the lower tail of the distribution to detect small changes for students being progress monitored over time. The Rasch analyses also allowed us to construct the forms so the average item difficulty was essentially equivalent across forms.

Alignment to Standards

We focused our alignment analyses for PS on a portion of Foundational Skills Standard Two in reading for Grades K and 1 because they specifically spell out expectations around student phoneme segmentation skills. Figure 6.1 displays the particular CCSS to which

alignment was evaluated. Our alignment study was unique in that we asked a group of carefully screened Grade K and 1 teachers to indicate the number of syllables and to segment each word of interest into its constituent sound parts (skill requirements identified by the targeted CCSS), so that we could evaluate both teacher agreement *and* the teacher's phonemic background knowledge. We used the accuracy of teacher responses to further evaluate the relationship between tested words and instructional practices. We reasoned that words deemed problematic for teachers might, in fact, represent gaps in teacher knowledge, and that such words could potentially be measuring something other than students' phonemic awareness skills. Words identified as problematic would need further examination.

Reading: Foundational Skills Standard 2

RF.K.2. Demonstrate understanding of spoken words, syllables, and sounds (phonemes).

A. Recognize and produce rhyming words.
B. Count, pronounce, blend, and segment syllables in spoken words.
C. Blend and segment onsets and rimes of single-syllable spoken words.
D. **Isolate and pronounce the initial, medial vowel, and final sounds (phonemes) in three-phoneme (consonant-vowel-consonant, or CVC) words.[1] (This does not include CVCs ending with /l/, /r/, or /x/.)**
E. Add or substitute individual sounds (phonemes) in simple, one-syllable words to make new words.

RF.1.2. Demonstrate understanding of spoken words, syllables, and sounds (phonemes).

A. Distinguish long from short vowel sounds in spoken single-syllable words.
B. Orally produce single-syllable words by blending sounds (phonemes), including consonant blends.
C. **Isolate and pronounce initial, medial vowel, and final sounds (phonemes) in spoken single-syllable words.**
D. **Segment spoken single-syllable words into their complete sequence of individual sounds (phonemes).**

Figure 6.1. CCSS Foundational Skills Standard Two in Reading for grade K (RF.K.2 D) and grade 1 (RF.1.2 C and D)

Teachers evaluated 98% of the unique words from the Grade K PS measures (169 words), and 99% of the unique words from the Grade 1 measures (158 words). Three teachers reviewed each word.

With respect to Grade K, we found weak evidence of alignment to standard RF.K.2 D. In general, the level of phoneme segmentation required by the measures exceeded that of the standard RF.K.2 D because nearly half of the words reviewed were comprised of more than three phonemes. Only 65 words (~49%) contained three phonemes, and of these only seven words were identified as consonant-vowel-consonant (CVC) as required by the targeted Grade K standard. The three-phoneme CVC words comprised less than 5% of the reviewed words.

Regarding Grade 1, we again found weak evidence of alignment to standard RF.1.2 C. While 127 (~80%) of the Grade 1 words reviewed were single-syllable, students' ability to isolate the initial, medial vowel, or final sounds of such words can only be inferred indirectly through teacher evaluation of student errors. On the other hand, evidence of alignment to standard RF.1.2 D appeared strong. As stated above, almost 80% of reviewed words were single-syllable. All test items on the Phoneme Segmenting measure require students to segment these words into their complete sequence of individual sounds or phonemes, with three tested words containing two phonemes (2.2%), 62 containing three phonemes (46.3%), and 69 containing four or five phonemes (51.5%).

A total of 31 words (~18%) were identified as problematic by the teacher panel due to: (a) non-representativeness of typical Grade K and/or 1 vocabularies, (b) extreme segmenting difficulty, and/or (c) concerns around pronunciation. The inclusion of extremely difficult words to segment (or pronounce) that might also not be representative of typical Grade K or 1 vocabulary and/or phonemic instruction might impact students' ability to give a correct response, and would suggest that these words may not be measuring the phonemic segmenting skills intended by the measure (or target standards).

Reliability

Alternate form reliability analyses were conducted by Wray et al. (2014), Alonzo and Tindal (2009), and Anderson et al. (2012). Alonzo and Tindal and Anderson et al. also conducted test-retest reliability analyses. Finally, generalizability and decision studies were conducted by Anderson et al. to explore how various facets of the measurement process impacted the reliability of the PS measures.

Alternate Form Reliability. Alonzo and Tindal (2009) administered PS forms 1, 3, and 5 to Grade 1 students on each of two separate occasions, spaced one week apart in the spring of 2008. The correlation between students' scores on each form ranged from .86 to .91, indicating a very strong relation. In other words, students' scores were quite stable regardless of the specific test form administered. Anderson et al. (2012) investigated the alternate form reliability of the Grade 1 PS measures using a similar approach, but with test forms 11, 12, 13, 14, 15, and 16. Overall, the correlations between forms ranged from .62 to .89, indicating the relation was slightly lower than that found by Alonzo and Tindal.

Wray et al. (2014) investigated alternate form reliability from tests taken on the same day, and across five time points collected approximately bi-monthly, for test forms 5-8, and 10-15. For tests taken the same day, correlations ranged from .81 to .90. Across time points, correlations ranged from .31 to .90, with the correlations generally decreasing as the time between assessments increased. Across studies, the correlations suggest students' scores are quite stable across forms. Please see the full reports for complete correlation tables for each study.

Test-Retest Reliability. Test-retest reliability was also investigated by Alonzo and Tindal (2009) and Anderson et al. (2012) with the same forms used to investigate alternate form reliability. However, rather than evaluating the correlation between students' scores across test forms, test-retest reliability is evaluated by examining the correlation between students' scores on the same test form across testing occasions, which were spaced one week apart. Theoretically students' skills should not change dramatically within a one-week time frame. Phoneme segmenting, however, is a skill that changes rapidly and it is possible that students' "true" segmenting ability could change marginally between testing occasions. We would therefore expect the test-retest correlations to be high, but perhaps not as high as the alternate form reliability correlations. Alonzo and Tindal found the test-retest correlations to range from .45 to .47, indicating a modest relation. Anderson et al. found the test-retest correlations ranged from .32 to .81, with a median of .57, indicating a low to strong relation. It is also important to keep in mind that the sample size for these analyses was quite small, ranging from 19-42.

Generalizability Study. In 2012, Anderson et al. conducted a generalizability study primarily to examine how specific facets of the measurement process related to the reliability of the measures. Across test forms, Anderson et al. found the G-coefficients ranged from .50 to .83. It is again important to note that the sample size was quite small, ranging from 19-42. For the analysis reporting the smallest G-coefficient, a non-trivial person by occasion interaction was found, while very little variance was uniquely attributable to the form itself. It is therefore possible that some event (e.g., birthday party, unusual weather, etc.) occurred during one of the testing occasions that differentially impacted students. In other words, some students may have responded to the event by behaving radically different, while other students were unaffected.

Validity Evidence

Criterion Validity

The bulk of the criterion validity evidence for the easyCBM© PS measures stems from two studies. Lai et al. (2010) evaluated the relation between the easyCBM© PS benchmarks and the Stanford-10, while Lai et al. (2013) explored the relation between the easyCBM© PS benchmarks and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) Phoneme Segmentation Fluency (PSF) measure and the Comprehensive Test of Phonological Processing (CTOPP) Elision measure. The Lai et al. (2010) study included a sample of approximately 2,000 Grade K students in the Pacific Northwest, while Lai et al. (2013) used a sample of just over 200 students. Lai et al. (2010) explored both the predictive and concurrent validity of PS, while Lai et al. (2013) explored only concurrent validity.

Predictive validity. Lai et al. (2010) used the fall and winter PS measures in both Grades K and 1 to predict students' spring performance on the Stanford-10. In Grade K, the fall and winter measures each accounted for approximately 10% of the variance in Stanford-10. Lai et al. also explored the utility of students' growth across the benchmark measures in predicting students' Stanford-10. Although students' growth was a much more modest predictor of their spring Stanford-10 achievement than their raw PS score, it generally remained a significant predictor, with higher positive slopes generally leading to higher Stanford-10 performance.

Wray et al. (2014) investigated the criterion validity of the easyCBM© early literacy measures by using multiple regression models. The letter names, letter sounds, and phoneme segmenting measures were used to predict performance on the Stanford Achievement Test, 10th edition (SAT-10). Across time points, the measures combined to account for 35-40% of the variance in Kindergarten SAT-10 Sounds and Letters performance, and 48-58% of the variance

in SAT-10 Word Reading performance. For Grade 1, the measures combined to account for 14-32% of the variance in SAT-10 Word Study Skills performance, and 49-56% of the variance in SAT-10 Word Reading Fluency performance. Perhaps unsurprisingly, the relations were strongest for measures that targeted constructs similar to those targeted by easyCBM©. See the full report for semi-partial correlations for the PS measure (i.e., the proportion of variance among the easyCBM© measures uniquely accounted for by PS).

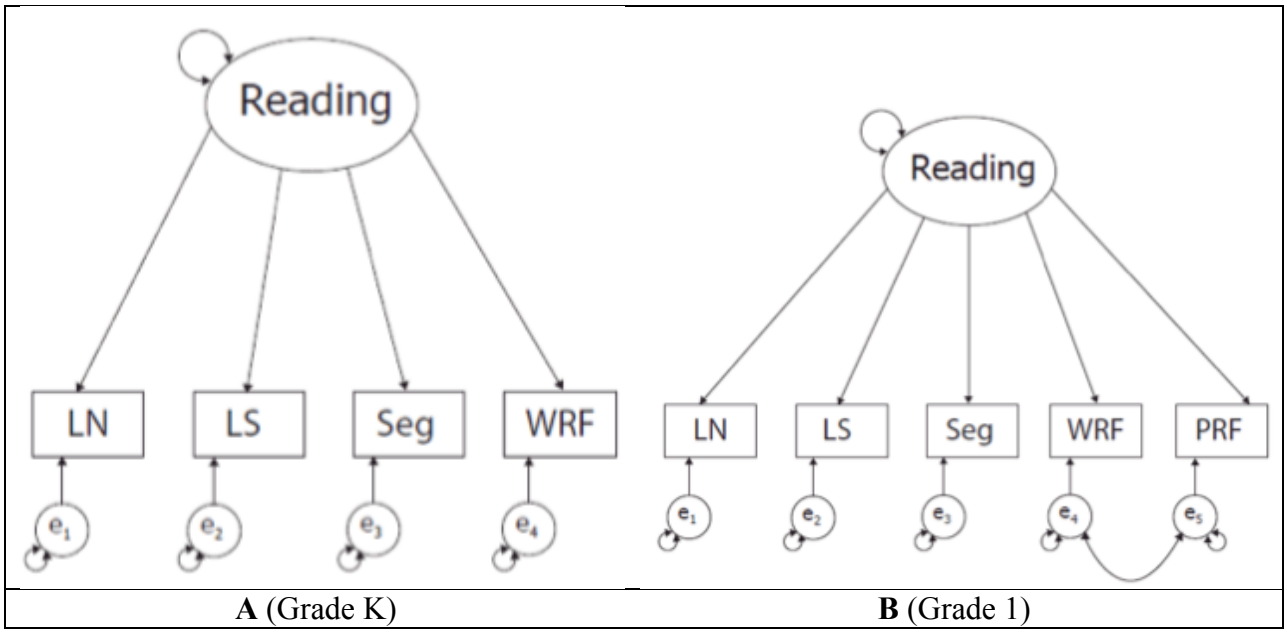
Concurrent validity. Both Lai et al. (2010) and Lai et al. (2013) investigated the concurrent validity (both tests administered at the same time) of the easyCBM© PS measures in Grade K. Lai et al. (2010) found that the spring measure accounted for a very small percentage of the variance in students' Stanford-10 scores in Grade K. These values were not unexpected. Segmenting is a skill that "tops out" relatively early. *A priori*, we did not expect the relation in spring to be as strong as in fall or winter, given that as students progress in developing reading skills they naturally stop focusing on discrete sounds associated with individual letters and combinations of letters within words and instead begin reading fluently, particularly with sight words.

Lai et al. 2013 explored the relation between the easyCBM© PS measures and the criterion measures using a "nonparametric" correlation statistic, Spearman's Rho. The correlation between easyCBM© PS and DIBELS PSF was quite high, at .85 for Grade K and .75 for Grade 1. When easyCBM© PS was compared to the CTOPP Elision subtest, the correlation was somewhat low for Grade K, at .39, and very low for Grade 1, at .05. The relation between the measures at Grade 1 was not statistically significant, indicating the easyCBM© PS measure did not provide information different from chance about how students would perform on the CTOPP Elision subtest. The high results between the easyCBM® PS measure and the DIBELS

PSF measure is likely related to the consistency between the measures in terms of content, formatting, and test administration procedures. The low correlation results for the CTOPP Elision subtest comparisons may be due to the fact that only one subtest of the CTOPP was included in the study due to logistical constraints.

Construct Validity

To gather construct validity evidence for the easyCBM© PS measures, Lai et al. (2010) conducted a series of confirmatory factor analyses. In Grade K, it was hypothesized that each of the early literacy measures (letter names, letter sounds, word reading fluency, and phoneme segmenting) each measured a portion of a unified, latent (i.e., unobservable), "Reading" trait (see Figure 6.1a). The same model was hypothesized in Grade 1, but with a passage reading fluency measure also included (see Figure 6.1b).



A (Grade K) **B** (Grade 1)

Figure 6.1. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM early literacy measure contributes to a single, latent (unobservable), "Reading" trait. LN = Letter Names, LS = Letter Sounds, Seg = Phoneme Segmenting, WRF = Word Reading Fluency, PRF = Passage Reading Fluency.

Lai et al. (2010) conducted a separate confirmatory factor analysis for each season within

each grade. The data were then evaluated for their fit to the hypothesized, one-factor model. Overall, the model fit statistics suggested the data had fair to good fit to the model for all seasons across both grades (for a full description of the model fit statistics and model fit criteria, see Lai et al. 2010, pages 8-9 and results in Tables 33-36). The PS measures had consistent and low to moderate "loadings" on the latent reading factor (.50s for grade Kindergarten and .30-.40 for grade one), suggesting that students reading ability moderately predicted their performance on the easyCBM PS measure.

References

Alonzo, J., & Tindal, G. (2007). The development of early literacy measures for use in a progress monitoring assessment system: Letter names, letter sounds and phoneme segmenting (technical report 39). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009). Alternate form and test-tetest reliability of easyCBM reading measures (technical report 0906). Eugene, OR: Behavioral Research and Teaching, University of Oregon. .

Anderson, D., Park, B. J., Lai, C., F., Alonzo, J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 1 (technical report 2016). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2013). easyCBM reading criterion related validity evidence: Grades k-1 (technical report 1309). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Nese, J. F. T., Jamgochian, E. M., Alonzo, J., & Tindal, G. (2010). Technical adequacy of the easyCBM primary-level reading measures (grades k-1), 2009-2010 version (technical report 1003). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Sáez, L., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). Phoneme segmenting alignment with the Common Core Foundational Skills Standard Two: Grades K-1 (Technical Report No. 1227). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Wray, K., Lai, C., F., Saez, L., Alonzo, J., & Tindal, G. (2014). *easyCBM Beginning Reading Measures: Grades K-1 Alternate Form Reliability and Criterion Validity With the SAT-10*

(Technical Report No. 1403). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Chapter 7: Word Reading Fluency

The technical adequacy evidence gathered for the easyCBM© Word Reading Fluency (WRF) measures to date suggests they are functioning as intended. An advanced statistical technique for scaling items (Rasch modeling) was used during the developmental process, providing information on the difficulty and functioning of each word in each WRF measure (Alonzo & Tindal, 2007). These results then guided test form creation to help ensure all 20 test forms at each grade were essentially equivalent in terms of difficulty and the distribution of easy and difficult items within each test form. The reliability of the measures has been shown by multiple measures to be quite high (Alonzo & Tindal, 2009; Anderson, Lai, Park, Alonzo, & Tindal, 2012; Anderson, Park, Lai, Alonzo, & Tindal, 2012; Park, Anderson, Alonzo, Lai, & Tindal, 2012). Studies investigating the validity of WRF were generally conducted as part of larger studies of the different measures available at Grade 3, where instruction has moved from away from word reading fluency and shifted to other areas (e.g., passage reading, vocabulary, comprehension), and the validity evidence from these studies was, predictably, weak to moderate (Anderson, Alonzo, & Tindal, 2011; Sáez et al., 2010). The exception was (Jamgochian et al., 2010), who did investigate the validity of the K-2 measures, and generally found good validity evidence. In sum, these studies suggest the technical adequacy evidence for the easyCBM© WRF measures is quite strong.

In what follows, we summarize the studies cited above examining the technical adequacy evidence for the easyCBM© WRF measures. We begin with the developmental process used, including creating alternate test forms of equivalent difficulty, as well as the alignment between the measures and the Common Core State Standards (CCSS). We then summarize reliability evidence, including alternate form, test-retest, and generalizability theory analyses. We conclude

by discussing validity evidence, including the relation between WRF and relevant criterion measures, and the degree to which WRF loads a latent "Reading" factor. For a conceptual discussion of the theoretical purpose for any analysis conducted, please see Chapter 3. Note that this document is intended to provide a summary of evidence, not to explain in detail the different studies from which the evidence was gleaned. For a complete account of each study, the full reports are available at www.brtprojects.org/publications/technical-reports.

Measurement Development

Initial item piloting used in the development of the WRF measures was conducted with a sample of students located in the Pacific Northwest during the 2006-2007 school year. Piloted items (words) came from a variety of cross-examined sources including, though not limited to, the Dolch word lists, a list of 'the first 1000 words' found in Chall and Dale's (1995) book of lists, and online grade-level word lists. Words with regular and irregular sound patterns and of a variety of lengths were included. Data were collected on each word, with a range of 144 to 2,654 student responses per word (linking items had significantly more responses because they were repeated across all forms in the calibration study). All items were scaled with a Rasch model (see Chapter 2 for a conceptual overview of this methodology). The difficulty of each item in all WRF test forms are reported by Alonzo and Tindal (2007).

Results from the Rasch analyses helped ensure test forms had adequate range (from easy to difficult items) to sufficiently classify students into risk categories, along with an adequate number of items on the lower tail of the distribution to detect small changes in students' WRF skills over time. The Rasch analyses also allowed us to construct the forms so the average item difficulty was essentially equivalent across forms. Specifically, we used the results of Rasch analyses to draw from the easiest bands of words to populate the first rows of each test form with

subsequent rows of words increasing in difficulty. We created alternate forms within each grade level by selecting words of comparable difficulty across test forms as a whole and also across individual rows between alternate forms.

Alignment to Standards

We focused our alignment study on specific sub-skills comprising Foundational Skills Standard Three, *Phonics and Word Recognition,* because this section of the CCSS outlined skill expectations consistent with the measurement goals of the WRF measure. Figure 7.1 shows the Grade K-3 CCSS sub-skill standards to which alignment was gauged in bold (RF.K.3 B and C – Grade K; RF.1.3 A, B, C, E, F and G – Grade 1; RF.2.3 A, B, C, D, and F – Grade 2; RF.3.3 B, C and D – Grade 3). To begin, we asked a group of trained K-3 teacher reviewers to provide us with information related to each targeted CCSS sub-skill. We reasoned that this method allowed us to evaluate reviewer agreement *and* the word reading knowledge of the participating teacher reviewers with respect to targeted Standard 3 sub-skills. We went beyond typical alignment methods (e.g., Webb, 1999) by using the accuracy of teacher reviewer responses to evaluate the relation between tested words and content on which teachers theoretically should have been instructing students (e.g., digraphs). We reasoned that words deemed problematic for teacher reviewers might represent gaps in teacher knowledge, and that such words could potentially be measuring something other than students' word reading fluency skills. Words deemed problematic for teachers would need further examination.

In total, 15 teacher reviewers (three Kindergarten, one first grade, four second grade, and four third grade) took part in the alignment study. Three teachers reviewed each word using a secure web-based alignment tool, with teacher experience and placement (general and special education) balanced to the extent possible. To simplify analyses and avoid redundancy, words

analyzed were drawn only from the first test form at each grade (K-3), and these words were used as a representative sample of words from all WRF measures. Words were collapsed into two grade bands (K-1 and 2-3) to create separate item review pools to which teacher reviewers were assigned. After removing common words (e.g., *I*), the Grade K-1 item pool included 57 Grade K words and 117 Grade 1 words. The Grade 2-3 item review pool included 172 Grade 2 words and 171 Grade 3 words.

RF.K.3. Know and apply grade-level phonics and word analysis skills in decoding words.

A. Demonstrate basic knowledge of letter-sound correspondences by producing the primary or most frequent sound for each consonant.
B. **Associate the long and short sounds with the common spellings (graphemes) for the five major vowels.**
C. **Read common high-frequency words** by sight (e.g., *the, of, to, you, she, my, is, are, do, does*).
D. Distinguish between similarly spelled words by identifying the sounds of the letters that differ.

RF.1.3. Know and apply grade-level phonics and word analysis skills in decoding words.

A. **Know the spelling-sound correspondences for common consonant digraphs** (two letters that represent one sound).
B. **Decode regularly spelled one-syllable words.**
C. **Know final -e and common vowel team conventions for representing long vowel sounds.**
D. Use knowledge that every syllable must have a vowel sound to determine the number of syllables in a printed word.
E. **Decode two-syllable words** following basic patterns by breaking the words into syllables.
F. **Read words with inflectional endings.**
G. **Recognize and read grade-appropriate irregularly spelled words.**

RF.2.3. Know and apply grade-level phonics and word analysis skills in decoding words.

A. **Distinguish long and short vowels when reading regularly spelled one-syllable words.**
B. **Know spelling-sound correspondences for additional common vowel teams.**
C. **Decode regularly spelled two-syllable words with long vowels.**
D. **Decode words with common prefixes and suffixes.**
E. Identify words with inconsistent but common spelling-sound correspondences.
F. **Recognize and read grade-appropriate irregularly spelled words.**

RF.3.3. Know and apply grade-level phonics and word analysis skills in decoding words.

A. Identify and know the meaning of the most common prefixes and derivational suffixes.
B. **Decode words with common Latin suffixes.**
C. **Decode multi-syllable words.**
D. **Read grade-appropriate irregularly spelled words**

Figure 7.1 CCSS Foundational Skills Standard Three in Reading for Grade K (RF.K.3 B and C), Grade 1 (RF.1.3 A, B, C, E, F and G), Grade 2 (RF.2.3 A, B, C, D, and F) and Grade 3 (RF.3.3 B, C and D) targeted by Sáez et al (2013).

With respect to Grade K, the WRF measure appeared closely aligned to standard RF.K.3.B, with 34 words (nearly 60%) containing long- or short-vowel sounds using one of the five major vowels (*a*, *e*, *i*, *o* and *u*). The remaining 40% of Grade K words contained vowel sounds deemed beyond those called for by RF.K.3.B. The WRF measure was also closely aligned with standard RF.K.3.C, with 32 words (56%) found among the 500 most common words in print (Pinnell & Fountas, 1998).

Overall, the Grade 1 WRF measure was most closely aligned to the standard R.F.1.3.B, which addresses students reading regularly spelled one-syllable words. Nearly 60% of words were monosyllabic with regular spellings, while 24% were identified as two-syllable words with regular spellings (R.F.1.3.D), both with a high level of reviewer agreement (≥ 98%). The WRF measure aligned to a lesser extent with the remaining Grade 1 CCSS sub-skills we targeted (Figure 7.1), with reviewer agreement less consistent (range = 55.6% – 78.6%). Teacher agreement was found to be near chance for standard RF.1.3.A, with common consonant digraphs correctly identified only 55.6% of the time, suggesting potential confusion about this standard in the sample of teachers who participated in the alignment study.

For Grade 2, the WRF measure was most closely aligned to the standards R.F.2.3.B and C, addressing long vowel sounds. Overall, 40% of the Grade 2 words analyzed were comprised of a long vowel sound composed of a common vowel team (e.g., *ai*, *ay*), while 56% were comprised of two-syllable words with a single vowel, vowel team, *y*-ending and/or a final-*e*. Reviewer agreement was moderate for these two CCSS sub-skills, 69.9% and 66.7%, respectively. In addition, 10% of the words were deemed aligned to standard RF.2.3F, with17 words identified as having irregular spellings.

The alignment of the Grade 3 WRF measure to the targeted CCSS sub-skill standards

appeared, overall, moderately weak. The strongest alignment was found with respect to standard RF.3.3C, with 26 words identified as multi-syllabic (reviewer agreement = 98.7%). However, only 12 words (7.8%) contained a Latin suffix (e.g., *-ity*, *-ion*), while 21 words (14%) were irregularly spelled, with 50% and 80.1% reviewer agreement, respectively. It is worth noting that by Grade 3, most students have progressed beyond reading words in isolation to reading connected text. Thus, the Grade 3 WRF measures are intended to be used for monitoring the progress of students who are well behind grade-level expectations for developing literacy. Given this design feature and intended use, the moderately weak alignment of the Grade 3 WRF measure to the CCSS sub-skill standards is not surprising.

Reliability

Alternate form and test-retest reliability for the easyCBM© WRF measures was investigated by Alonzo and Tindal (2009), Park et al. (2012), Anderson, Lai et al. (2012), and Anderson, Park et al. (2012). Generalizability and decision studies were also conducted to explore how various facets of the measurement process impact the reliability of the WRF measures (Anderson, Lai et al., 2012; Anderson, Park et al., 2012; Park et al., 2012).

Alternate Form Reliability. Alonzo and Tindal (2009) administered WRF forms 1, 3, and 5 to Grade 1 and 3 students on each of two separate occasions, spaced one week apart in the spring of 2008. The correlation between students' scores on each form ranged from .95-.96 for Grade 1 and .87-.93 for Grade 3, indicating a very strong relation. In other words, students' scores were quite stable regardless of the specific test form administered. Anderson, Park et al. (2012), Anderson, Lai et al. (2012), and Park et al. (2012) investigated the alternate form reliability of the WRF measures in Grades 1, 2, and 3 respectively, using a similar approach, but with different test forms. Forms 11, 12, 14, and 15 were used for Grade 1, and forms 11-16 were

used Grades 2 and 3 in these studies. Overall, the correlations between forms ranged from .89 to .97 for Grade 1, .92-.95 for Grade 2, and .72-.92 for Grade 3, indicating a very strong relation among the measures, with one exception in Grade 3. Complete correlation tables are reported in the full reports.

Test-Retest Reliability. Test-retest reliability was also investigated with the same forms used to investigate alternate form reliability in each study cited above. Alonzo and Tindal (2009) reported test-retest correlations to ranging from .94 to .95 for Grade 1 and .92-.94 for Grade 3, indicating a very strong relation. Anderson Lai, et al. (2012) and Anderson Park, et al. (2012) largely replicated these results finding test-retest correlations ranging from .87 to .95, with medians of .92 for Grade 1 and .93 for Grade 2. Park et al. found correlations slightly lower for Grade 3, ranging from .67- .92, with a median of .78 for Grade 3. All results thus indicated alternate forms had a strong relation with each other. In interpreting these results, one should note that the sample size for these analyses was quite small, ranging from 17-62, thereby increasing the potential for a small number of participants' responses impacting the results substantially

Generalizability Study. Anderson, Park et al. (2012), Anderson, Lai et al. (2012), and Park et al. (2012) conducted Generalizability Theory studies to examine how specific facets of the measurement process related to the reliability of the measures. Across test forms, Anderson, Park et al. found G-coefficients ranging from .96-.98 for Grade 1, while Anderson, Lai et al. found a range of .96-.99 for Grade 2. Park et al. found G-coefficients ranged from.74-.95 for Grade 3. It is again important to note that the sample size was quite small, ranging from 17-62 students.

Validity Evidence

Criterion Validity

The majority of the criterion validity evidence for easyCBM© WRF measures comes from four studies. Tindal, Nese, and Alonzo (2009) compared Grade 3 fall easyCBM© WRF measures with the Oregon state test administered in the spring. Jamgochian et al. (2010) compared easyCBM© WRF Grade 2 measures with the Stanford Achievement Tests, tenth edition (SAT-10) in reading. Sáez et al. (2010) compared easyCBM© WRF measures to the spring 2010 Grade 3 OAKS Reading assessment. Finally, Anderson et al. (2011) evaluated the relation between the easyCBM© WRF benchmark measures and the Grade 3 OAKS Reading assessment for three public school districts in Oregon.

Tindal et al. (2009) explored the validity of the fall WRF measures to predict spring state standardized test scores for two school districts in Oregon, reporting statistically significant correlations moderate in magnitude (> .60) across samples. Jamgochian, et al. (2010) explored predictive validity by calculating correlations between the Grade 2 fall and winter easyCBM WRF measures and the SAT-10 administered in the spring, both individually and in combination. They also explored the predictive utility of students' rate of growth (slope) in a year using Hierarchical Linear Modeling (HLM). The sample included approximately 2,200 students. Both Fall and Winter WRF were significantly correlated with SAT-10, with Pearson correlations of .15. Regression analyses were conducted for each of the measures separately and combined by season. Both the fall 2009 and Winter 2010 WRF measures uniquely explained approximately 2% of the variance in SAT-10 scores when they were included in the model with other easyCBM© benchmark measures. Rate of growth for students in grade two in WRF was low but positive across the first through the third quartiles, with students scoring, on average,

.23, .31, and .11, points higher on the SAT-10 for every one point of extra growth made on WRF, respectively.

Sáez et al. (2010) conducted two predictive validity analyses with the Grade 3 WRF measures. The first used the fall and winter benchmark measures to predict the state test administered in the spring, while the second used students' rate of growth (estimated from a hierarchical linear model) during the school year as a predictor of the same state test. The study included between 849 and 988 Oregon public school students, with sample sizes varying by season. Fall WRF scores accounted for approximately 36% of the variance in OAKS reading scores, while winter WRF scores accounted for approximately 49% of the total variance. As is to be expected, when other measures of students' reading ability (passage reading fluency, vocabulary, and comprehension) were included in the model, the predictive utility of WRF scores decreased from fall, to winter, to spring, as the proportion of unique variance in state test score accounted for by the different easyCBM© measures shifted somewhat across the different seasons. The spring 2010 OAKS Reading results were correlated at approximately .60 with the fall easyCBM results, and .61 with the winter easyCBM results. In terms of the predictive utility of students' rate of growth (slope), those who began the year in the normative first quartile of achievement were the only group with a moderate rate of growth, with students scoring, on average, .50 points higher on OAKS for each additional point of growth made. Other quartiles had a low relation between growth and OAKS (2^{nd} quartile r = .07, 3^{rd} quartile r = -.07, 4^{th} quartile r =.18), suggesting a possible ceiling effect for the WRF measure in higher performance brackets.

Anderson et al. (2011) found the Grade 3 Fall WRF measure to generally have a strong relation with the OAKS reading assessment, with correlations of .65 and .75 in two separate

districts. The third district investigated was the exception, exhibiting essentially no relation (-.03). The Winter WRF measure had a similar relation with OAKS, with a correlation of .60 in District 1, .69 in District 2, and .25 in District 3. No clear explanation for the difference observed in District 3 data was found, but sampling error is clearly possible.

Concurrent validity. Jamgochian et al. (2010) conducted a concurrent validity analysis by calculating correlations between the SAT-10 and the spring 2010 easyCBM© WRF measures, with a sample of 2,154 students. The authors concluded that the Grade 2 easyCBM WRF measure was not significantly correlated with the SAT-10. In addition, Sáez et al. (2010), conducted a different concurrent validity analysis, comparing the spring 2010 easyCBM© WRF measure to the spring 2010 Reading OAKS, Oregon's statewide assessment. With a Pearson correlation of .53, the spring 2010 WRF scores uniquely accounted for approximately 28% of the variance in spring 2010 OAKS reading scores. In a final concurrent validity study on WRF, Anderson et al. (2011) found the correlation between the Grade 3 spring WRF measure and the spring OAKS reading assessment was quite high across districts, at .70 in District 1, .60 in District 2, and .70 in District 3.

Construct Validity

To gather construct validity evidence for the easyCBM© Grades 2-3 WRF measures, a series of confirmatory factor analyses (CFAs) were conducted. In all CFA models, we hypothesized a one-factor structure, where one general *Reading* construct was measured. For the Grade 2 data, *Reading* was measured by WRF, Passage Reading Fluency (PRF) and a latent *Comprehension* variable. For Grade 3, a similar model was examined, with the addition of a vocabulary (VOC) measure. The latent *Comprehension* variable was measured by 12 multiple

choice reading comprehension (MCRC) items for Grade 2, and 20 MCRC items for grade. See Figures 7.2-7.3 for the hypothesized models.

For the CFAs across all grades and time points, one of the factor loadings for comprehension was constrained to be 1.0, along with the factor loadings for reading and comprehension latent constructs, to identify the model. All other factor loadings and variances were freely estimated. A weighted least squares estimator was used with the *Mplus* software (WLSMV; Muthén & Muthén, 1998-2007). Model fit was evaluated using the Tucker Lewis Index (TLI), Comparative Fit Index (CFI), and Root-Mean Square Error of Approximation (RMSEA). In particular, with binary and continuous model variables, CFI and TLI values ≥ 0.95, and RMSEA values ≤ 0.05 were considered indications of good model fit to the data (Yu, 2002).

Separate CFA models were conducted for each seasonal benchmark assessment for Grade 2, and only fall and spring assessments for Grade 3. The data were evaluated for their fit to the hypothesized, one-factor model. Overall, the model fit statistics suggested the data had good fit to the model for all seasons across all grades (see Table 7.1 below). The WRF measures also had consistent and strong factor loadings on the latent *reading* factor, suggesting that students' reading ability was a strong predictor of their performance on the easyCBM© WRF measure.

Table 7.1. *CFA Results for Grades 2 and 3.*

Grade	Time Point	*N*	Fit Statistics			Factor Loadings			
						Reading			Comprehension
			CFI	TLI	RMSEA	WRF	PRF	VOC	
2	F	1701	0.997	0.996	0.033	0.910	0.958	-	0.906
	W	1959	0.997	0.996	0.03	0.945	0.983	-	0.835
	S	1793	0.999	0.999	0.017	0.918	0.988	-	0.775
3	F	782	0.994	0.993	0.022	0.947	0.961	0.822	0.843
	S	876	0.991	0.990	0.027	0.881	0.995	0.815	0.700

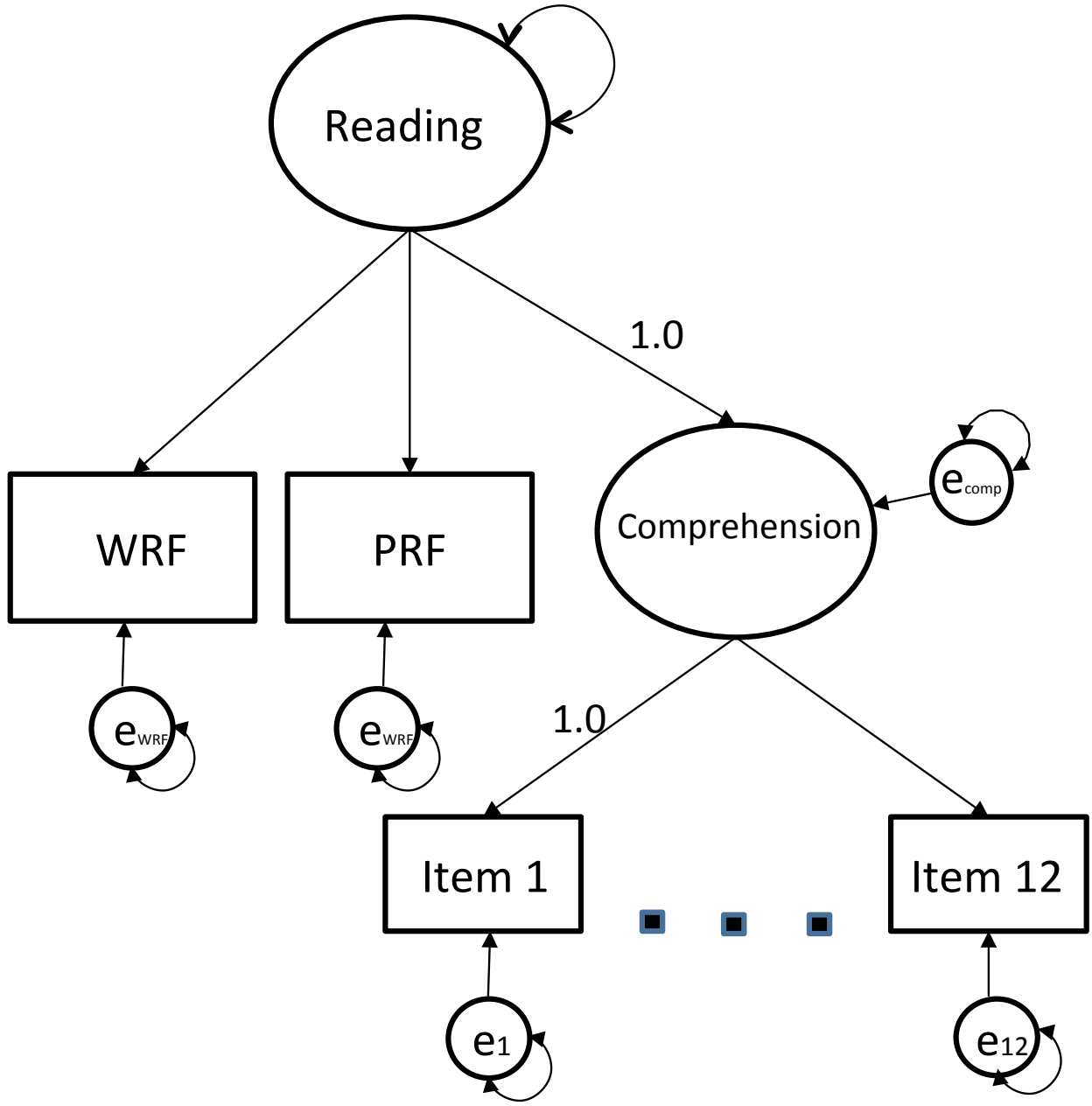


Figure 7.2. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM Grade Two measure contributes to a single, latent (unobservable), *Reading* trait. WRF = Word Reading Fluency, PRF = Passage Reading Fluency.

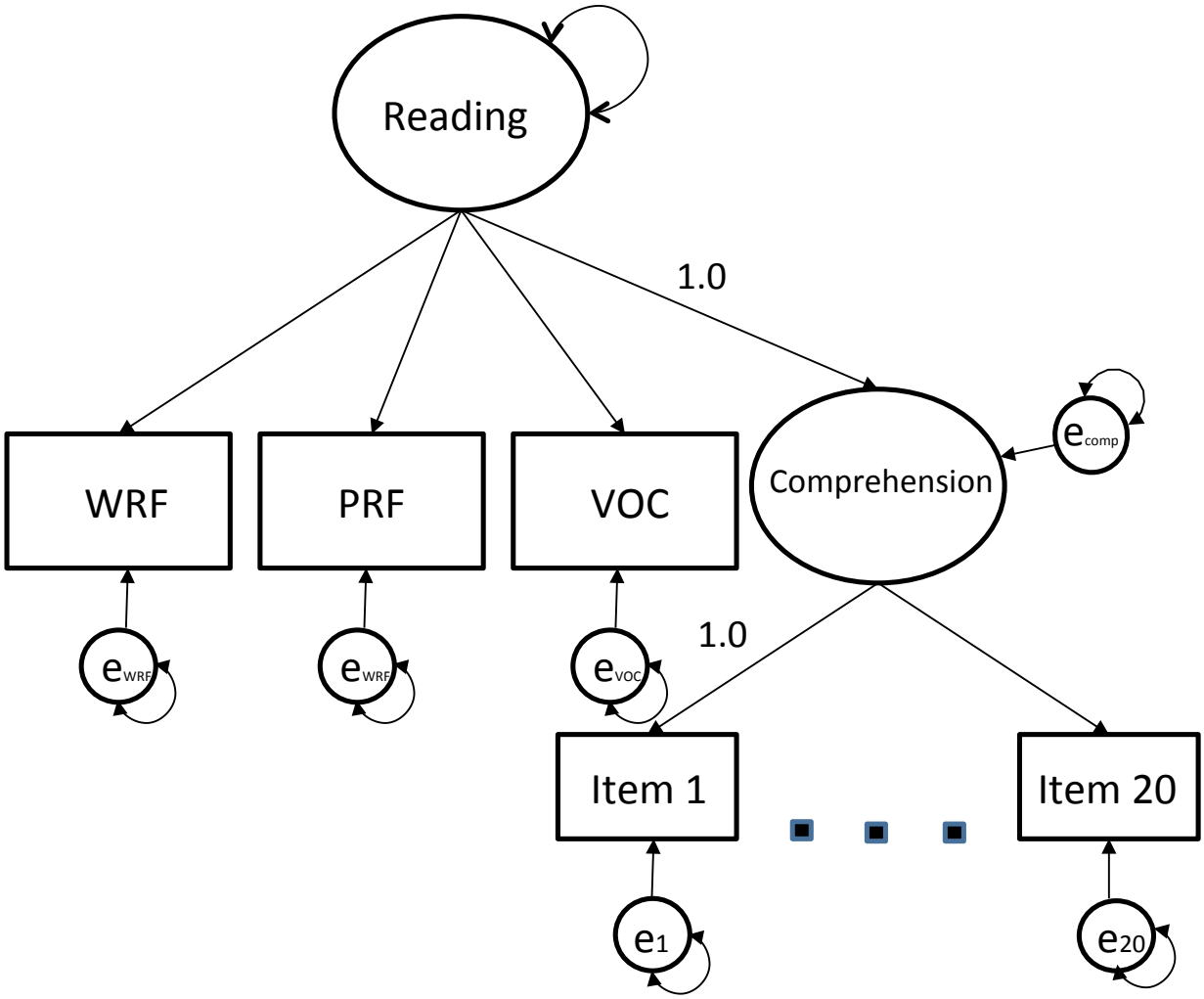


Figure 7.3. Theoretical One-Factor Confirmatory Factor Analysis: Grade 2. The model hypothesizes that each easyCBM Grade Three measure contributes to a single, latent (unobservable), *Reading* trait. WRF = Word Reading Fluency, PRF = Passage Reading Fluency, VOC = Vocabulary.

References

Alonzo, J., & Tindal, G. (2007). The development of word and passage reading fluency measures in a progress monitoring assessment system (technical report 40). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009). Alternate form and test-tetest reliability of easyCBM reading measures (technical report 0906). Eugene, OR: Behavioral Research and Teaching, University of Oregon. .

Anderson, D., Alonzo, J., & Tindal, G. (2011). easyCBM reading criterion related validity evidence: Oregon state test 2009-2010 (technical report 1103). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Lai, C. F., Park, B. J., Alonzo, J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 2 (technical report 1217). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Park, B. J., Lai, C., F., Alonzo, J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 1 (technical report 2016). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale–Chall readability formula*. Cambridge, MA: Brookline Books.

Jamgochian, E. M., Park, B. J., Nese, J. F. T., Lai, C. F., Sáez, L., Anderson, D., et al. (2010). Technical adequacy of the easyCBM grade 2 reading measures (technical report 1004). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Muthén, L. K., & Muthén, B. O. (1998-2007). Mplus User's Guide (Fifth ed.). Los Angeles, CA.

Park, B. J., Anderson, D., Alonzo, J., Lai, C. F., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 3 (technical report 1218). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Pinnell, G. S., & Fountas, I. (1998). *Word matters: Teaching phonics and spelling in the reading/writing classroom*. Portsmouth, NH: Heinemann.

Sáez, L., Irvin, P. S., Alonzo, J., & Tindal, G. (2013). Alignment with the Common Core State Standards: easyCBM K-3 Word Reading (Technical Report No. 1303). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Sáez, L., Park, B. J., Nese, J. F. T., Jamgochian, E. M., Lai, C. F., Anderson, D., et al. (2010). Technical adequacy of the easyCBM reading measures (Grades 3-7), 2009-2010 version (technical report 1005). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Tindal, G., Nese, J. F., & Alonzo, J. (2009). Criterion-related evidence using easyCBM® reading measures and student demographics to predict state test performance in grades 3-8 (technical report 0910). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Research Monograph No. 18). Madison, WI: University of Wisconsin-Madison, National Institute for Science Education.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Doctoral Dissertation, University of California Los Angeles.

Chapter 8: Passage Reading Fluency

The technical adequacy evidence gathered for the easyCBM© Passage Reading Fluency (PRF) measures to date suggests they are functioning largely as intended. During the developmental process, empirical data from pilot studies were used to evaluate the difficulty of each test form, with revisions made as needed to bring all forms into alignment in terms of difficulty (Alonzo, Park, & Tindal, 2008b; Alonzo & Tindal, 2007b, 2008b). Alternate form and test-retest reliability has been shown to be very high, while generalizability theory studies suggest the measures are quite reliable (Alonzo, Lai, Anderson, Park, & Tindal, 2012; Alonzo & Tindal, 2009; Anderson, Lai, Park, Alonzo, & Tindal, 2012; Anderson, Park, Lai, Alonzo, & Tindal, 2012; Lai, Park, Anderson, Alonzo, & Tindal, 2012; Park, Anderson, Alonzo, Lai, & Tindal, 2012). Multiple studies have shown the measures have a high relation with state tests and other relevant criteria (see Table 8.1). The sensitivity and specificity of the measures in predicting state test proficiency has also been shown to be high (see Table 8.1) Finally, Alonzo, Park, and Tindal (2013) showed the measures load strongly on a latent "reading" factor. In sum, these studies suggest the technical adequacy evidence for the easyCBM© PRF measures is quite strong.

In what follows, we summarize the studies cited above examining the technical adequacy evidence for the easyCBM© PRF measures. We begin with the developmental process used, including creating alternate test forms of equivalent difficulty. We then summarize reliability evidence, including alternate form, test-retest, and generalizability theory analyses. Finally, we conclude by discussing validity evidence, including the relation between PRF and relevant criterion measures, and the degree to which PRF loads a latent "Reading" factor. For a conceptual discussion of the theoretical purpose for any analysis conducted, please see Chapter

3. Note that this document is intended to provide a summary of evidence, not to explain in detail the different studies from which the evidence was gleaned. For a complete account of each study, the full reports are available at www.brtprojects.org/publications/technical-reports.

Measure Development

The development of PRF measure took place in two cycles, with Grade 1-4 passages written during the 2006-2007 school year and piloted in the spring of 2007 and Grade 5-8 passages written during the 2007-2008 school year and piloted in the spring of 2008. The development process began with a team of graduate students and former educators writing 20 alternate passages for each of grades K-8, and was followed by a four-stage review process. Throughout the review process, education researchers with experience in assessment development and measurement along with in-service teachers with grade-level reading instruction experience reviewed and revised each passage for grammatical correctness, grade-level appropriateness of content and style, and readability using the Flesch-Kinkaid index feature available on Microsoft Word.

Piloting was conducted using a sample of students located in the Pacific Northwest. Unlike most other measures in the easyCBM© assessment system, PRF measures were analyzed using classical test theory, as no discrete items were available for Rasch modeling. Differences in the average number of words read correctly were analyzed using a repeated measure analysis of variance (ANOVA). Correlations among test forms administered at the same time were also evaluated. These analyses provided empirical evidence on the equivalence, or lack of equivalence, of forms and guided subsequent revisions. In each grade, the nine passages that were most similar in difficulty were identified, and additional passages that were easier or more difficult were revised to bring them into closer alignment. Descriptive statistics including the

mean number of words read correctly per minute and correlations between each of the 20 alternate forms within each grade are presented in Alonzo and Tindal (2007a), Alonzo, Park, and Tindal (2008a) and Alonzo and Tindal (2008a). The piloting, statistical analyses, and review and revision process helped ensure test forms had adequate range (easy and difficult words) to sufficiently classify students into risk categories, along with an adequate reach to the lower tail of the distribution of student fluency performance to detect small changes for students being progress monitored over time. Descriptive and correlation analyses also allowed developers to construct the forms so the average passage difficulty was essentially equivalent across same-grade-level test forms.

Reliability

Alternate form and test-retest reliability for the easyCBM© Passage Reading Fluency (PRF) measures was investigated by Alonzo and Tindal (2009), Alonzo et al. (2012), Anderson, Lai et al. (2012), Anderson, Park et al (2012), Park et al. (2012), and Lai et al. (2012). Each study, with the exception of Alonzo and Tindal, also conducted generalizability and decision studies to explore how various facets of the measurement process impacted the reliability of the PRF measures. Across all studies, the sample size was relatively small, ranging from 19-42 depending on the specific test form administered for Grade 1, 34-50 for Grade 2, 17-53 for Grade 3, 20-74 for Grade 4, and 19-87 for Grade 5.

Alternate Form Reliability. Alonzo and Tindal (2009) administered PRF forms 1, 3, and 5 to Grade 1 students on each of two separate occasions, spaced one week apart in the spring of 2008. The correlation between students' scores on each form ranged from .95-.97, indicating a very strong relation. In other words, students' scores were quite stable regardless of the specific test form administered. Similarly, Alonzo and Tindal found the correlations in Grade 3 ranged

from .94-.95, Grade 5 ranged from .87-.96, and Grade 8 ranged from .87-.95, all indicating a very strong relation. Anderson, Park et al. (2012) also investigated the alternate form reliability of the Grade 1 PRF measures using a similar approach, but with test forms 11, 12, 13, 14, 15, and 16. Correlations ranged from .93-.98, again indicating a very strong relation. Anderson, Lai et al. (2012) investigated the same form numbers, but at Grade 2, and found similar results (range .91-.95); as did Park et al. (2012) in Grade 3 (range .92-.96), Alonzo et al. (2012) in Grade 4 (range .83-.98), and Lai et al (2012) in Grade 5 (range .85-.98). These numbers suggest that students' scores are very stable across test forms. See the full reports for complete correlation tables.

Test-Retest Reliability. Test-retest reliability was also investigated in each of the above studies, with the same forms used to investigate alternate form reliability. Alonzo and Tindal (2009) found the test-retest correlations to range from .91-.97 for Grade 1, indicating a very strong relation. Also for Grade 1, Anderson et al. found correlations ranging from .83-.98, with a median of .97, again indicating a very strong relation. Test-retest correlations for Grade 2 ranged from .88-.96, with a median of .94, Grade 3 ranged from .84-.94, with a median of .90, Grade 4 ranged from .86-.96, with a median of .95, and Grade 5 ranged from .88-.94 (with form 14 at .54), with a median of .91 (Alonzo et al., 2012; Anderson, Lai, et al., 2012; Lai et al., 2012; Park et al., 2012).

Generalizability Study. In 2012, all the above studies (except Alonzo and Tindal, 2009) also included generalizability analyses primarily to examine how specific facets of the measurement process related to the reliability of the measures. Across test forms, the G-coefficients ranged from .91-.99 for Grade 1, .97-.98 for Grade 2, .95-.97 for Grade 3, .94-.98 for Grade 4, and .90-.97 for Grade 5 (Alonzo et al., 2012; Anderson, Lai et al., 2012; Anderson, Park

et al., 2012; Lai et al., 2012; Park et al., 2012). These results largely confirm the results of the test-retest and alternate form reliability analyses, suggesting the measures are highly stable.

Validity Evidence

Criterion Validity

Multiple studies have investigated the criterion validity of the easyCBM© PRF measures. Table 1 displays a summary of these studies, including the grade levels investigated, whether predictive or concurrent validity (or both) was investigated, and the overall results found. Each of these studies is summarized in more detail below. See the full reports for a complete description of the study details.

Table 8.1
Studies investigating the validity of easyCBM© PRF

Study	Grades	Predictive/Concurrent	Summary of results
Tindal, Nese, and Alonzo (2009)	3-8	Both	r = .55 to .69 Model R^2 = .47 to .69
Jamgochian et al. (2010)	2	Both	r = .19 to .22
Sáez et al. (2010)	3-7	Both	r = .35 to .66 Model R^2 = .41 to .48
Anderson, Alonzo, and Tindal (2011a)	3-8	Both	r = -.05[a] to .67
Anderson, Alonzo, and Tindal (2011b)	3-7	Both	r = .55 to .71
Park, Anderson, Irvin, Alonzo, and Tindal (2011)	3-8	Predictive	Sensitivity = .71 to .89 Specificity = .70 to .81 AUC = .82 to .91
Park, Irvin, Anderson, Alonzo, and Tindal (2011)	3-8	Both	Sensitivity = .75 to .90 Specificity = .70 to .81 AUC = .80 to .91
Irvin, Park, Anderson, Alonzo, and Tindal (2011)	3-8	Predictive	Sensitivity = .70 to .85 Specificity = .43 to .81 AUC = .62 to .87
Anderson, Park, Irvin, Alonzo, and Tindal (2011)	3-8	Predictive	Sensitivity = .70 to .85 Specificity = .51 to .71 AUC = .66 to .84

(Lai, Alonzo, & Tindal, 2013a)	2-5	Concurrent	r = .88 to .95

Note. A state test was used as the criterion in all studies with the exception of Jamgochian et al. and Lai et al., who used the Stanford-10 and DIBELS tests respectively.
[a] Value represents an outlier; Median = .57

Predictive validity. Tindal et al. (2009) explored the predictive validity of the easyCBM© PRF measures by comparing the fall and winter measures to the state standardized test scores for two school districts in Oregon. The sample included approximately 2,000 total students per grade. The correlations between PRF and the Oregon Assessment of Knowledge and Skills (OAKS), the state test used in Oregon, were generally high, ranging from .55 to .69. Multiple regression analyses were also conducted, which included demographic control variables (gender, historically low/high achieving, economic disadvantage, and special education) and two additional reading variables (vocabulary and multiple-choice reading comprehension). The overall model accounted for 47-69% of the variance in OAKS reading scores. The fall easyCBM© PRF measure uniquely accounted for approximately 2-7% of the variance in OAKS across grades, *beyond* what was accounted for by the demographic and other reading variables. The winter easyCBM© PRF measure uniquely accounted for approximately 8-16% of the variance in OAKS. In models not including vocabulary, the winter PRF measure uniquely accounted for between 20-29% of the variance in OAKS.

Jamgochian et al. (2010) explored the predictive validity of the easyCBM© PRF measures by examining the relation between the Stanford-10 (SAT-10) and the fall 2009 and winter 2010 easyCBM© PRF measures, respectively. The sample ranged from 2,205-2,236 students. Fall and winter PRF were significantly correlated with the SAT-10 at .19 and .22, respectively. Simple linear regression analyses were also conducted. The PRF measures accounted for 4% and 5% of the variance in the SAT-10 in fall and winter, respectively.

Jamgochian et al. also explored the predictive utility of students' rate of growth (slope) by correlating an estimate of students' growth across the year (hierarchical linear model random effect) with their SAT-10 performance in the spring. The authors found correlations ranging from -0.11 for the top quartile of student performance, to 0.60 for the bottom quartile. In other words, students' rate of growth had a strong relation with SAT-10 performance for students performing below expectations, but was less related for students performing above expectations.

Sáez et al. (2010) conducted two studies to explore the predictive validity of PRF. The first examined the correlation from fall 2009 and winter 2010 PRF performance with spring 2010 OAKS performance. The second used rate of growth (slope) as a predictor of the same outcome, and also explored the diagnostic efficiency of the measures (sensitivity/specificity). The study included approximately 2,000 students per grade; all participants were Oregon public school students in Grades 3-7. The correlations for the fall PRF measure with OAKS ranged from .35-.64, while the winter ranged from .35-.65. Correlations generally decreased as grade level increased. Sáez et al. (2010) conducted analyses with both the full sample of students, and by demographic subgroup. For the full sample regression analyses, fall PRF scores accounted for approximately 42-45% of the variance in OAKS. The winter measure accounted for 41-47% of the variance in OAKS scores. See the full reports for results by demographic group. The authors also explored the predictive utility of growth rates, and explored growth "cut scores" for most accurately predicting performance on OAKS. These results can also be found in the full report.

Anderson, Alonzo et al. (2011a) report correlations between the winter PRF measure and the spring state test used in Washington state, the Measures of Student Progress (MSP), ranging from .46 to .64. Anderson, Alonzo et al. (2011b) report similar results for the relation of PRF

with OAKS, with correlations ranging from .58 to .71 for the fall measure, and .56 to .70 for the winter.

Park, Anderson et al. (2011) examined the diagnostic efficiency of the fall and winter easyCBM© PRF measures relative to meeting or not meeting proficiency on the spring OAKS assessment in Grades 3-8. The sample included approximately 2,000 students per grade level from three districts in Oregon. The fall and winter results from sensitivity analyses ranged from .71 to .88. The fall and winter results for specificity ranged from .70 to .81. The overall correct classification results for fall and winter ranged from .72 to .81. The Area Under the ROC Curve statistic (AUC) for fall and winter ranged from .82 to .91. Irvin et al. (2011) replicated the Park, Anderson, et al. (2011) study in Washington state with a sample of approximately 1,200 students per grade. The fall and winter results from sensitivity analyses ranged from .70 to .85. The fall and winter results for specificity ranged from .51 to .71. The overall correct classification results for fall and winter ranged from .61 to .75. The Area Under the ROC Curve statistic (AUC) for fall and winter ranged from .66 to .84. These results suggest that easyCBM© PRF measures are *good* to *excellent* at discriminating between students who will and will not reach proficiency on the OAKS and MSP.

Park, Irvin et al. (2011) and Irvin et al. (2011) conducted a cross-validation study of the cut scores to optimally predict OAKS and MSP performance, respectively, in Grades 3-8. The studies extended the work of Park, Anderson et al. (2011) and Anderson, Park et al. (2011) by exploring the stability of the cut scores across two randomly selected groups. The authors noted that the cut scores appeared quite stable overall, with the average difference in cut scores between groups approximately 1.50 correct words per minute (CWPM). The 95% confidence intervals for AUC statistics overlapped between the groups for PRF, as well. The consistency

between the optimal cut scores combined with the lack of significant differences between AUC statistics in all measurement occasions and grades provide strong evidence for the cut scores derived.

Concurrent validity. Tindal et al. (2009) also conducted a concurrent validity analysis comparing spring easyCBM© PRF to spring OAKS in Grades 3 to 8. Multiple regression analyses again included demographic control variables (gender, historically low/high achieving, economic disadvantage, and special education) and two additional reading variables (vocabulary and multiple-choice reading comprehension). The overall model accounted for 54-69% of the variance in OAKS reading scores. The spring easyCBM© PRF measure uniquely accounted for approximately 2-12% of the variance in OAKS across grades, *beyond* what was accounted for by the demographic and other reading variables.

Jamgochian et al. (2010) conducted a concurrent validity analysis by calculating correlations between the SAT-10 and the Grade 2 spring 2010 easyCBM© measures. The authors concluded that the Grade 2 easyCBM© PRF measure was not significantly correlated with the SAT-10, with a Pearson correlation of .05. The measure was also analyzed with a regression analysis, which included word reading fluency and multiple-choice reading comprehension in the model. While the model explained 55% of the variance in SAT-10 scores, the easyCBM© PRF measure uniquely explained approximately 1.6% of the variance in SAT-10 scores.

Sáez et al. (2010) explored the concurrent validity of PRF by examining the relation between spring OAKS and PRF. The sample included approximately 2,000 students per grade. Correlations ranged from .39-.66, generally decreasing as grade increased. Regression analyses were also performed both with the full sample and by demographic subgroups. For the full

sample, PRF accounted for approximately 42-48% of the variance in OAKS. See the full report for results disaggregated by subgroup.

Anderson, Alonzo et al. (2011a) reported concurrent correlations between the spring PRF measure and the MSP ranging from -.05 to .67, with a median of .57. Anderson, Alonzo et al. (2011b) reported similar results for the concurrent relation of PRF with OAKS, with correlations ranging from .52 to .70.

Park, Anderson et al. (2011) and Anderson, Park et al. (2011) reviewed the diagnostic efficiency of the spring easyCBM PRF relative to students meeting or not meeting proficiency on the OAKS and MSP, respectively, in Grades 3-8. The Park, Anderson et al. sample included approximately 2,000 students per grade level, while Anderson, Park et al. included approximately 1,200 students per grade. Park, Anderson et al. found sensitivity ranging from .79 to .89, specificity ranging from .70 to .80, classification consistency ranging from .73 to .81, and AUC ranging from .83 to .90. Anderson, Park et al. found sensitivity ranging from .71 to .81, specificity ranging from .60 to .70, classification consistency ranging from .64 to .74, and AUC ranging from .72 to .82. These studies suggest that the easyCBM© PRF measures are *fair* to *good* at discriminating which students will and will not meet proficiency on OAKS and the MSP, respectively.

Lai, Alonzo, and Tindal (2013b) examined the concurrent relation between PRF and the DIBELS ORF assessment, 6th edition. Approximately 200 students per grade participated in the study. High correlations were reported across all grades, ranging from .88-.95.

Construct Validity

To gather construct validity evidence for the easyCBM© Grades 2-7 PRF measures, a series of confirmatory factor analyses (CFAs) were conducted (Alonzo et al., 2013). In all CFA

models, we hypothesized a one-factor structure, where one general *Reading* construct was measured. For the Grade 2 data, *Reading* was measured by Word Reading Fluency (WRF), PRF and a latent *Comprehension* variable. For Grade 3, a similar model was examined, with the addition of a vocabulary measure. For Grades 4-7, *Reading* was measured by PRF, Vocabulary (VOC) and the latent *Comprehension* variable. The latent *Comprehension* variable was measured by 12 Multiple Choice Reading Comprehension (MCRC) items for Grade 2, and 20 MCRC items for Grades 3-7. See Figures 8.1-8.3 for the hypothesized models.

For the CFAs across all grades and time points, one of the factor loadings for comprehension was constrained to be 1.0, along with the factor loadings for reading and comprehension latent constructs, to identify the model. All other factor loadings and variances were freely estimated. A weighted least squares estimator was used with the *Mplus* software (WLSMV; Muthén & Muthén, 1998-2007). Model fit was evaluated using the Tucker Lewis Index (TLI), Comparative Fit Index (CFI), and Root-Mean Square Error of Approximation (RMSEA). In particular, with binary and continuous model variables, CFI and TLI values ≥ 0.95, and RMSEA values ≤ 0.05 were considered indications of good model fit to the data (Yu, 2002).

Separate CFA models were conducted for each seasonal benchmark assessment for Grade 2, while only fall and spring assessments for Grades 3-7. The data were evaluated for their fit to the hypothesized, one-factor model. Overall, the model fit statistics suggested the data had good fit to the model for all seasons across all grades (see Table 8.2 below). The PRF measures also had moderate to strong factor loadings on the latent *Reading* factor (.70-.90), suggesting that students' reading ability was a strong predictor of their performance on the easyCBM© PRF measure.

Table 8.2 – CFA Results for Grades 2-7

Grade	Time Point	*N*	Fit Statistics			Factor Loadings			
						Reading			
			CFI	TLI	RMSEA	WRF	PRF	VOC	Comprehension
2	F	1701	0.997	0.996	0.033	0.910	0.958	-	0.906
	W	1959	0.997	0.996	0.03	0.945	0.983	-	0.835
	S	1793	0.999	0.999	0.017	0.918	0.988	-	0.775
3	F	782	0.994	0.993	0.022	0.947	0.961	0.822	0.843
	S	876	0.991	0.990	0.027	0.881	0.995	0.815	0.700
4	F	1887	0.979	0.977	0.026	-	0.852	0.831	0.895
	S	2048	0.984	0.982	0.022	-	0.779	0.768	0.803
5	F	2037	0.981	0.979	0.022	-	0.856	0.753	0.862
	S	2123	0.980	0.978	0.023	-	0.781	0.712	0.824
6	F	1240	0.961	0.957	0.025	-	0.791	0.742	0.850
	S	1148	0.987	0.986	0.016	-	0.790	0.715	0.844
7	F	791	0.962	0.957	0.026	-	0.769	0.690	0.842
	S	848	0.978	0.975	0.018	-	0.707	0.521	0.904

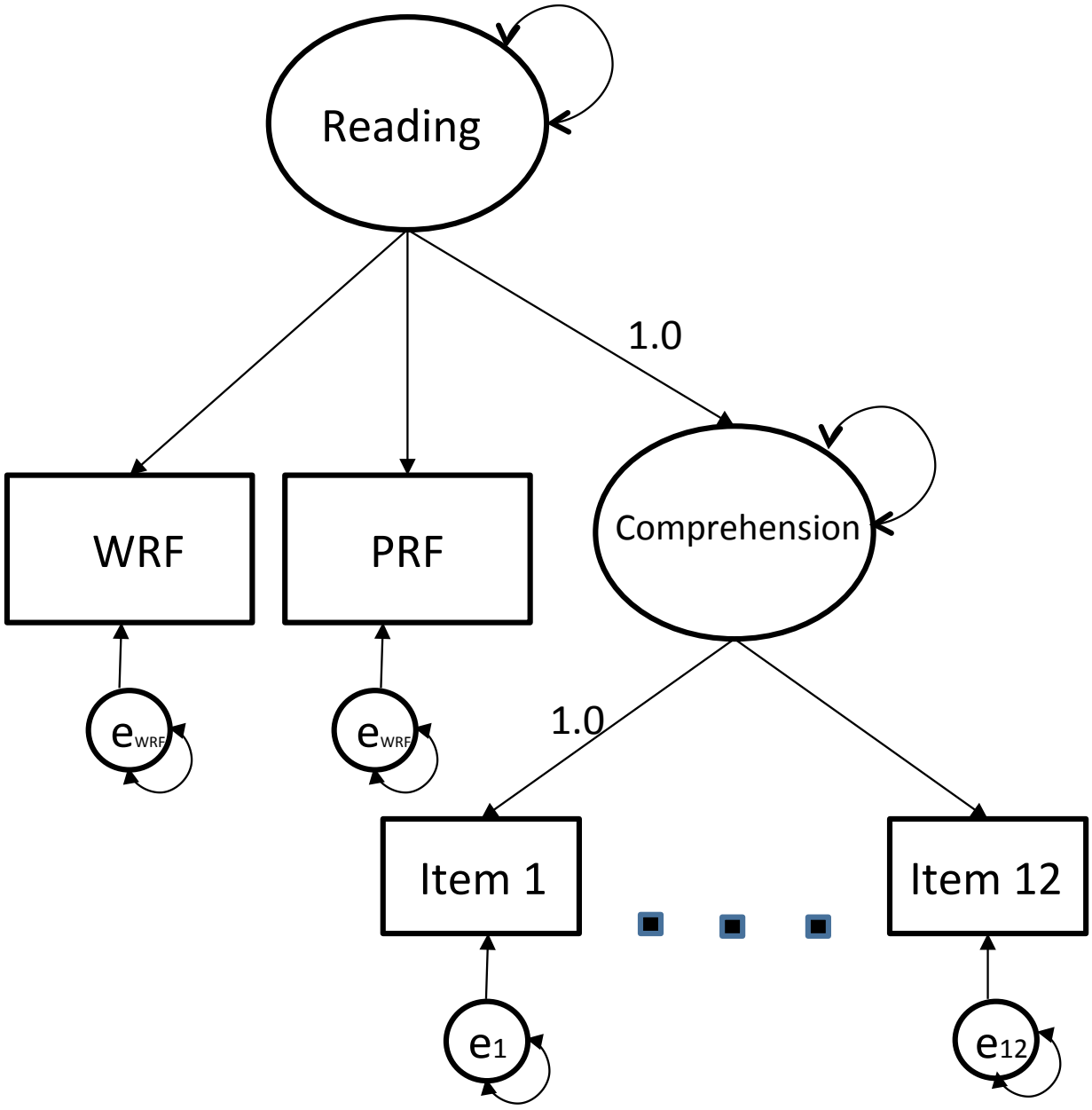


Figure 8.1. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM Grade Two measure contributes to a single, latent (unobservable), *Reading* trait. WRF = Word Reading Fluency, PRF = Passage Reading Fluency.

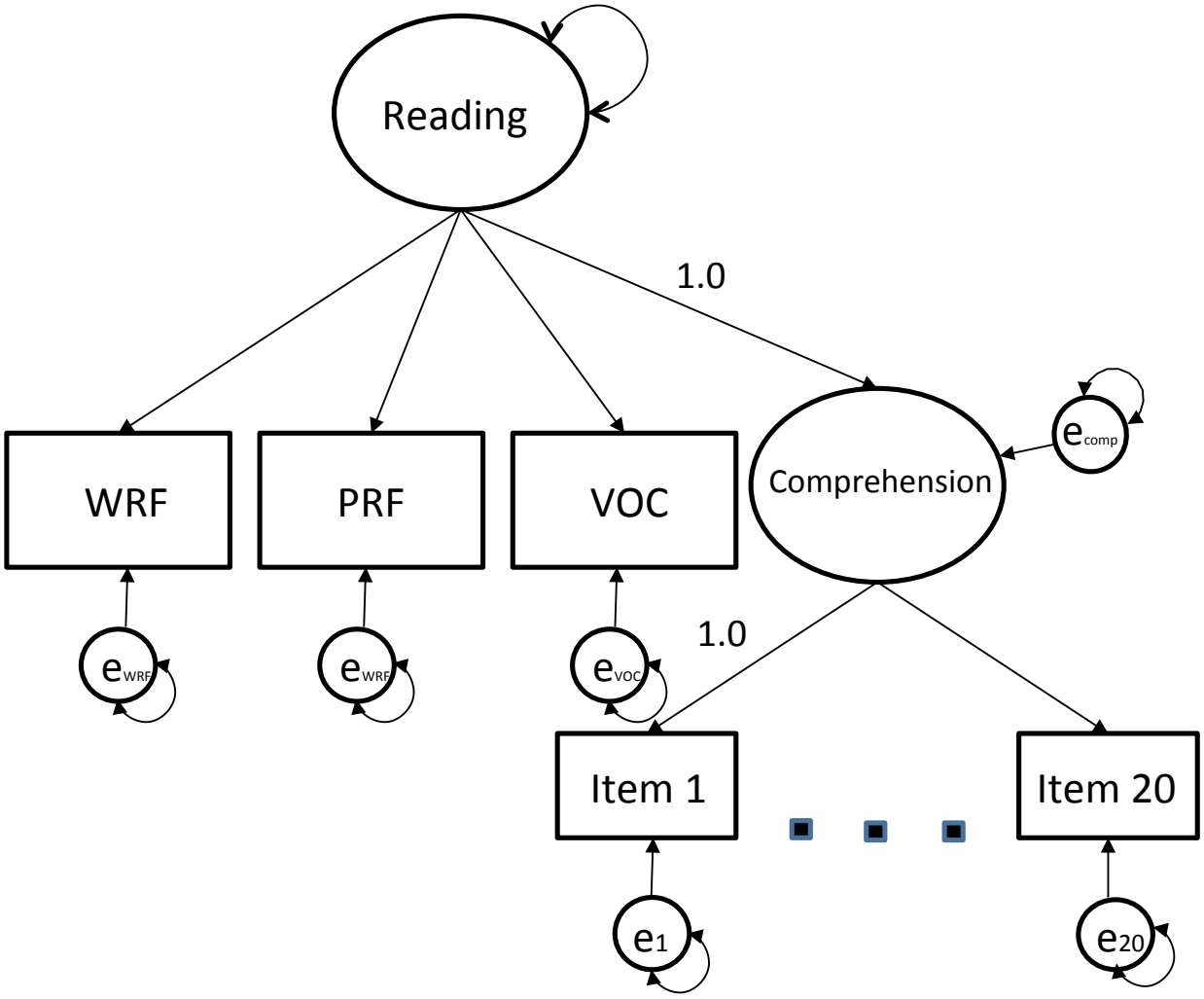


Figure 8.2. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM Grade Three measure contributes to a single, latent (unobservable), *Reading* trait. WRF = Word Reading Fluency, PRF = Passage Reading Fluency, VOC = Vocabulary.

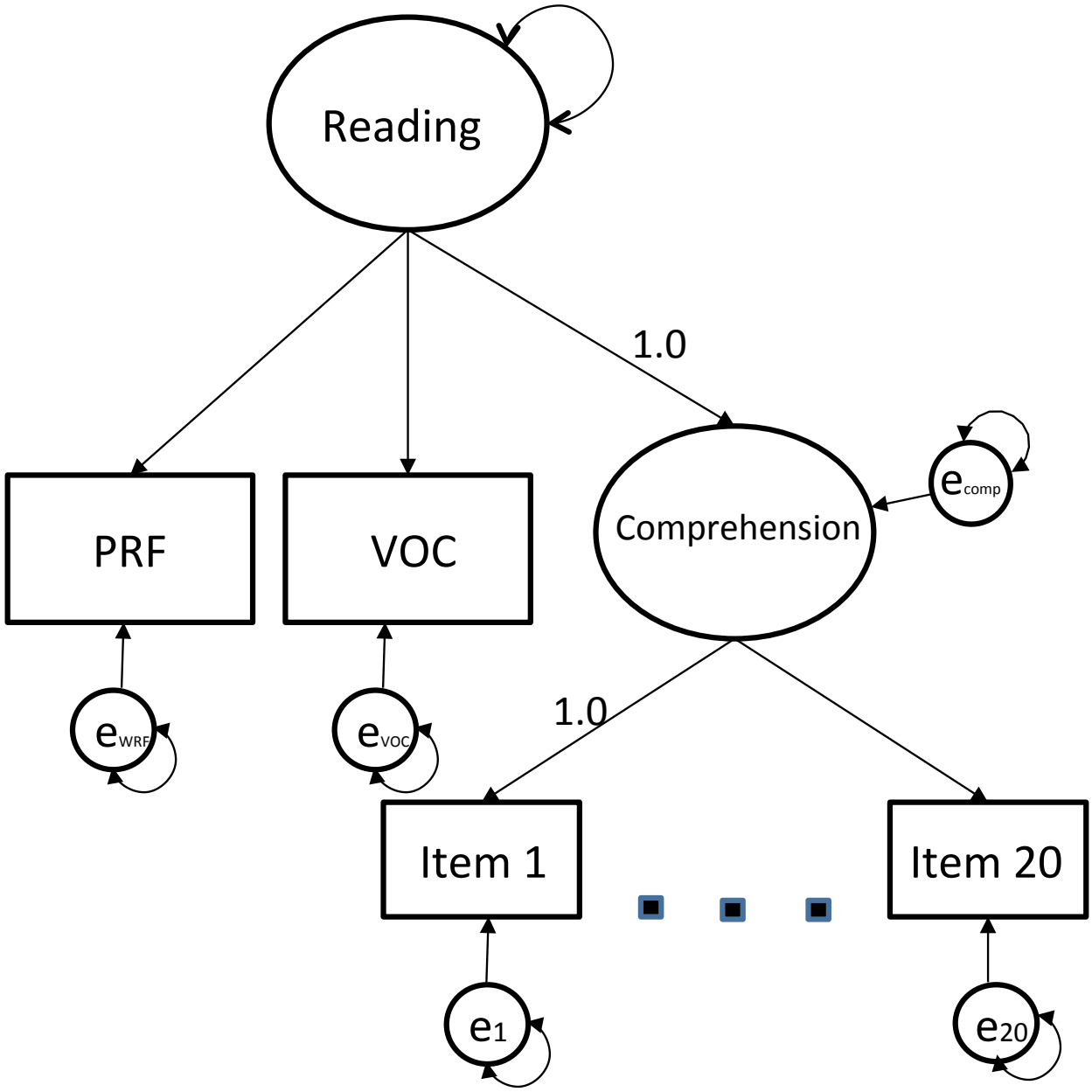


Figure 8.3. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM Grade Four-Seven measure contributes to a single, latent (unobservable), *Reading* trait. PRF = Passage Reading Fluency, VOC = Vocabulary.

References

Alonzo, J., Lai, C. F., Anderson, D., Park, B. J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 4 (technical report 1219). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2008a). The development of middle school passage Reading fluency measures in a progress monitoring assessment system (technical report 46). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2008b). The development of middle school passage Reading fluency measures in a progress monitoring assessment system (Technical Report No. 46). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2013). Examining the internal structure of the easyCBM reading measures, Grades K-5 (technical report 1302). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2007a). The development of word and passage reading fluency measures in a progress monitoring assessment system (technical report 40). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2007b). The development of word and passage reading fluency measures in a progress monitoring assessment system (Technical Report No. 40). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2008a). The development of fifth-grade passage reading fluency measures for use in a progress monitoring assessment system (technical report 43). Eugene, OR: Behavioral Research and Teaching.

Alonzo, J., & Tindal, G. (2008b). The development of fifth-grade passage reading fluency measures in a progress monitoring assessment system (Technical Report No. 43). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009). Alternate form and test-tetest reliability of easyCBM reading measures (technical report 0906). Eugene, OR: Behavioral Research and Teaching, University of Oregon. .

Anderson, D., Alonzo, J., & Tindal, G. (2011a). easyCBM reading criterion related to validity evidence: Washington state test 2009-2010 (technical report 1101). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2011b). easyCBM reading criterion related validity evidence: Oregon state test 2009-2010 (technical report 1103). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Lai, C. F., Park, B. J., Alonzo, J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 2 (technical report 1217). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Park, B. J., Irvin, P. S., Alonzo, J., & Tindal, G. (2011). Diagnostic efficiency of easyCBM reading: Washington State (technical report 1107). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Park, B. J., Lai, C., F., Alonzo, J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 1 (technical report 2016). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Park, B. J., Anderson, D., Alonzo, J., & Tindal, G. (2011). A cross-validation of easyCBM reading cut scores in Washington: 2009-2010 (technical report 1109). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Jamgochian, E. M., Park, B. J., Nese, J. F. T., Lai, C. F., Sáez, L., Anderson, D., et al. (2010). Technical adequacy of the easyCBM grade 2 reading measures (technical report 1004). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2013a). easyCBM reading criterion related validity evidence: Grades 2-5 (technical report 1310). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2013b). easyCBM reading criterion related validity evidence: Grades k-1 (technical report 1309). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Park, B. J., Anderson, D., Alonzo, J., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 5 (technical report 1220). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Muthén, L. K., & Muthén, B. O. (1998-2007). Mplus User's Guide (Fifth ed.). Los Angeles, CA.

Park, B. J., Anderson, D., Alonzo, J., Lai, C. F., & Tindal, G. (2012). An examination of test-retest, alternate form reliability, and generalizability theory study of the easyCBM reading assessments: Grade 3 (technical report 1218). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2011). Diagnostic efficiency of easyCBM reading: Oregon (technical report 1106). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Irvin, P. S., Anderson, D., Alonzo, J., & Tindal, G. (2011). A Cross-validation of easyCBM Reading cut scores in Oregon: 2009-2010 (technical report 1108). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Sáez, L., Park, B. J., Nese, J. F. T., Jamgochian, E. M., Lai, C. F., Anderson, D., et al. (2010). Technical adequacy of the easyCBM reading measures (Grades 3-7), 2009-2010 version (technical report 1005). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Tindal, G., Nese, J. F., & Alonzo, J. (2009). Criterion-related evidence using easyCBM® reading measures and student demographics to predict state test performance in grades 3-8 (technical report 0910). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Doctoral Dissertation, University of California Los Angeles.

Chapter 9: Vocabulary

The easyCBM© Vocabulary measures are relatively new to the system, becoming operational for the first time during the 2011-12 school year. As such, technical adequacy documentation is limited relative to other easyCBM© measures. In this chapter, we discuss the developmental process used, including scaling items with a Rasch model to obtain information on the difficulty and functioning of each Vocabulary item, which guided test form creation to help ensure all 20 test forms at each grade are essentially equivalent in terms of difficulty and the distribution of easy and difficult items within each test form. We further discuss a study by Lai, Alonzo, and Tindal (2013), who conducted a concurrent criterion validity study, laying the groundwork for an evidence base supporting the easyCBM© Vocabulary measures. For a conceptual discussion of the theoretical purpose for any analysis conducted, please see Chapter 3. Note that this document is intended to provide a summary of evidence, not to explain in detail the different studies from which the evidence was gleaned. For a complete account of each study, the full reports are available at www.brtprojects.org/publications/technical-reports.

Measure Development

Item piloting was conducted with a nationwide sample of students during the 2010-2011 school year. Data were collected via a secure online piloting system for approximately 5,600 vocabulary items across Grades 2-8. During the pilot, each student was administered between 17 (Grade 2) and 20 (Grades 3-8) vocabulary items. Among these items, five within each grade were common (anchor items), with the remaining items selected from the full pool of grade-level items based on a conditional random sampling algorithm that selected items based on the number of students having previously responded to the item. The five anchor items allowed for all items

to be calibrated on a common scale. Approximately 1,200 students participated in the Grade 2 piloting, with approximately 2,800 participating in each of Grades 3-8.

All items were scaled with a Rasch model (see Chapter 2 for a conceptual overview of this methodology). The difficulty and model fit of each item in all Vocabulary test forms, including distractor analyses of how answer options functioned, are reported in a series of grade-specific technical reports by Alonzo, Anderson, Park, and Tindal (2012a, 2012b, 2012c, 2012d, 2012e, 2012f, 2012g). Results from the Rasch analyses helped ensure test forms had adequate range (from easy to difficult items) to sufficiently classify students into risk categories, along with an adequate number of items on the lower tail of the distribution to detect small changes for students being progress monitored over time. The Rasch analyses also allowed us to construct the forms so the average item difficulty was essentially equivalent across forms. Further, Rasch analysis of both item and distractor options provided robust evidence of item functioning, with those deemed to be inappropriately functioning not placed on operational forms.

Reliability

Wray, Alonzo, and Tindal (2014) investigated the internal consistency of the easyCBM© Vocabulary measures for the fall and winter benchmark across all grades (2-8) using a large extant dataset (*n* range = 17,382 to 30,598). Cronbach's Alpha and Split-half reliabilities (first half/second half) as well as item level statistics of the top and bottom 27th percentiles, were estimated. Table 9.1 below reports internal and split-half reliabilities. Cronbach's Alpha ranged from .76 to .84 and had a median of .81 for all vocabulary measures in both the fall and winter. Split-half reliabilities ranged from .61 to .75 for the first and second half of the measures, with a median of .66 and .69, respectively. The correlation between the two halves ranged from .58 to .72 with a median correlation of .64. Relative to the top/bottom reliability, all items performed as

expected, with higher percentile students getting the items correct more often than the lower percentile groups. Complete statistics can be found in the full report.

Table 9.1

Internal Reliability: easyCBM© Vocabulary

		Split-half Reliability		
Grade/Time	Cronbach's Alpha	1st Half	2nd Half	Correlation
2/F	.83	.69	.75	.70
2/W	.84	.73	.73	.72
3/F	.84	.72	.75	.69
3/W	.82	.70	.72	.67
4/F	.82	.71	.72	.65
4/W	.81	.65	.71	.64
5/F	.79	.62	.68	.63
5/W	.78	.61	.68	.59
6/F	.80	.67	.69	.62
6/W	.81	.65	.64	.65
7/F	.79	.66	.67	.64
7/W	.78	.64	.66	.61
8/F	.79	.66	.69	.61
8/W	.76	.65	.64	.58

Validity Evidence

Concurrent Validity

Lai et al. (2013) correlated the easyCBM© Vocabulary assessment with the Gates-MacGinitie Word Knowledge assessments in Grades 2-5. The sample included students from ten schools in one Oregon school district. Approximately 100-200 students per grade participated in the study. Correlations between the tests varied across grades, ranging from .39-.76. The low-moderate correlations were attributed to differences in assessment targets, as the easyCBM© measures sample from the Oregon State Standards for vocabulary, while the Gates-Macginitie specifically samples idioms, parts of speech, and word meaning.

Construct Validity

To gather construct validity evidence for the easyCBM© grades three through seven Vocabulary (VOC) measures, a series of confirmatory factor analysis (CFA) models were conducted. In all of the CFA models across time points, we hypothesized a one-factor structure, where one general "*reading*" construct was measured. For the grade three data, "*reading*" was measured by Word Reading Fluency (WRF), Passage Reading Fluency (PRF), Vocabulary (VOC) and the latent variable (*comprehension*). For grades four through seven, "*reading*" was measured by PRF, VOC and the *comprehension* latent variable. In all models, the latent variable *comprehension* was measured by 20 Multiple Choice Reading Comprehension (MCRC). See Figures 9.1-9.2 for the hypothesized models.

For the CFAs across all grades and time points, one of the factor loadings for comprehension was constrained to be 1.0, along with the factor loadings for reading and comprehension latent constructs, to identify the model. All other factor loadings and variances were freely estimated. A weighted least squares estimator was used with the *Mplus* software

(WLSMV; Muthén & Muthén, 1998-2007). Model fit was evaluated using the Tucker Lewis Index (TLI), Comparative Fit Index (CFI), and Root-Mean Square Error of Approximation (RMSEA). In particular, with binary and continuous model variables, CFI and TLI values ≥ 0.95, and RMSEA values ≤ 0.05 were considered indications of good model fit to the data (Yu, 2002).

Separate CFA models were conducted for fall and spring benchmarks. The data were then evaluated for their fit to the hypothesized, one-factor model. Overall, the model fit statistics suggested the data had good fit to the model for all seasons across all grades (see Table 9.2 below). The VOC measures also had moderate to moderately strong factor loadings on the latent *Reading* factor (.50-.80), suggesting that students' reading ability was a strong predictor of their performance on the easyCBM© VOC measure.

Table 9.2 *CFA Results for Grades 3-7.*

Grade	Time Point	*N*	Fit Statistics			Factor Loadings			
						Reading			Comprehension
			CFI	TLI	RMSEA	WRF	PRF	VOC	
3	F	782	0.994	0.993	0.022	0.947	0.961	0.822	0.843
	S	876	0.991	0.990	0.027	0.881	0.995	0.815	0.700
4	F	1887	0.979	0.977	0.026	-	0.852	0.831	0.895
	S	2048	0.984	0.982	0.022	-	0.779	0.768	0.803
5	F	2037	0.981	0.979	0.022	-	0.856	0.753	0.862
	S	2123	0.980	0.978	0.023	-	0.781	0.712	0.824
6	F	1240	0.961	0.957	0.025	-	0.791	0.742	0.850
	S	1148	0.987	0.986	0.016	-	0.790	0.715	0.844
7	F	791	0.962	0.957	0.026	-	0.769	0.690	0.842
	S	848	0.978	0.975	0.018	-	0.707	0.521	0.904

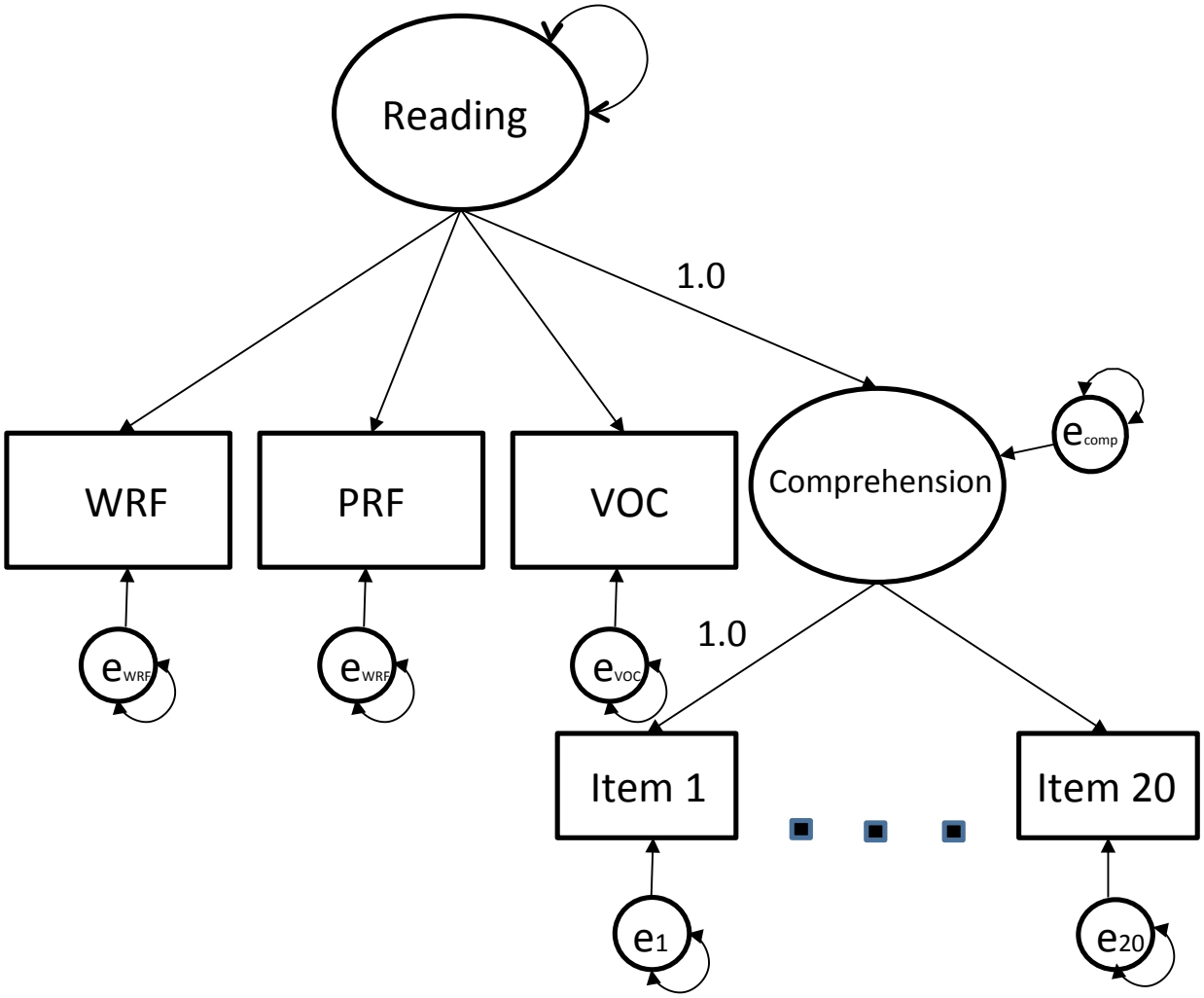


Figure 9.1. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM Grade Three measure contributes to a single, latent (unobservable), *Reading* trait. WRF = Word Reading Fluency, PRF = Passage Reading Fluency, VOC = Vocabulary.

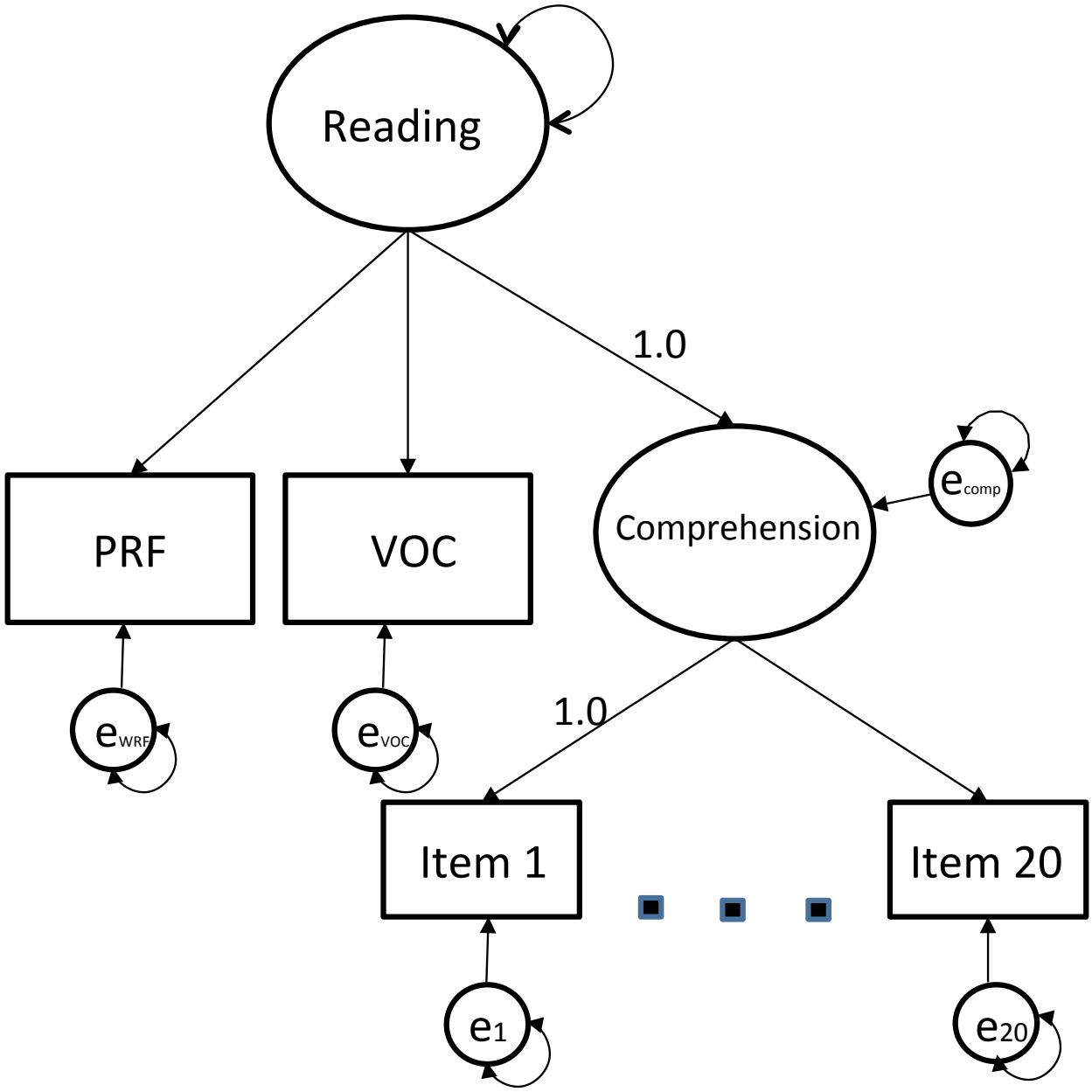


Figure 9.2. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM Grade Four-Seven measure contributes to a single, latent (unobservable), *Reading* trait. PRF = Passage Reading Fluency, VOC = Vocabulary.

References

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012a). The development of CBM vocabulary measures: Grade 2 (Technical Report No. 1209). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012b). The development of CBM vocabulary measures: Grade 3 (Technical Report No. 1210). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012c). The development of CBM vocabulary measures: Grade 4 (Technical Report No. 1211). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012d). The development of CBM vocabulary measures: Grade 5 (Technical Report No. 1212). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012e). The development of CBM vocabulary measures: Grade 6 (Technical Report No. 1213). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012f). The development of CBM vocabulary measures: Grade 7 (Technical Report No. 1214). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012g). The development of CBM vocabulary measures: Grade 8 (Technical Report No. 1215). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2013). easyCBM reading criterion related validity evidence: Grades 2-5 (technical report 1310). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Muthén, L. K., & Muthén, B. O. (1998-2007). Mplus User's Guide (Fifth ed.). Los Angeles, CA.

Wray, K., Alonzo, J., & Tindal, G. (2014). *Internal consistency of the easyCBM Measures, Grades 2-8*. Eugene, OR: University of Oregon, Behavorial Research and Teaching.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Doctoral Dissertation, University of California Los Angeles.

Chapter 10: Multiple Choice Reading Comprehension

The technical adequacy evidence gathered for the easyCBM© Multiple Choice Reading Comprehension (MCRC) measures to date suggests they are functioning largely as intended. An advanced statistical technique for scaling items (Rasch modeling) was used during the developmental process, providing information on the difficulty and functioning of each item. These results then guided test form creation to help ensure all 20 test forms were essentially equivalent in terms of difficulty and the distribution of easy and difficult items within each test form. As part of a battery of assessments, including Passage Reading Fluency and Vocabulary, the MCRC measures uniquely contribute to the variance in statewide large-scale assessments of accounted for. Evidence gathered since the measures' release indicate that the easyCBM© MCRC measures have a moderate degree of validity for measuring students' comprehension skills within a response to intervention framework, and that they are particularly relevant for students whose oral reading fluency skill and vocabulary knowledge are near or at grade level.

In what follows, we summarize the studies examining the technical adequacy evidence for the easyCBM© MCRC measures. We begin with the developmental process used, including creating alternate test forms of equivalent difficulty. We then summarize reliability evidence, including alternate form, and test-retest. Finally, we conclude by discussing validity evidence, including the relation between MCRC and relevant criterion measures, and the degree to which MCRC loads on a latent "Reading" factor. For a conceptual discussion of the theoretical purpose for any analysis conducted, please see Chapter 3. Note that this document is intended to provide a summary of evidence, not to explain in detail the different studies from which the evidence was gleaned. For a complete account of each study, the full reports are available at www.brtprojects.org/publications/technical-reports.

Measure Development

The development of the MCRC measure was an iterative process from 2005-2010. Alternate test forms for Grades 3 and 4 were created during the 2005-2006 school year, with measures piloted in spring 2006. Alternate test forms for Grades 2 and 5 were created during the 2006-2007 school year, with measures piloted in spring 2007. Alternate test forms for Grade 6-8 were created from 2008-2010 with measures piloted in spring 2010. Across all grades, measurement development began with the creation of original narratives based on common story specifications (i.e., length of stories, characters, setting and plot) provided by the lead researcher and hired professional item writer(s) who had expertise in instrument development and language arts. Story specification details provided guidance to authors of the passages (elementary and secondary school teachers and/or graduate students in the College of Education at the University or Oregon) and help reduce potential variance in test scores unrelated to the students' reading comprehension. Once the original narratives were created, professional item writer(s) wrote a number of multiple-choice items, each targeting literal, inferential and evaluative comprehension skills. Twelve items were written for each test forms in Grade 2, while 20 items were written for each test form in Grades 3-8. The number of test forms varied by grade, with 20 alternate forms available in Grades 2-5, and 11, 17, and 15 test forms available in Grades 6-8, respectively (including the 3 benchmark forms at each grade). Items on each grade-level test were ordered beginning with the easiest literal comprehension item and continuing with items of increasing difficulty, ending with the item designed to be the most difficult.

During development, the technical adequacy of the MCRC measures was evaluated in two primary ways: (a) content review of the stories and test items and (b) statistical analyses of data obtained when the measures were piloted. Content and grade-level appropriateness of each

of the MCRC measures were analyzed in four ways: (a) grade-level appropriateness of vocabulary, (b) adequate story structure for the types of items called for in the test specification documents, (c) lack of bias in language or story, and (d) formatting. The lead researcher, a professional item writer, and either an educator with appropriate grade-level teaching experience or graduate research assistants with experience in special education and assessment development conducted all content reviews. Included was a concurrent review and revision of the narrative stories as items were written.

A common-person/common item piloting design, in which test forms overlapped across all test takers, was conducted using samples of convenience. Typically these samples were recruited from large school districts in the Pacific Northwest, though for Grades 6-8 the convenience sample was comprised of teachers from across the United States who responded to a pilot invitation through the official easyCBM website.

After piloting, all items were scaled with a Rasch model (see Chapter 2 for a conceptual overview of this methodology). Item options were revised based on the Rasch modeling results. Item-level fit statistics were used to identify items that did not adequately fit the measurement model, and then the distractor-level difficulty statistic (average estimated ability of students who selected each answer option) was used to inform revisions. The difficulty and model fit of each item along with distractor analysis results are reported by test form in a series of grade-specific technical reports (Alonzo, Liu, & Tindal, 2007, 2008; Alonzo & Tindal, 2008; Park, Alonzo, & Tindal, 2011). Results from the Rasch analyses helped ensure test forms had adequate range (from easy to difficult items) to sufficiently classify students into risk categories, along with an adequate number of items on the lower tail of the distribution to detect small changes for students being progress monitored over time. The Rasch analyses also allowed us to construct

the forms so the average test difficulty was essentially equivalent across grade-level test forms.

Reliability

Split-half and top-bottom reliability studies, as well as Rasch Analyses for the easyCBM© MCRC measures, were conducted in Grades 2-7 by Lai, Irvin, Alonzo, Park, and Tindal (2012), Lai, Irvin, Park, Alonzo, and Tindal (2012), Park, Irvin, Alonzo, Lai, and Tindal (2012), Park, Irvin, Lai, Alonzo, and Tindal (2012), Irvin, Alonzo, Park, Lai, and Tindal (2012), and Irvin, Alonzo, Lai, Park, and Tindal (2012). Grades 2-5 investigated test forms 8-16, with the exception of grade 4, which dropped forms 11, 12, and 13 due to insufficient sample size (n = 24-25). Grade 6 used forms 1-6, 9, and 10, and Grade 7 used forms 6 and 10-17. The measures were given on 3 separate occasions approximately one week apart to a total of 1,032 students (n = 715 from district A and n = 317 from district B).

Split-half Reliability. Split-half reliabilities for grades 2 through 7 can be found in Table 10.1 below. Overall, the coefficients indicated moderate internal consistency across grades and test forms.

Table 10.1

Split-half Reliabilities of the easyCBM © MCRC Measures for Grade 2-7

Grade	Range	Median
2	.56-.87	.73
3	.43-.81	.63
4	.38-.67	.57
5	.29-.83	.60
6	.39-.75	.51
7	.12-.63	.45

Top-Bottom Reliability. Top-bottom reliability compares the proportion of students correctly responding to an item for students in the lower 23rd percentile to the proportion of students correctly responding to an item for students in the top 78th percentile. For Grade 2, the lower group ranged from .10-.93 on all forms but Form 14. For this form, all students got 2 of 10 items correct and the percent correct for the remaining items ranged from .13-.88. The upper percentile group got all of the questions right on form 8, 9, 11, and 12. On form 10 they all got 10 of 12 correct and ranged from .24-.95 for the remaining two items. On forms 13, 15, and 16 the students got 7 of 12 correct and ranged from .71-.96 for the remaining questions. On form 14, they got 9 of 12 correct and .77-.92 correct for the remaining 4 items.

For Grade 3, all students in the lower percentile group correctly responded to 0-3 of the 20 items across all forms. The proportion of students responding correctly to the remaining items ranged from .08-.92. Between 3-15 items were correctly responded to by all students in the higher percentile group across forms. The proportion of students correctly responding to the remaining items ranged from .25-.93. Between 0-2 items were correctly responded to by all students in the lower percentile group in Grade 4, across forms, with .11-.91 responding correctly for all remaining items. All students in the upper percentile group correctly responded to 5-9 items, with the remaining items ranging from .33-.95. All students in the Grade 5 lower percentile correctly responded to 0-1 items across forms, with the remaining items ranging from .06-.94. All students in the upper percentile group correctly responded to 0-20 items across forms, with the remaining items ranging from .25-.97. For the Grade 6 lower percentile group, the proportion of correct responses across items and forms ranged from .09 to.81. All students in the upper percentile group correctly responded to 2-8 items, while the remaining items ranged from .10 to .98. Finally, all Grade 7 students in the lower percentile group correctly responded to

0-2 items across forms, while the remaining items ranged between .06 and .96. All students in the upper percentile group correctly responded to 2-6 items across forms, with the remaining items ranging from .08 to .97.

Rasch Analyses. Results from the Rasch analyses can be found in Table 10.2. The results of the person reliabilities for grade 3-7 were lower than we would have liked, potentially because of the low number of items in each measure (20). However, the item reliabilities tended to be higher, possibly due to a more appropriate number of students participating.

Table 10.2

Rasch Analyses of the easyCBM © MCRC Measures for Grade 2-7

Grade	Person reliability range	Item reliability range	*N*
2	.00-.59	.39-.94	44-52
3	.36-.64	.79-.90	40-48
4	.56-.67	.86-.92	44
5	.25-.72	.80-.90	52-83
6	.00-.83	≥.90	49-79
7	.12-.66	.83-.96	77-109

Validity Evidence

Criterion Validity

The majority of the criterion validity evidence for the easyCBM© MCRC measures has been gathered in Grades 3-8, given that state test performance served as the criterion. Several studies provide predictive and concurrent validity evidence for easyCBM© MCRC measures. Table 10.3 provides a summary of these studies.

Table 10.3
Studies investigating the validity of easyCBM© MCRC

Study	Grade(s)	Predictive/Concurrent	Results
Tindal, Nese, and Alonzo (2009)	3-8	Both	Predictive $R^2 = .47\ to\ .68$ $sr^2 = 2\%\ to\ 10\%$ Concurrent $R^2 = .54\ to\ .65$ $sr^2 = 2\%\ to\ 8\%$
Jamgochian et al. (2010)	2	Both	Predictive $R^2 = .47$ $sr^2 = 3\%\ to\ 9\%$ Concurrent $R^2 = .55$ $sr^2 = 12\%$
Sáez et al. (2010)	3-7	Both	Predictive $R^2 = .44\ to\ .67$ $sr^2 = 1\%\ to\ 6\ \%$ Concurrent $R^2 = .55$ to .61 $sr^2 = 1\%\ to\ 9\%$
Anderson, Alonzo, & Tindal (2010)	3-8	Both	Predictive *r* = .41 to .71 $R^2 = .09\ to\ .45$ Concurrent *r* = .37 to .66 $R^2 = .11\ to\ .37$

Table 10.3 (continued)
Studies investigating the validity of easyCBM© MCRC

Study	Grade(s)	Predictive/Concurrent	Results
Anderson, Alonzo, & Tindal (2011)	3-7	Both	Predictive r = .50 to .70 $R^2 = .09\ to\ .45$ Concurrent r = .37 to .66 $R^2 = .11\ to\ .37$
Park, Anderson, Irvin, Alonzo, & Tindal (2011)	3-8	Both	Predictive AUC .79 to .86 Concurrent AUC .83 to .87
Anderson, Park, Irvin, Alonzo, & Tindal (2011)	3-8	Both	Predictive AUC .74 to .85 Concurrent AUC .76 to .84
Park, Irvin, Anderson, Alonzo, & Tindal (2011)	3-8	Cross-validation	ADC = .44
Irvin, Park, Anderson, Alonzo, & Tindal (2011)	3-8	Cross-validation	ADC = .59
Lai, Alonzo, and Tindal (2013)	2-5	Concurrent	r = .41 to .70

Note. sr^2 is the squared semi-partial correlation; ADC = the average difference in cut scores. A state test was used as the criterion in all studies with the exception of Jamgochian et al. and Lai et al., who used the Stanford-10 reading and Gates-MacGinitie reading comprehension tests, respectively.

Predictive validity. Tindal et al. (2009) explored the predictive validity of the MCRC measures by comparing fall and winter performance to spring state standardized test scores in Oregon, the Oregon Assessment of Knowledge and Skills (OAKS), for two school districts. The authors report correlations greater than or equal to .46. The MCRC measures were significant predictors of OAKS reading test scores in all cases ($p < .01$). The standardized beta weights between MCRC measures and OAKS reading assessment results ranged from .16 to .40. The variance accounted for by the multiple regression model, which included gender, ethnicity,

economic disadvantage, special education, Title 1, and easyCBM© Passage Reading Fluency (PRF) scores in District 1 and gender, historically high/low achieving, economic disadvantage, special education, PRF, and easyCBM© Vocabulary (VOC) scores in District 2, ranged from .47 to .68. The unique variance in OAKS scores accounted for by MCRC ranged from 2%-10% across both districts. While VOC generally explained the most unique variance, MCRC measures consistently explained more unique variance than all of the demographic variables. The authors caution that the correlation between easyCBM© predictors was high, which indicates multicollinearity in the regression model that may have affected the estimates.

Jamgochian, et al. (2010) conducted a similar study at Grade 2, using the SAT-10 as the spring criterion for fall and winter MCRC performance. The authors also explored the predictive utility of MCRC growth, using a two-level hierarchical linear growth model (HLM), with level one representing time and level two representing the student. The sample included 1,696 students in the fall and 2,039 students in winter. A multiple regression analysis was conducted using fall and winter measures, which included easyCBM© Word Reading Fluency (WRF) and PRF scores, in addition to MCRC. The model predicted 47% of the variance in spring SAT-10 scores in both fall and winter. In a linear regression model, the fall MCRC uniquely accounted for 0.5% of the variance in SAT-10 scores; winter MCRC uniquely accounted for 0.6% of the variance. The fall 2009 MCRC measure uniquely accounted for approximately 3% of the variance in SAT-10 scores ($p < .01$). The winter 2010 MCRC measure uniquely accounted for approximately 9% of the variance in SAT-10 scores ($p < .01$). The average MCRC growth rate for Grade 2 students was calculated by quartiles of normative performance on the fall MCRC measure (i.e., four separate samples). Students in the first and second quartiles made moderate and positive growth, with standardized coefficients of .68 and .67, respectively. It was not possible to estimate the

third quartile due to convergence errors. Rate of growth for the top quartile was low, but positive, with a standardized coefficient of .18.

Sáez et al. (2010), conducted three analyses exploring the predictive validity of the MCRC measures: (a) bivariate correlations between the fall and winter MCRC measures and spring OAKS performance, (b) within-year MCRC growth as a predictor of OAKS performance with regression, and (c) diagnostic efficiency of the fall and winter measures predicting meeting proficiency on the OAKS. The study included approximately 2,000 Oregon public school students per grade level. The analyses included results that were disaggregated by gender, ethnicity, and special education status; however, only overall results are reported here. The Pearson's bivariate correlation coefficient for fall MCRC with spring OAKS ranged from .55 to .67. The winter MCRC correlations ranged from .44 to .61. In a simple linear regression, fall MCRC scores accounted for 30% to 45% of the variance in spring 2010 OAKS reading scores; winter MCRC scores accounted for 19% to 37%. In a multiple regression analysis that included WRF (Grade 3 only), PRF, and VOC, the fall MCRC scores uniquely accounted for 1% to 6% of the variance in OAKS reading; winter MCRC scores uniquely accounted for 2% to 9%. In slope comparisons for MCRC, the 1st through 3rd quartiles had moderate standardized coefficients, ranging from .48 to .63. In other words, a one standard deviation increase in MCRC growth corresponded to, on average, approximately half a standard deviation increase in OAKS performance. Students' growth generally had weak diagnostic utility, with the area under the receiver operating characteristic curve (AUC) ranging from .00 to .97. Interestingly, the diagnostic utility of students' Grade 3 growth (.78 to .97) was much greater than other grades' (.00 to .43), where the results suggest alternate measures might be more appropriate if being used in a growth model. The overall low diagnostic utility was not surprising, given the relatively

small amount of growth students exhibited, on average, as well as the low variance among students' growth.

Anderson et al. (2010) found bivariate correlations between fall MCRC and the spring state test used in Washington State, the Measures of Student Progress (MSP), ranged from .52 to .65 across Grades 3-8, while winter measures correlated from .41 to .71. Anderson et al. (2011) found correlations between the fall MCRC measure and the OAKS in Grades 3-7 to range from .50 to .70, while the winter measure correlated from .52 to .67. Linear regression analyses that included fall MCRC measures accounted for 16%-45% of the variance in the MSP, while winter accounted for 9%-30% (Anderson et al., 2010). In Oregon the fall measures accounted for 16%-45% in OAKS, while winter accounted for 9%-30%. In both studies, the measures used in the lower grades generally accounted for more variance than the measures used in the upper grades.

Park, Anderson et al. (2011) reviewed the diagnostic efficiency of the fall and winter easyCBM© MCRC measures compared to the spring OAKS status of meeting/not meeting state performance expectations in Grades 3-8, while Anderson et al. (2011) did the same for the MSP. The OAKS sample included approximately 2,000 students per grade, while the MSP sample included approximately 1,200. Given the optimal cut score, the fall and winter results for sensitivity ranged from .70 to .86 for OAKS, and .68 to .85 for the MSP. Specificity ranged from .66 to .79 for OAKS and .57 to .76 for the MSP. The AUC for fall and winter ranged from .79 to .86 for OAKS and .74 to .85 for the MSP. A perfect AUC statistic is 1.0; an AUC that is no better than chance is .50. These results suggest that easyCBM MCRC measures are good to excellent discriminators of future OAKS performance.

Park, Irvin, Anderson, Alonzo, & Tindal (2011) conducted a cross-validation of the Park, Anderson et al. (2011) study using easyCBM© fall and winter MCRC measures to predict

meeting/not meeting proficiency on OAKS in Grades 3-8. Irvin, Park, Anderson, Alonzo, and Tindal (2011) conducted a similar study with the MSP, cross-validating the resuls of the Park, Anderson et al. (2011) study. In each cross-validation study, the authors randomly split a large sample into two groups to examine the stability of the cut scores across the randomly-selected groups. In both reports, the authors note that the cut scores appear quite stable overall, with the average difference in cut scores between groups of .44 for OAKS and .59 for the MSP. The 95% confidence intervals for AUC statistics overlapped between the groups for MCRC in each study. The consistency between the optimal cut scores combined with the lack of significant differences between AUC statistics in all measurement occasions and grades provide strong evidence for the cut scores derived. See the full reports for specific cut scores.

Concurrent validity. Tindal et al. (2010) explored the concurrent validity of the MCRC measures by comparing spring performance to spring state standardized test scores for two school districts in Oregon. Pearson correlations between the MCRC measures and the OAKS were > .50. The MCRC results significantly predicted state reading test scores in all cases, and all results reported here were significant ($p < .01$). The standardized beta weights between MCRC measures and OAKS reading assessment results ranged from .14 to 24. The variance accounted for by the multiple regression model, which included gender, ethnicity, economic disadvantage, special education, PRF, and VOC scores in District 1 and gender, historically high/low achieving, economic disadvantage, special education, PRF, and VOC in District 2, ranged from .54 to .65. The unique variance in OAKS scores explained by MCRC measures ranged from 2%-8% across both districts. While VOC generally explained the most unique variance, MCRC measures consistently explained more unique variance than all of the demographic variables. The authors caution that the correlation between easyCBM© predictors

was high, which indicated multicollinearity in the regression model that may have affected the coefficient estimates.

Jamgochian et al. (2010) explored the concurrent validity of the MCRC measures in Grade 2 by correlating MCRC scores with the SAT-10 in the spring. The sample included 2,154 students. Regression analyses were conducted for each of the measures separately and combined by season. In a multiple regression model that included WRF and PRF, the model accounted for 55% of the variance in SAT-10 scores, with the MCRC measures accounting for the most unique variance, approximately 12%, which was significant ($p < .01$).

Sáez et al. (2010) also explored the concurrent validity of the MCRC measures. The spring 2010 easyCBM© MCRC measure was compared to the spring 2010 OAKS. The analysis included results disaggregated by gender, ethnicity, and program status; however, only overall results are reported here. The spring 2010 MCRC measures correlated between .55 and .61 with the spring OAKS. In a simple linear regression model, spring MCRC scores accounted for 28% to 66% of the variance in spring 2010 OAKS reading scores. In a multiple regression model, MCRC scores uniquely accounted for 1% to 9% of the variance in OAKS, along with WRF (Grade 3 only), PRF, and VOC.

Anderson et al. (2010) determined that concurrent correlations between the spring MCRC measure and the spring MSP in grades 3-8 ranged from .37 to .66, while Anderson et al. (2011) found the correlations to range from .52 to .68 for OAKS across Grades 3-7. A linear regression analysis that included spring MCRC measures accounted for 11-37% of the variance in both MSP and OAKS scores (Anderson et al. 2010, 2011). Lower-grade MCRC results generally accounted for MSP outcomes better than upper-grade MCRC results.

Park, Anderson et al. (2011) reviewed the diagnostic efficiency of the spring easyCBM© MCRC measures compared to meeting proficiency on OAKS in Grades 3-8, while Anderson, Park et al. did the same with the MSP. The Oregon sample included approximately 2,000 students per grade level, while the Washington sample included approximately 1,200. Given the optimal cut score, the spring results from sensitivity range from .71 to .89 for OAKS and .69 to .79 for the MSP. The spring results for specificity range from .70 to .80 for OAKS and .64 to .77 for the MSP. The AUC statistic ranged from .83 to .87 for OAKS and .76 to .84 for the MSP. These results suggest that easyCBM© MCRC measures are *good* to *excellent* discriminators of whether students will reach proficiency on OAKS.

Finally, Lai et al. (2013) compared the easyCBM© MCRC measures to the Gates-MacGinitie Reading Comprehension measures concurrently, with a sample of approximately 200 students per grade. The easyCBM© MCRC measures showed moderate correlations with the Gates-MacGinitie across grades, with correlations ranging from .41 to.70. The authors attribute these differences to assessment targets, as the easyCBM© MCRC assessments target students' literal, inferential, and evaluative comprehension skills, while the Gates-MacGinitie Reading Comprehension measures target only literal comprehension.

Construct Validity

To gather construct validity evidence for the easyCBM© MCRC measures, a series of confirmatory factor analysis (CFA) models were conducted using extant data from Grades 2-7. In all CFA models, we hypothesized a one-factor structure, where a general *Reading* construct was measured. For the Grade 2 data, *Reading* was measured by Word Reading Fluency (WRF), Passage Reading Fluency (PRF), and MCRC. The MCRC factor indicator was itself a latent factor, with item responses serving as the factor indicators. For Grade 3, a similar model was

estimated, but with a vocabulary measure included as a factor indicator. For Grades 4-7, *Reading* was measured by PRF, VOC and the *comprehension* latent variable. The hypothesized models are displayed in Figures 10.1-10.3.

For the CFAs across all grades and time points, one of the factor loadings for comprehension was constrained to be 1.0, along with the factor loadings for reading and comprehension latent constructs, to identify the model. All other factor loadings and variances were freely estimated. A weighted least squares estimator was used with the *Mplus* software (WLSMV; Muthén & Muthén, 1998-2007). Model fit was evaluated using the Tucker Lewis Index (TLI), Comparative Fit Index (CFI), and Root-Mean Square Error of Approximation (RMSEA). In particular, with binary and continuous model variables, CFI and TLI values ≥ 0.95, and RMSEA values ≤ 0.05 were considered indications of good model fit to the data (Yu, 2002).

Separate CFA models were conducted for each seasonal benchmark assessment for Grade 2, while only fall and spring assessments for Grades 3-7 were analyzed. The data were evaluated for their fit to the hypothesized, one-factor model. Overall, the model fit statistics suggested the data had good fit to the model for all seasons across all grades (see Table 10.4 below). The latent *Comprehension* factor had moderately high to high factor loadings (.70-.91) on *Reading*, suggesting that students' reading ability was a strong predictor of their performance on the easyCBM© MCRC measure.

Table 10.4
Confirmatory Factor Analysis Results

Grade	Time Point	*N*	Fit Statistics			Factor Loadings			
						Reading			
			CFI	TLI	RMSEA	WRF	PRF	VOC	Comprehension
2	F	1701	0.997	0.996	0.033	0.910	0.958	-	0.906
	W	1959	0.997	0.996	0.03	0.945	0.983	-	0.835
	S	1793	0.999	0.999	0.017	0.918	0.988	-	0.775
3	F	782	0.994	0.993	0.022	0.947	0.961	0.822	0.843
	S	876	0.991	0.990	0.027	0.881	0.995	0.815	0.700
4	F	1887	0.979	0.977	0.026	-	0.852	0.831	0.895
	S	2048	0.984	0.982	0.022	-	0.779	0.768	0.803
5	F	2037	0.981	0.979	0.022	-	0.856	0.753	0.862
	S	2123	0.980	0.978	0.023	-	0.781	0.712	0.824
6	F	1240	0.961	0.957	0.025	-	0.791	0.742	0.850
	S	1148	0.987	0.986	0.016	-	0.790	0.715	0.844
7	F	791	0.962	0.957	0.026	-	0.769	0.690	0.842
	S	848	0.978	0.975	0.018	-	0.707	0.521	0.904

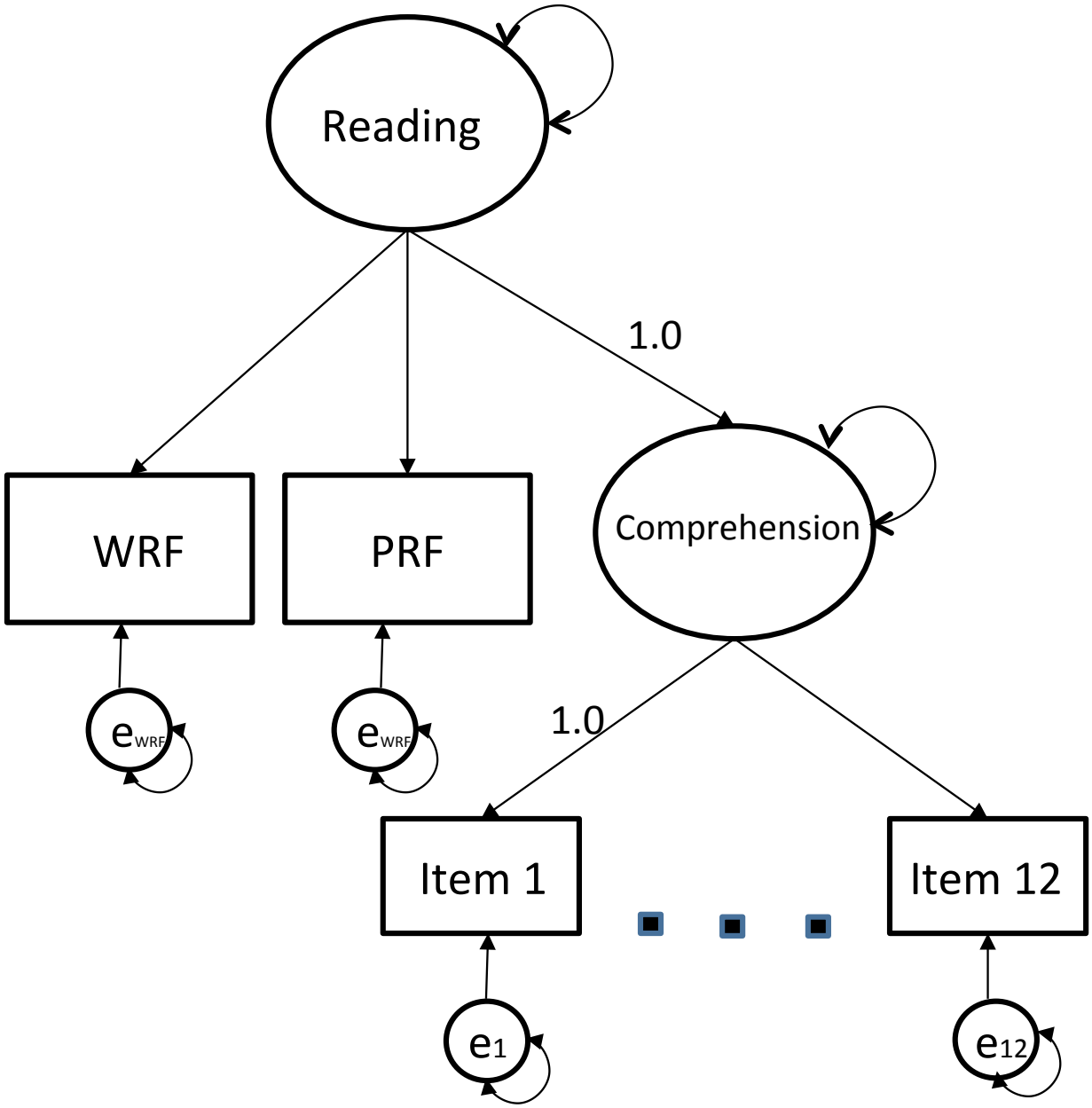


Figure 10.1a. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM Grade Two measure contributes to a single, latent (unobservable), *Reading* trait. WRF = Word Reading Fluency, PRF = Passage Reading Fluency.

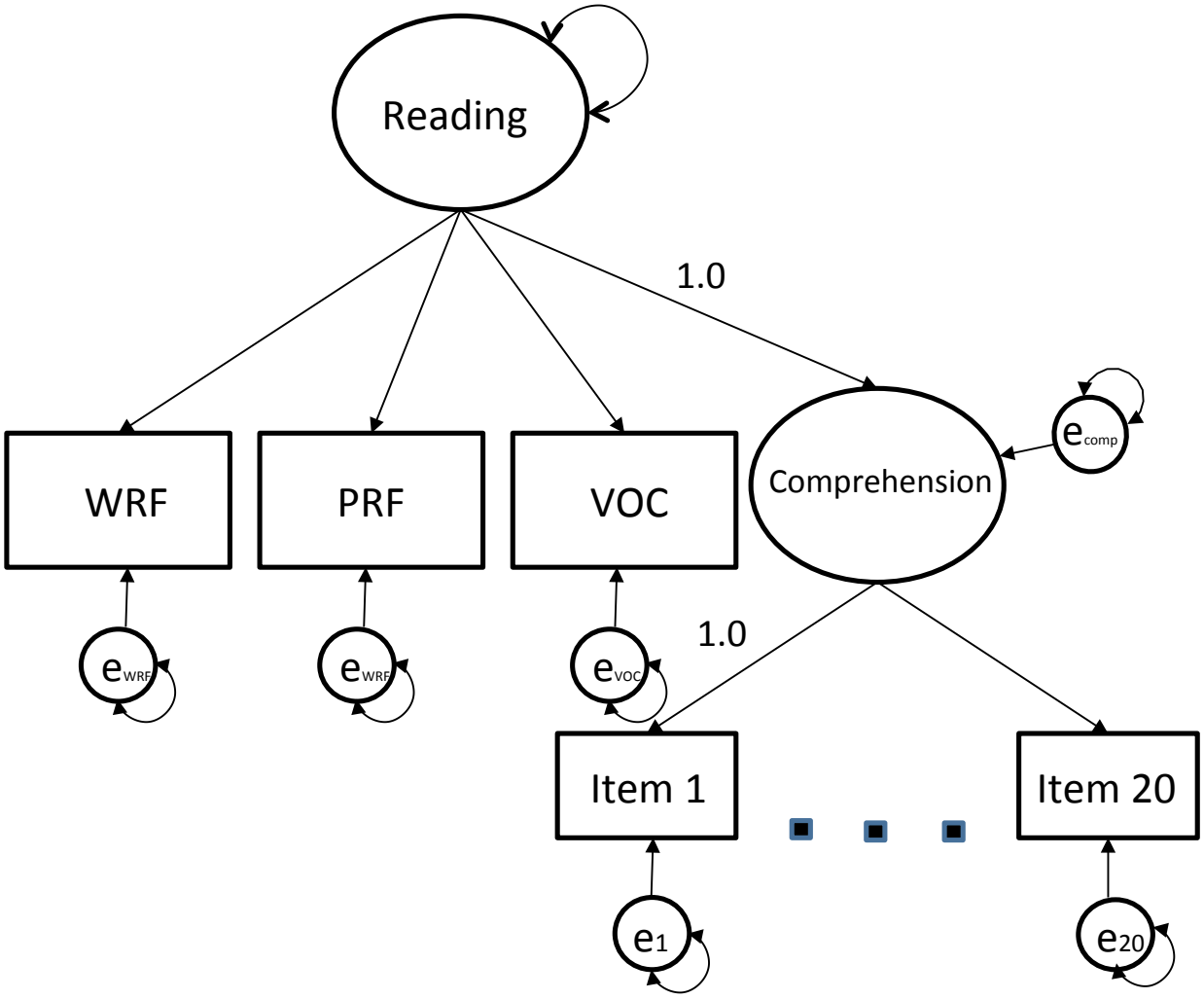


Figure 10.2. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM Grade Three measure contributes to a single, latent (unobservable), *Reading* trait. WRF = Word Reading Fluency, PRF = Passage Reading Fluency, VOC = Vocabulary.

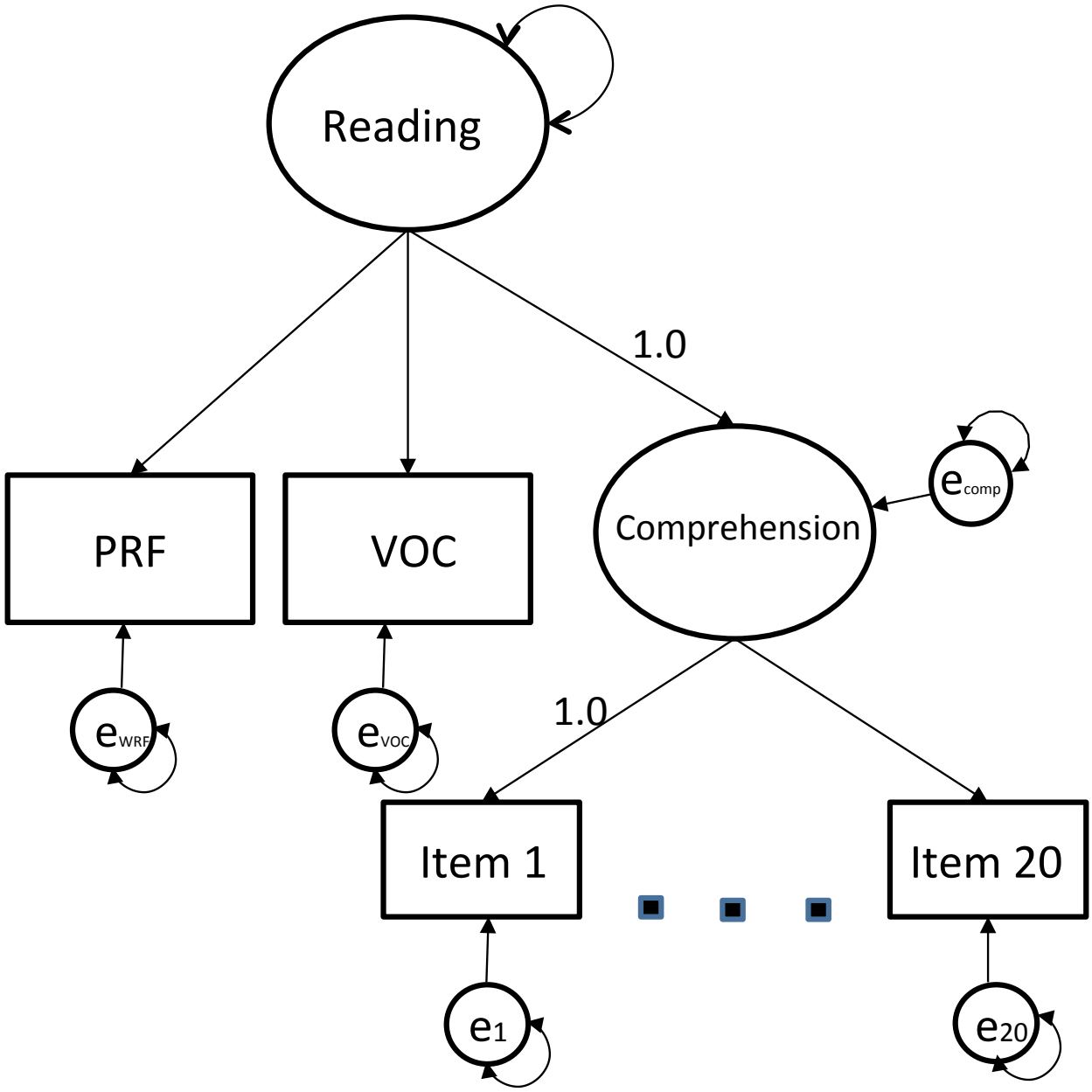


Figure 10.3. Theoretical One-Factor Confirmatory Factor Analysis. The model hypothesizes that each easyCBM Grade Four-Seven measure contributes to a single, latent (unobservable), *Reading* trait. PRF = Passage Reading Fluency, VOC = Vocabulary.

References

Alonzo, J., Liu, K., & Tindal, G. (2007). Examining the technical adequacy of reading comprehension measures in a progress monitoring assessment system (Technical Report No. 41). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Liu, K., & Tindal, G. (2008). Examining the technical adequacy of second-grade reading comprehension measures in a progress monitoring assessment system (Technical Report No. 0808). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2008). Examining the technical adequacy of fifth-grade reading comprehension measures in a progress monitoring assessment system (Technical Report No. 0807). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Alonzo, J., Lai, C. F., Park, B. J., & Tindal, G. (2012). Analyzing the reliability of the easyCBM reading comprehension measures: Grade 7 (technical report 1206). Eugegne, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Alonzo, J., Park, B. J., Lai, C. F., & Tindal, G. (2012). Analyzing the reliability of the easyCBM reading comprehension measures: Grade 6 (technical report 1205). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Jamgochian, E. M., Park, B. J., Nese, J. F. T., Lai, C. F., Sáez, L., Anderson, D., et al. (2010). Technical adequacy of the easyCBM grade 2 reading measures (technical report 1004). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2013). easyCBM reading criterion related validity evidence: Grades 2-5 (technical report 1310). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Irvin, P. S., Alonzo, J., Park, B. J., & Tindal, G. (2012). Analyzing the reliability of the easyCBM reading comprehension measures: Grade 2 (technical teport 1201). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012). Analyzing the reliability of the easyCBM reading comprehension measures: Grade 3 (technical report 1202). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Muthén, L. K., & Muthén, B. O. (1998-2007). Mplus User's Guide (Fifth ed.). Los Angeles, CA.

Park, B. J., Alonzo, J., & Tindal, G. (2011). The development and technical adequacy of seventh-grade reading comprehension measures in a progress monitoring assessment system (Technical Report No. 1102). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Irvin, P. S., Alonzo, J., Lai, C. F., & Tindal, G. (2012). Analyzing the reliability of the easyCBM reading comprehension measures: Grade 4 (technical report 1203). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Irvin, P. S., Lai, C. F., Alonzo, J., & Tindal, G. (2012). Analyzing the reliability of the easyCBM reading comprehension measures: Grade 5 (technical report 1204). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Sáez, L., Park, B. J., Nese, J. F. T., Jamgochian, E. M., Lai, C. F., Anderson, D., et al. (2010). Technical adequacy of the easyCBM reading measures (Grades 3-7), 2009-2010 version (technical report 1005). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Tindal, G., Nese, J. F., & Alonzo, J. (2009). Criterion-related evidence using easyCBM® reading measures and student demographics to predict state test performance in grades 3-8

(technical report 0910). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Doctoral Dissertation, University of California Los Angeles.

Chapter 11: CCSS Reading

The technical adequacy evidence gathered for the easyCBM© CCSS Reading measures to date suggests they are functioning largely as intended. An advanced statistical technique for scaling items (Rasch modeling) was used during the developmental process, providing information on the difficulty and functioning of items. These results then guided test form creation to help ensure all 20 test forms in each grade were essentially equivalent in terms of difficulty and the distribution of easy and difficult items within each test form Alonzo, Park, and Tindal (2012a, 2012b, 2012c, 2012d, 2012e, 2012f). Guerreiro, Alonzo, and Tindal (2014) found the reliability to be quite high, with Cronbach's alpha ranging from .83 to .90 across grades, and split half reliabilities ranging from .73 to .86. Alonzo, Park, and Tindal (2013a, 2013b) conducted confirmatory factor analyses (CFA) and found a model including both the CCSS Reading and MCRC measures as the same construct (identified as 'comprehension' and contrasted with early literacy and vocabulary measures) fit the data well.

In what follows, we summarize the studies cited above examining the technical adequacy evidence for the easyCBM© CCSS measures. For a conceptual discussion of the theoretical purpose for any analysis conducted, please see Chapter 3. Note that this document is intended to provide a summary of evidence, not to explain in detail the different studies from which the evidence was gleaned. For a complete account of each study, the full reports are available at www.brtprojects.org/publications/technical-reports.

Measure Development

A team of educators with experience instructing students struggling in reading wrote all items for the CCSS Reading tests. A research team with training in measurement and assessment development reviewed items prior to piloting in fall 2011. Piloting was conducted using a sample

of convenience from voluntary teacher sign up through the easyCBM© assessment system. Students from eight states (Oregon, Washington, Montana, Florida, Texas, Illinois, California, and Wisconsin) participated in the pilot study in December 2011. Students responded to 25 items clustered evenly into five subtests. Subtests were randomly assigned to students to reduce the potential for group effects biasing the results. Approximately 900 items were piloted in each of grades 3-8.

During scaling, each subtest of five items was collapsed into a raw sum score ranging from 0 (no items correct) to 5 (all items correct). Each raw sum score was then treated as an item. This process helped guard against "testlet" effects that can bias scaling results. All raw sum items were scaled with a partial credit Rasch model. Students responded to common items during piloting so all items could be calibrated concurrently on the same scale within each grade. The difficulty and model fit of each raw sum item are reported by subtest and operational benchmark in a series of grade-specific technical reports by Alonzo et al. (2012a, 2012b, 2012c, 2012d, 2012e, 2012f). Results from the Rasch analyses helped ensure test forms had adequate range (from easy to difficult items) to sufficiently classify students into risk categories, along with an adequate number of items on the lower tail of the distribution to detect small changes for students being progress monitored over time. The Rasch analyses also allowed us to construct the forms so the average item (subtest) difficulty was essentially equivalent across grade-level test forms.

Reliability

Guerreiro et al. (2014) investigated the internal consistency of the easyCBM© CCSS Reading measures in Grades 3-8 by evaluating Cronbach's alpha at both the total test and split-half levels. Total test reliability ranged from .83 to .90 across grades and seasons. Split-half

reliability ranged from .73 to .86, while the correlation between split-halves ranged from .56 to 74. These results suggest the measures consistently measured the same underlying trait. Reliabilities for each grade and season are reported in Table 11.1 below.

Table 11.1

Internal Reliability: CCSS Reading

Grade/time	Cronbach's alpha	Split-half reliability		
		1st half	2nd half	Correlation
3/F	.90	.81	.86	.74
3/W	.87	.73	.82	.69
4/F	.88	.77	.83	.67
4/W	.87	.74	.81	.67
5/F	.84	.76	.75	.60
5/W	.85	.81	.73	.59
6/F	.89	.78	.85	.67
6/W	.87	.74	.85	.65
7/F	.83	.75	.75	.56
7/W	.86	.73	.80	.65
8/F	.88	.75	.84	.68
8/W	.88	.76	.81	.66

Validity Evidence

Criterion Validity

Lai, Alonzo, and Tindal (2013) compared the easyCBM© CCSS Reading measures to the Gates-MacGinitie Reading Comprehension assessments in Grades 2-5. The sample included approximately 100-200 students per grade from ten schools in one school district in Oregon. Correlations between the easyCBM CCSS Reading measure and the Gates-MacGinitie Reading Comprehension measure varied across grades, with values ranging from .41-.71. The authors attribute the low to moderate correlations to differences in assessment targets, as the easyCBM© CCSS Reading measures target the Common Core State Standards for literature, informational text, and literacy in science and technical subjects, while the Gates-Macginitie measure targets literal comprehension. The authors also note missing data concerns for this study, particularly at Grade 3, related to an unexpectedly severe flu season that resulted in multiple absences during the study data collection window.

Construct Validity

To gather construct validity evidence for the easyCBM© CCSS Reading measures, Alonzo et al. (2013a, 2013b) conducted confirmatory factor analysis (CFA) to examine the internal structures of the Grades K-5 easyCBM© CCSS reading assessments.

For Grades K-2, Alonzo and colleagues (2013a) compared four CFA models, as represented in Figures 11.1-11.4. The authors were primarily interested in potential testlest effects. Note that "Comp Part 1" represents easyCBM© CCSS Reading measures, while "Comp Part 2" represents the easyCBM© MCRC measures. Model comparison results suggested that a three factor model without a testlet effect fit the data best (Figure 11.3). This findings suggest that the easyCBM© measures assess three constructs of reading: (1) early literacy, as measured

by phoneme segmenting, letter names, and letter sounds, (2) fluency, as measured by word and passage reading, and (3) comprehension, as measured by both the CCSS and MCRC comprehension measures. For a full description of the model fit criteria, see Alonzo, Park and Tindal (2013a).

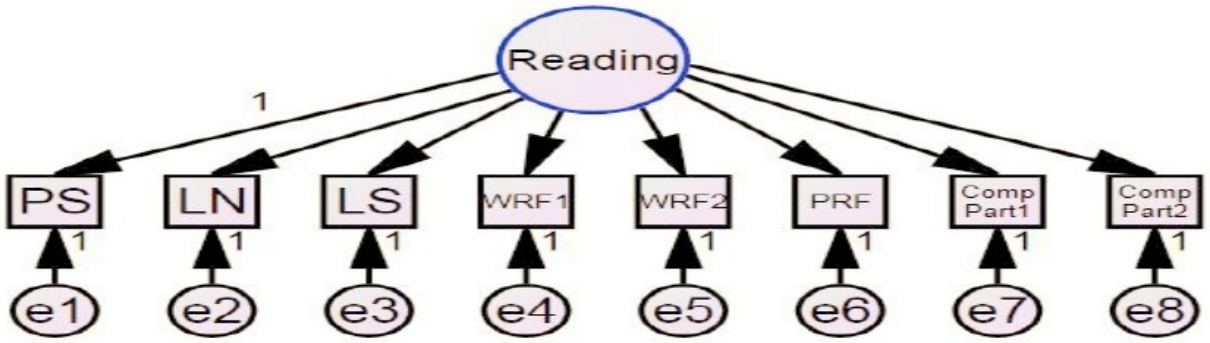


Figure 11.1. One factor CFA model without testlet effect for grades K-2.

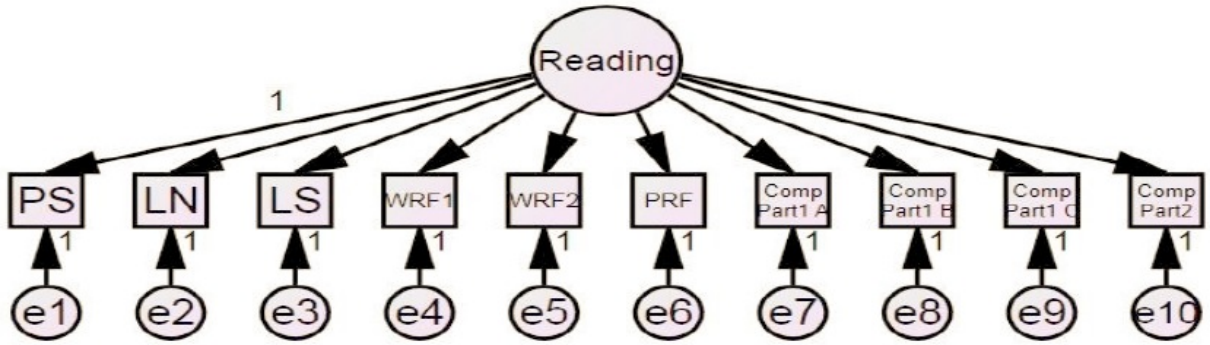


Figure 11.2. One factor CFA model with testlet effect for grades K-2. Note that the CCSS Reading measures have been disaggregated into its three sub-domains, labeled as a, b, and c.

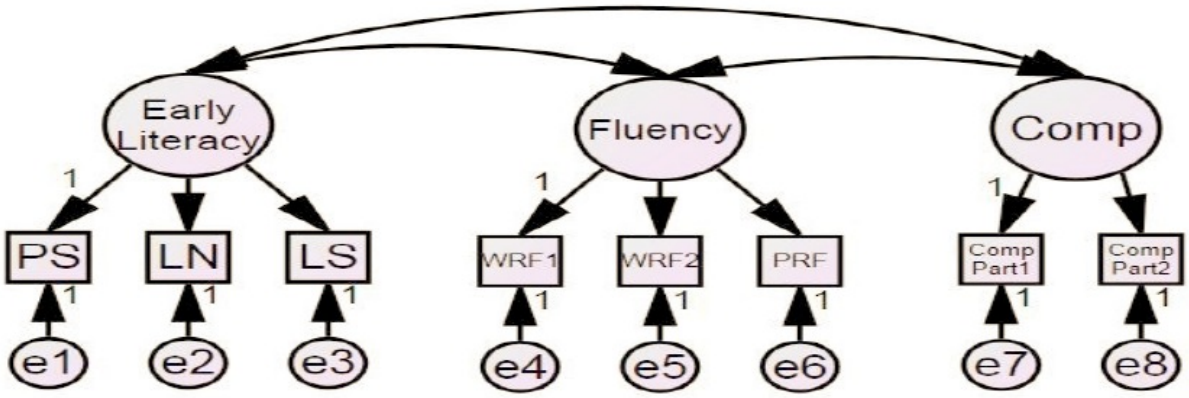


Figure 11.3. Three factor CFA model without testlet effect for Grades K-2. Note that this model fit the data best.

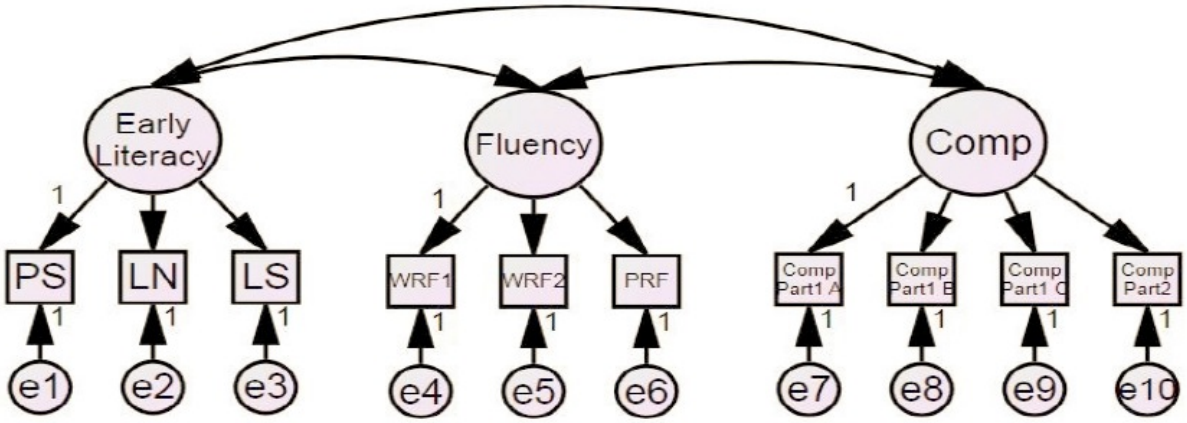


Figure 11.4. Three factor CFA model with testlet effect for Grades K-2.

Alonzo et al. (2013a) conducted similar CFA studies to evaluate the underlying structure for Grades 3-5. Four factor structures were again examined, as represented in Figures 11.5-11.8. For the Grade 3 and 4 data, the one-factor model without testlet effects fit the data best. The one factor models suggest that the easyCBM© measures assess one general "reading" construct, as measured by word reading fluency, passage reading fluency, and reading comprehension measures. For grade five, the two-factor model without a testlet effect fit the data best. The two-factor model suggests that there are two constructs: (a) fluency, as measured by word reading and passage reading fluency measures, and (b) reading comprehension, as measured by the

easyCBM© CCSS and MCRC measures. For a full description of the model fit criteria, see Alonzo, Park and Tindal (2013a).

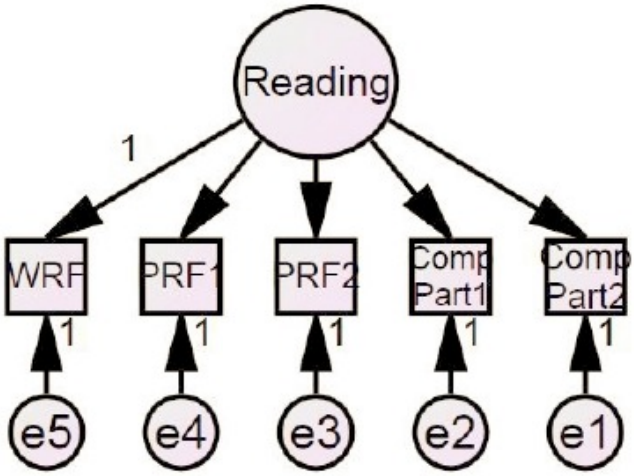


Figure 11.5. One factor CFA model without testlet effect for grades 3-5.

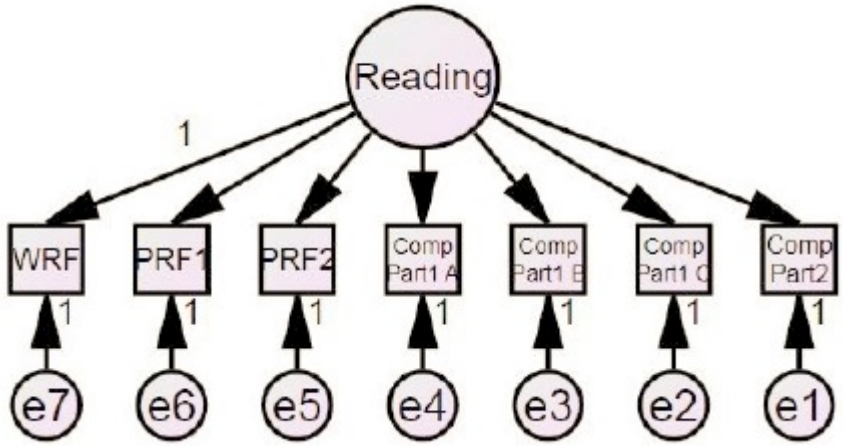


Figure 11.6. One factor CFA model with testlet effect for grades 3-5.

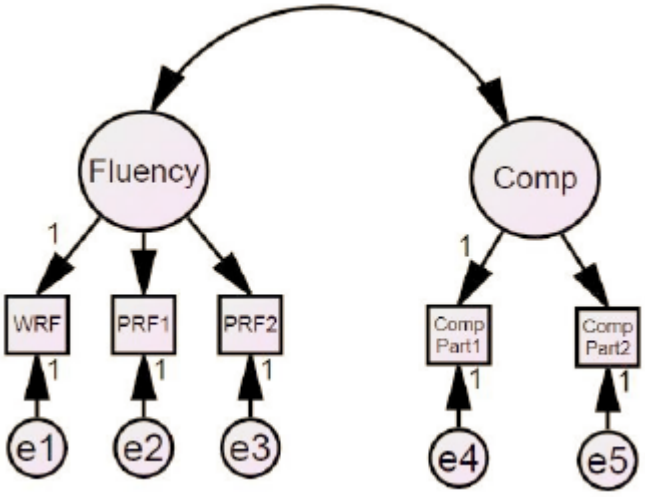


Figure 11.7. Two factor CFA model without testlet effect for grades 3-5.

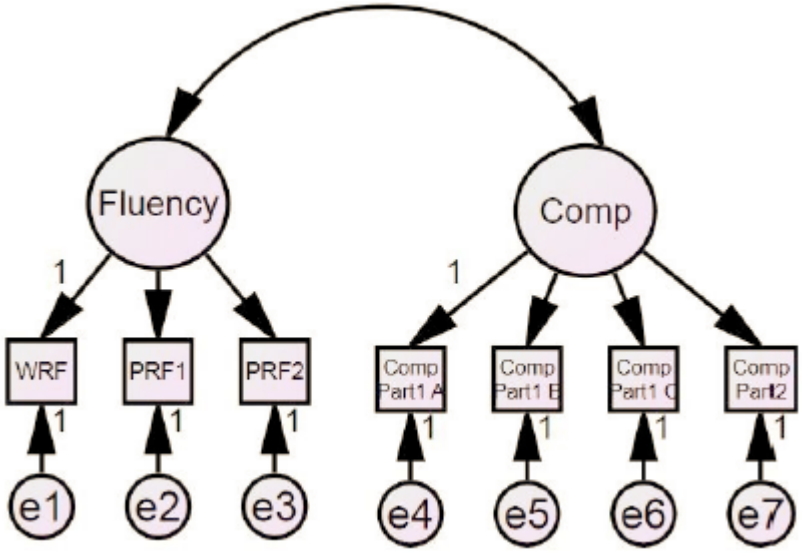


Figure 11.8. Two factor CFA model with testlet effect for grades 3-5.

In another study, the same authors (2013b) conducted a similar CFA study examining only the Grades 3-5 easyCBM© CCSS measures, without evaluating testlet effects and the fluency measures. The CCSS measures include item prompts based on genre of item prompt: *Read to Perform a Task*, *Informational Text*, and *Short Literary Text* targeting literal comprehension. Model fit for one-factor, two-factor, and three-factor models were compared. In the one-factor model, the authors hypothesized that the measures would load on a single factor ('literal reading comprehension') regardless of the passage genre. In the two-factor models, it

was hypothesized that that the measures would load on two factors, based on genre of the passage used in the prompt (i.e. *Read to Perform a Task* as Factor 1, and *Informational Text*/*Short Literary Text* as Factor 2, or *Read to Perform a Task/ Short Literary Text* as Factor 1 and *Informational Text* as Factor 2 or *Short Literary Text* as Factor 1 and *Read to Perform a Task/Informational Text* as Factor 2). The three-factor model constrained the factor structure by genre of item prompt: *Informational Text*, *Short Literary Text*, and *Read to Perform a Task*.

Results indicated that the one-factor model resulted in the best fit, although at least one of the two-factor models at each grade level also produced reasonable fit statistics. The Grade 3 and 4 CCSS measures had moderate to strong "loadings" on the latent reading factor (.60-.80 for grade three, .50-.70 for grade four), suggesting that students' literal reading comprehension ability moderately predicted their performance on the easyCBM© CCSS measure. The grade five measures had low to moderately strong "loadings" on the latent factor, ranging from .10-.70, with a median of 0.58. For a full description of the model fit statistics and model fit criteria, see Alonzo, Park and Tindal (2012b).

References

Alonzo, J., Park, B. J., & Tindal, G. (2012a). The development of the easyCBM CCSS Reading assessments: Grade 3 (Technical Report No. 1221). Eugene, OR: University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012b). The development of the easyCBM CCSS Reading assessments: Grade 4 (Technical Report No. 1222). Eugene, OR: University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012c). The development of the easyCBM CCSS Reading assessments: Grade 5 (Technical Report No. 1223). Eugene, OR: University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012d). The development of the easyCBM CCSS Reading assessments: Grade 6 (Technical Report No. 1224). Eugene, OR: University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012e). The development of the easyCBM CCSS Reading assessments: Grade 7 (Technical Report No. 1225). Eugene, OR: University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2012f). The development of the easyCBM CCSS Reading assessments: Grade 8 (Technical Report No. 1226). Eugene, OR: University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2013a). An examination of the internal structures of the easyCBM CCSS reading measures (technical report 1304). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Park, B. J., & Tindal, G. (2013b). An examination of the internal structures of the gr. K-5 easyCBM CCSS reading measures: A construct validity study (technical report 1305). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Guerreiro, M., Alonzo, J., & Tindal, G. (2014). *Internal Consistency of the easyCBM CCSS Reading Measures: Grades 3-8* (Technical Report No. 1406). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2013). easyCBM reading criterion related validity evidence: Grades 2-5 (technical report 1310). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Chapter 12: NCTM Math Measures

The technical adequacy evidence gathered for the easyCBM© NCTM Math measures to date suggests they are functioning largely as intended. An advanced statistical technique for scaling items (Rasch modeling) was used during the developmental process, providing information on the difficulty and functioning of each item. These results then guided test form creation to help ensure all test forms within a grade were essentially equivalent in terms of difficulty and the distribution of easy and difficult items within each test form. Nese, Lai, Anderson, Jamgochian et al. (2010) showed that the NCTM items were reasonably well-aligned with the National Council of Teachers of Mathematics (NCTM) Focal Point Standards. The internal consistency of the measures has been shown to be quite strong across multiple studies (Anderson, Lai et al., 2010; Nese, Lai, Anderson, Jamgochian et al., 2010). Multiple studies have also shown the measures to have a strong relation with external criteria (e.g., state test performance) both predictively and concurrently (see Table 12.5). Finally, Rasch analyses conducted by Anderson, Lai et al. (2010) and factor analyses conducted by Nese, Lai, Anderson, Jamgochian et al. (2010) suggest the items consistently measured a unidimensional *Math* construct. In sum, these studies suggest the technical adequacy evidence for the easyCBM© NCTM Math measures is quite strong.

In what follows, we summarize the technical adequacy evidence for the easyCBM© NCTM Math measures. We begin with the developmental process used, including creating alternate test forms of equivalent difficulty, as well as the alignment between the measures and the NCTM Focal Point Standards. We also report the results of an exploratory study of the alignment between the NCTM measures and the Common Core State Standards. We then summarize the reliability evidence collected to date, and conclude by discussing criterion and

construct validity evidence. For a conceptual discussion of the theoretical purpose for any analysis conducted, please see Chapter 3. Note that this document is intended to provide a summary of evidence, not to explain in detail the different studies from which the evidence was gleaned. For a complete account of each study, the full reports are available at www.brtprojects.org/publications/technical-reports.

Measure Development

The development of the easyCBM© math measures aligned to the NCTM Focal Points was an iterative process over 2007-2008. In each of Grades K-8, mathematics item writing for the NCTM Math measures began in fall 2007. Eight item writers were recruited from across the state of Oregon who had experience in teaching and assessing students in mathematics; a majority of the item writers had worked extensively with students in special education programs. It is important to note that the NCTM Math measures were designed to assess both the general education student population and the "2% student population" – the 20% of students with identified disabilities who are assessed using grade-level content standards but with modified academic achievement standards to reduce required cognitive complexity (U.S. Department of Education, 2005). Such students typically demonstrate very low academic performance, receive special education services, and would also likely receive significant support in general education contexts. Item writers were trained and given specific guidelines on how to approach and complete the item writing process, including two major points that emphasized: (a) the importance of writing math items requiring reduced cognitive complexity, while (b) preserving the integrity of the math items for use within standards-based curricula by connecting them to grade-level content standards (grade-level NCTM Focal Point Standards).

Six University of Oregon researchers with experience in assessment development and test

item creation conducted a formal two-month long item review process. During this item review process, and using universal design for assessment as a primary guide (Johnstone, Altman, & Thurlow, 2006), the six researchers studied specific item characteristics individually and as a group, including general clarity and alignment with the standards, formatting, wording, and appropriateness of answer choices. Researchers finalized the item bank in August 2008, at which point all items were uploaded into the easyCBM© online system and paired with appropriate graphics created by a professional graphics artist working in collaboration with the item writers.

Piloting took place over approximately one month in November-December 2008 with approximately 2,800 students in each of Grades K-8. Data from each grade level were analyzed separately with all items scaled with a Rasch model (see Chapter 2 for a conceptual overview of this methodology). A common-person/item equating plan was used, whereby the first 20 items were randomly assigned to each student, while the final 5 items (representing a range of difficulty and all NCTM Focal Points within a given grade) were held constant across all students. We analyzed, on average, 1,100 items in each of Grades K-8.

Poorly-functioning math items, based on Rasch modeling results, were removed from consideration for inclusion in operational test forms. We initially used the item-level fit statistic to identify items that did not adequately fit the measurement model, and then the distractor-level difficulty statistic (average estimated ability of students who selected each correct or incorrect answer option) as a basis for exclusion. More specifically, for a given math item that was *overfit* (those items having an estimated Mean Square Outfit < 0.49) or *underfit* (those items having an estimated Mean Square Outfit > 1.51) to be retained in the item bank, distractor analysis had to indicate the item was functioning appropriately (i.e., students with the highest average estimated ability selected the correct answer choice, while students with lower estimated ability selected

distractors). Otherwise, such items were removed from the item bank and from consideration for inclusion.

The difficulty and model fit of each item along with distractor analysis results are reported for all piloted items in a series of grade-specific technical reports (Alonzo, Lai, & Tindal, 2009a, 2009b, 2009c; Alonzo & Tindal, 2009a, 2009b; Lai, Alonzo, & Tindal, 2009a, 2009b, 2009c, 2009d). Results from the Rasch analyses helped ensure test forms had adequate range (from easy to difficult items) to sufficiently classify students into risk categories, along with an adequate number of items on the lower tail of the distribution to detect small changes for students performing below grade level and being progress monitored over time. The Rasch analyses also allowed us to construct the forms so the average test difficulty was essentially equivalent across grade-level test forms.

In total, for each of Grades K-8, we developed three seasonal benchmarks and 30 grade-level progress monitoring test forms. Benchmarks originally consisted of 48 items, but were later refined to a 45-item tests by removing the three items from each test with the weakest technical characteristics. Benchmark forms targeted all NCTM Focal Points at each grade level. Progress monitoring forms consisted of 16 items, with 10 test forms targeting each NCTM Focal Point at each grade level. NCTM Math measures were released for use in schools in Fall 2009.

Alignment to Standards

We conducted two alignment studies in relation to the NCTM Math measures. The first study analyzed the alignment of items to the NCTM Focal Points Standards and was conducted in November 2009 to January 2010. The second study analyzed the alignment between the NCTM Math items and the Common Core State Standards (CCSS), and was conducted September 2011 to June 2012.

NCTM alignment. Judgments of individual math items were collected on a sampling of NCTM Math test forms in Grades K-1 and 3-8 using the alignment model outlined by Webb (1997, 2002). All grade-level benchmark test forms were included in the study, as well as 5, 9, or 10 progress monitoring test forms for each grade-level Focal Point. Thirteen teacher raters with experience in math teaching and learning were trained and then judged the alignment between items and target Focal Point Standards in each selected grade. Also, for Grades 3-8, raters judged the depth of knowledge (Webb, 2002) for each standard and associated math item. At least two expert raters judged the alignment of items in each test form.

Across grades, focal points, and test forms, the ratings of easyCBM© math items aligned to NCTM Focal Point Standards were generally strong. Table 12.1 displays the percent of items within benchmark and progress monitoring test forms linked to standards across all teacher raters. Depth of knowledge (DOK) ratings appeared more subjective, with findings inconclusive. Full results from the first study of alignment for the NCTM Math measures are available in a detailed technical report by Nese, Lai, Anderson, Park et al. (2010).

Table 12.1. *Percentage of math items linked to standards across benchmark and progress monitoring test forms*.

Table 12.1 – *NCTM Alignment*

	Benchmark Focal Points			Progress Monitoring Focal Points		
Grade K	Numbers Operations	Geometry	Measure-ment	Numbers Operations	Geometry	Measure-ment
	88%	98%	67%	92%	98%	65%
Grade 1	Numbers Operations	Geometry	Numbers Operations and Algebra	Numbers Operations	Geometry	Numbers Operations and Algebra
	94%	84%	94%	95%	91%	84%
Grade 3	Numbers Operations	Geometry	Numbers Operations and Algebra	Numbers Operations	Geometry	Numbers Operations and Algebra
	83%	79%	79%	96%	72%	90%
Grade 4	Numbers Operations	Measure-ment	Numbers Operations and Algebra	Numbers Operations	Measure-ment	Numbers Operations and Algebra

	92%	88%	90%	92%	97%	96%
Grade 5	Numbers Operations	Geometry Measurement Algebra	Numbers Operations Algebra	Numbers Operations	Geometry Measurement Algebra	Numbers Operations Algebra
	92%	77%	90%	99%	81%	79%
Grade 6	Numbers Operations	Algebra	Numbers Operations Ratios	Numbers Operations	Algebra	Numbers Operations Ratios
	88%	96%	100%	88%	74%	94%
Grade 7	Numbers Operations Algebra Geometry	Geometry Measurement Algebra	Numbers Operations Algebra	Numbers Operations Algebra Geometry	Geometry Measurement Algebra	Numbers Operations Algebra
	75%	92%	88%	80%	88%	88%
Grade 8	Algebra	Geometry Measurement	Data Analysis Numbers Operations Algebra	Algebra	Geometry Measurement	Data Analysis Numbers Operations Algebra
	65%	42%	77%	66%	58%	73%

CCSS alignment. Because the NCTM Math measures were originally written to align with the NCTM Focal Points Standards in each of Grade K-8, we anticipated gaps in the measures' alignment with the CCSS. Thus, results from this second alignment study were intended to serve both as a gauge of the alignment of the NCTM Math measures with the CCSS, and also as the basis for planned new item development over 2012-2013 to create new easyCBM© math measures closer alignment with the CCSS. We conducted the CCSS alignment study in two distinct phases, with results from Phase 1 serving to inform the design of Phase 2. Included in both phases of the study were all items from the three easyCBM© seasonal NCTM benchmark mathematics assessments in each of Grades K-8.

For Phase 1, beginning in September 2011, nine teachers with experience in math teaching and learning and the CCSS were recruited nationwide, one in each of Grades K-8. Phase 1 teachers were given hard copies of all items and asked to gauge item alignment to both on- and prior-grade CCSS. Because these measures were designed to assess a range of math knowledge and skills, and were written, in particular, to be appropriate for students performing

well below grade-level expectations, including those with disabilities, we reasoned that many NCTM Math items might align to prior-grade CCSS. Teachers were instructed to indicate if a given item was aligned to one or more than one on- and/or prior-grade CCSS. Strength of rating alignments were not gathered in Phase 1 of the alignment study.

An additional four teachers in each of Grades K-8 were recruited for Phase 2 of the alignment study, which began in January 2012. Teachers were asked to give strength of alignment ratings for both on- and prior-grade CCSS using a 3-point Likert scale (0 = not at all linked, 1 = somewhat linked, and 2 = direct link). In instances where a given item was deemed *not aligned* to a given CCSS, teachers were asked whether the item assessed a pre-requisite skill to that CCSS. In instances where a given item was deemed *aligned* to a given on- or prior-grade CCSS, teachers were asked to indicate the strength of alignment (1 = *somewhat linked*, or 2 = *direct link*) to the aligning standard.

We found ratings of NCTM Math items to the CCSS reasonable, though gaps in alignment were also observed as anticipated. Generally, benchmark math items appeared more strongly aligned with on-grade CCSS compared to prior-grade. Additionally, alignment at the CCSS domain level appeared stronger than at the individual standard level. Overall alignment results with respect to *on-grade* CCSS domains and standards are presented in Table 12.2, with more detailed alignment results, including those broken down by seasonal grade-level benchmark, presented in three grade-specific technical reports (Irvin, Park, Alonzo, & Tindal, 2012a, 2012b; Park, Irvin, Alonzo, & Tindal, 2012). Table 12.2 presents CCSS standard codes that are over- or under-represented in each grade level. For all codes, the grade level is listed first, followed by the domain and standard. For example, K.G.2 represents the CCSS Math standard for *Kindergarten*, in the *Geometry* domain, for *standard two*.

Table 12.2

NCTM Math alignment with the CCSS and related math item development

Grade	Common Core State Standards	
	Overrepresented	Underrepresented and 12-13 Item Development Focus
K	K.G.2, K.G.6, K.MD.2	K.CC.1, K.CC.3, K.CC.5, K.CC.7, K.G.3, K.G.5, K.MD.3, K.NBT.1, K.OA.1, K.OA.2, K.OA.3, K.OA.4
1	1.G.2, 1.NBT.1, 1.OA.1	1.G.3, 1.MD.1, 1.MD.2, 1.MD.3, 1.NBT.3, 1.NBT.5, 1.NBT.6, 1.OA.2, 1.OA.4, 1.OA.5, 1.OA.7, 1.OA.8
2	2.MD.1, 2.NBT.5	2.G.1, 2.G.2, 2.G.3, 2.MD.2, 2.MD.4, 2.MD.5, 2.MD.9, 2.MD.10, 2.NBT.2, 2.NBT.3, 2.NBT.6, 2.NBT.7, 2.NBT.8, 2.NBT.9, 2.OA.2, 2.OA.3, 2.OA.4
3	3.G.2, 3.NF.1	3.G.1, 3.MD.1, 3.MD.2, 3.MD.3, 3.MD.4, 3.MD.5, 3.MD.6, 3.MD.7, 3.MD.8, 3.NBT.1, 3.NBT.2, 3.NBT.3, 3.NF.2, 3.OA.5, 3.OA.6, 3.OA.8
4	4.MD.2, 4.NF.6	4.G.1, 4.G.2, 4.G.3, 4.MD.4, 4.MD.5, 4.MD.6, 4.MD.7, 4.NBT.1, 4.NBT.2, 4.NBT.3, 4.NBT.4, 4.NBT.6, 4.NF.1, 4.NF.2, 4.NF.3, 4.NF.4, 4.NF.5, 4.OA.1, 4.OA.4
5	5.MD.3, 5.NBT.7, 5.NF.1	5.G.1, 5.G.2, 5.G.3, 5.G.4, 5.MD.1, 5.MD.2, 5.NBT.1, 5.NBT.2, 5.NBT.3, 5.NBT.4, 5.NBT.5, 5.NF.2, 5.NF.3, 5.NF.4, 5.NF.5, 5.NF.6, 5.NF.7, 5.OA.1, 5.OA.2, 5.OA.3
6	6.EE.2, 6.RP.3	6.EE.8, 6.EE.9, 6.G.1, 6.G.2, 6.G.3, 6.G.4, 6.NS.1, 6.NS.2, 6.NS.3, 6.SP.1, 6.SP.2, 6.SP.3, 6.SP.4
7	7.G.4, 7.NS.3	7.EE.2, 7.G.2, 7.G.3, 7.G.5, 7.SP.1, 7.SP.2, 7.SP.3, 7.SP.4, 7.SP.8
8	8.F.3, 8.F.4	8.EE.1, 8.EE.2, 8.EE.3, 8.EE.4, 8.EE.5, 8.F.2, 8.F.5, 8.G.1, 8.G.2, 8.G.3, 8.G.8, 8.G.9, 8.NS.1, 8.NS.2, 8.SP.1, 8.SP.2, 8.SP.3, 8.SP.4

Reliability

The internal consistency of the easyCBM© NCTM Math measures has been investigated by Anderson, Lai et al (2010) and Nese, Lai, Anderson, Jamgochian et al. (2010), using Cronbach's alpha and split-half reliability analyses. Anderson, Lai et al (2010) analyzed extant data from the seasonal benchmark assessments in Grades K-2 (fall, winter, and spring) over the 2009-2010 school year. Nese, Lai, Anderson, Jamgochian et al (2010) analyzed extant data from the Grades 3-8 benchmark assessments during the same time period to grades 3-8.

For all time points and grades in each study, Cronbach's alpha ranged from .78 - .91, indicating *acceptable* to *strong* reliability (see Table 12.3). For split-half reliability, coefficients ranged from .71 to .89, with a median of .82, which indicated *acceptable* to *strong* reliability (see Table 12.4).

Table 12.3
Cronbach's Alpha of the NCTM math measure for Grades K-8

Grade	*N*	Fall	Winter	Spring
K	3511	.83	.85	.87
1	3785	.78	.86	.89
2	3675	.80	.85	.82
3	4269	.82	.85	.86
4	4282	.87	.86	.87
5	4343	.85	.88	.91
6	4455	.86	.88	.91
7	4270	.89	.89	.90
8	4413	.86	.90	.89

Table 12.4

Split-half Reliability (Spearman-Brown Analysis) of the NCTM math measure for Grades K-8

Grade	Fall	Winter	Spring
K	.80	.82	.82
1	.73	.79	.85
2	.75	.86	.79
3	.76	.81	.82
4	.84	.81	.85
5	.83	.85	.88
6	.84	.84	.89
7	.70	.71	.81
8	.86	.80	.84

Validity Evidence

Criterion Validity

Multiple studies have investigated the criterion validity of the easyCBM© NCTM measures. Table 12.5 summarizes eight studies investigating criterion-related validity evidence for the measures. All studies looked at both predictive and concurrent validity. Each of these studies is reviewed in more detail below. For a complete description of each study, please see the full reports.

Table 12.5
Studies Investigating Criterion-related Validity Evidence

Study	Grades	Summary of results
Nese, Lai, Anderson, Jamgochian et al. (2010)	3-8	Model R^2 = .48 to .73 Sensitivity = .73 to .88 Specificity = .75 to .88 AUC = .83 to .93
Anderson, Alonzo, and Tindal (2010a)	3-8	Sensitivity = .79 to .90 Specificity = .65 to .84 AUC = .84 to .94
Anderson, Alonzo, and Tindal (2010b)	3-8	Sensitivity = .78 to .93 Specificity = .69 to .85 AUC = .86 to .92
Anderson, Alonzo, and Tindal (2010c)	3-8	Model R^2 = .48 to .68

		r = .70 to .82
Anderson, Alonzo, and Tindal (2010d)	3-8	Model R^2 = .48 to .67 r = .68 to .83
Anderson, Lai et al. (2010)	K-2	Model R^2 = .39 to .54
Anderson, Alonzo, and Tindal (2011a)	3-8	Sensitivity = .80 to .94 Specificity = .71 to .84 AUC = .85 to .92
Anderson, Alonzo, and Tindal (2011b)	3-8	Sensitivity = .77 to .92 Specificity = .71 to .86 AUC = .82 to .94

Predictive validity. Nese, Lai, Anderson, Jamgochian et al. (2010) examined the predictive validity of the easyCBM© NCTM measures by comparing the fall and winter measures to the spring 2010 administrations of the mathematics portion of the Oregon Assessment of Knowledge and Skills (OAKS) in three districts in Oregon, and the spring 2010 administration of the mathematics portion of the Measures of Student Progress (MSP) in one district in the state of Washington. The sample included approximately 3,600 students per grade level in Oregon and 650 students per grade level in Washington. Results were analyzed by ethnicity, but only overall results are reported here. The fall and winter simple linear regression model accounted for 58-73% of the variance in OAKS math test scores, and 56-72% of the MSP math test scores, with variance accounted for generally increasing with grade level.

The authors also explored the predictive utility of within-year growth estimates, split by quartile, as well as the diagnostic efficiency of the tests for predicting whether or not students would meet proficiency on the state test. For students in the bottom quartile of normative performance, standardized coefficients for the predictive utility of the slope ranged from .47 to .82; for students in the second quartile, standardized coefficients ranged from .39 to.65; for students in the third quartile, standardized coefficients ranged from .38 to.83; and for students in the fourth quartile, standardized coefficients ranged from -.47 to.63 (the negative growth results

were isolated to one grade in one state, and were likely sample specific). These coefficients imply that for every standard deviation increase in NCTM scores, there was a corresponding increase of approximately .5 to .75 standard deviation increase in OAKS scores. In terms of diagnostic efficiency, the area under the receiver operating characteristic curve (AUC) statistics ranged from .86 to .92 for the OAKS and .83 to .93 for the MSP. Sensitivity for the optimal cut score ranged from.73 to .88, for the OAKS and .75 to .90 for the MSP. Specificity statistics for the optimal cut score ranged from .76 to .87 for the OAKS and .75 to .88 for the MSP.

Anderson, Alonzo et al. (2010a) established optimal cut scores for the fall and winter measures in terms of meeting proficiency on the OAKS, while Anderson, Alonzo et al. (2010b) did the same for the MSP. Results in each report are provided by ethnicity category, but only overall results are reported here. For the fall and winter administrations, the sensitivity of the optimal cut score ranged from .78 to .92 for the OAKS and .79 to .90 for the MSP. Specificity results ranged from .69 to .82 for the OAKS and .71 to .84 for the MSP. The overall classification accuracy results ranged from .72 to .81 for the OAKS, and .75 to .85 for the MSP. Finally, the AUC ranged from .86 to .91 for the OAKS, and .84 to .93 for the MSP.

Anderson, Alonzo et al. (2010c) explored the predictive validity of the easyCBM© NCTM measures in Grades 3-8 by examining the relation between the fall and winter benchmark measures and OAKS scores during the 2009-10 school year, while Anderson, Alonzo et al. (2010d) did the same for the MSP. The Oregon sample included 1,707 students per grade, while the Washington sample included approximately 640 students per grade. Fall and winter easyCBM© correlations ranged from 70 to .81 for OAKS and .68 to .83 for the MSP. A multiple regression model of the three seasonal easyCBM© NCTM measures accounted for 64% to 75% of the total variance in spring OAKS scores and 59% to 75% in spring MSP scores. In simple

regression analyses, the fall measures accounted for 48% to 65% of the variance in OAKS, and 49% to 67% of the variance in spring MSP. Simple linear regression results for winter ranged from approximately 48% to 68% for OAKS and 52% to 67% for the MSP.

Anderson, Lai et al. (2010) explored the predictive validity of the easyCBM© NCTM measures in Grades K-2 by comparing the fall and winter measures to the TerraNova 3, which was administered in May. The sample included students from 76 schools in 54 cities across 26 states in 4 geographic locations. Students in grades K (n = 2,400), 1 (n = 3,782), and 2 (n = 2,940) were administered levels 10, 11, and 12 of the math portion of the TerraNova, respectively. In simple linear regression models, the fall measure accounted for 39%-54% of the variance in TerraNova, while the winter measure accounted for 27%-54% of the variance in TerraNova. These results were significant ($p < .05$). The authors also used hierarchical linear modeling (HLM) to obtain within-year growth estimates, split by quartile. Across the three grade levels included, the standardized coefficients ranged from .58-.68 for quartile 1, .29-.51 for quartile 2, .44-.74 for quartile 3, and .46-.82 for quartile 4. These coefficients imply that for every standard deviation increase in NCTM scores, there was a corresponding increase of approximately .5 to .75 standard deviations in TerraNova scores.

Anderson et al. (2011a) and Anderson et al. (2011b) conducted cross-validation studies of the cut scores to optimally predict OAKS and MSP performance, respectively, in Grades 3-8. The studies extend the work conducted by Anderson, Alonzo et al. (2010c) and Anderson, Alonzo et al. (2010d) by exploring the stability of the cut scores across two randomly-selected groups, each including approximately 2,000 students. The authors noted that the cut scores appeared quite stable overall, with the cut scores performing consistently across groups. The 95% confidence intervals for AUC statistics also overlapped between the randomly-selected

groups, suggesting the measure was equally predictive for each group. The consistency between the optimal cut scores combined with the lack of significant differences between AUC statistics in all measurement occasions and grades provide strong evidence for the cut scores derived. Sensitivity statistics for the optimal cut score in Oregon ranged from .80 to .94, while specificity ranged from .71 to .86. In Washington, sensitivity ranged from .77 to .91, while specificity ranged from .71 to .86. AUC statistics for the fall and winter predictions in Oregon ranged from .85 to .91, and from .82 to .94 for Washington.

Concurrent validity. Nese, Lai, Anderson, Jamgochian et al. (2010) evaluated the concurrent validity of the easyCBM© NCTM measures by comparing the spring measures to the spring 2010 administrations of the OAKS and the MSP, respectively. Simple linear regression analyses demonstrate that the measures accounted for 52%-67% of the variance in OAKS and 48%-67% of the variance in MSP scores.

Anderson, Alonzo et al. (2010c) explored the concurrent validity of the easyCBM© NCTM measures by examining the relation between spring benchmarks in 2011 and OAKS scores, while Anderson, Alonzo et al. (2010d) did the same for the MSP. Correlations with OAKS scores ranged from .73 to .82, and from .68 to .81 for the MSP. Simple linear regression models accounted for 52% to 67% of the variance in OAKS, and 48% to 67% of the variance in the MSP.

Anderson, Lai et al. (2010) also explored the concurrent validity of the easyCBM© NCTM measures at Grades K-2 by comparing the spring measure to the TerraNova 3, which was administered in May. The spring regression models were significant across the three grades evaluated, with the model accounting for 52-53% of the variance in TerraNova ($p < .05$).

Anderson, Alonzo et al. (2010a) established optimal cut scores for the spring measure in terms of meeting proficiency on the OAKS in Grades 3-8, while Anderson, Alonzo et al. (2010b) did the same for the MSP. Results in each report are provided by ethnicity category, but only overall results are reported here. Sensitivity for the optimal cut score ranged from .80 to .93 for OAKS, and .80 to .89 for the MSP. Specificity ranged from .73 to .85 for OAKS, and .65 to .78 for the MSP. The overall classification accuracy results ranged from .77 to .84 for OAKS, and .73 to .83 for the MSP. Finally, AUC statistics ranged from .89 to .92 for OAKS, and .88 to .94 for the MSP.

Anderson et al. (2011a) and Anderson et al. (2011b) also conducted cross-validation studies of the optimal cut scores for meeting proficiency on OAKS and the MSP, respectively, in Grades 3-8. Sensitivity statistics for Oregon ranged from .81 to .91, while specificity ranged from .73 to .84. In Washington, sensitivity statistics ranged from .82 to .92, while specificity ranged from .73 to .84. AUC statistics in Oregon ranged from .89 to .92., and from .87 to .94 for Washington.

Construct Validity

To gather construct validity evidence for the easyCBM© NCTM math measures, Anderson, Lai et al. (2010) and Nese, Lai, Anderson, Jamgochian et al. (2010) conducted a series of analyses, including Rasch analyses, confirmatory factor analyses (CFA), and bivariate correlational analyses. These analyses were used to examine item-level information, the internal structures of the Grades K-8 easyCBM© Math assessments respectively, and their relationships with year-end state math tests.

Grades K-2. For Grades K-2, Anderson, Lai, and colleagues (2010) first examined the fit of the items to the Rasch model, which assumes unidimensionality (one factor measuring the

latent construct of math ability). Poorly fitting items indicate a departure from the one-factor assumption. The authors then compared two CFA models to evaluate construct validity of the math measures: a one-factor model and a three-factor model measuring three NCTM focal points. See Figures 12.1 and 12.2 below for the two hypothesized measurement models.

Results from the Rasch analysis for the fall, winter, and spring measures in Grades K and 1 suggested the items fit moderately well to the one-factor assumption, with mean square outfit values ranging from .50-1.30. For Grade 2, the mean square outfit values ranged from .60-1.79. The chi-square difference test between a three-factor model and the one-factor model indicated that the three-factor models did not result in significantly better fit (only fall measures were evaluated). Additionally, there were moderate to high correlations between the factors on the three-factor model (.70-.90), providing further evidence for the one-factor model.

Grades Three-Eight. Similar CFA studies were conducted to evaluate the underlying structure of the Grade 3-8 easyCBM© math measures. Again, two models were compared: a one-factor and a three-factor model (see Figures 12.1 and 12.2). Using the same criteria as the analyses for the earlier grades, the chi-square difference test between a three-factor model and the one-factor model indicated that the three-factor models did not result in significantly better fit for the Grade 3-8 data. Additionally, there were moderate to high correlations between the factors on the three-factor model (.60-.80), providing further evidence for the one-factor model.

A series of bivariate correlational analyses (disaggregated by subgroups for ethnicity for each separate grade-level analysis) between the fall, winter, and spring measures and the OAKS and MSP were also examined as another source of construct validity evidence. Results indicated that the easyCBM© math measures consistently displayed moderate to strong correlations with the year-end state math tests ranging from approximately .60-.80.

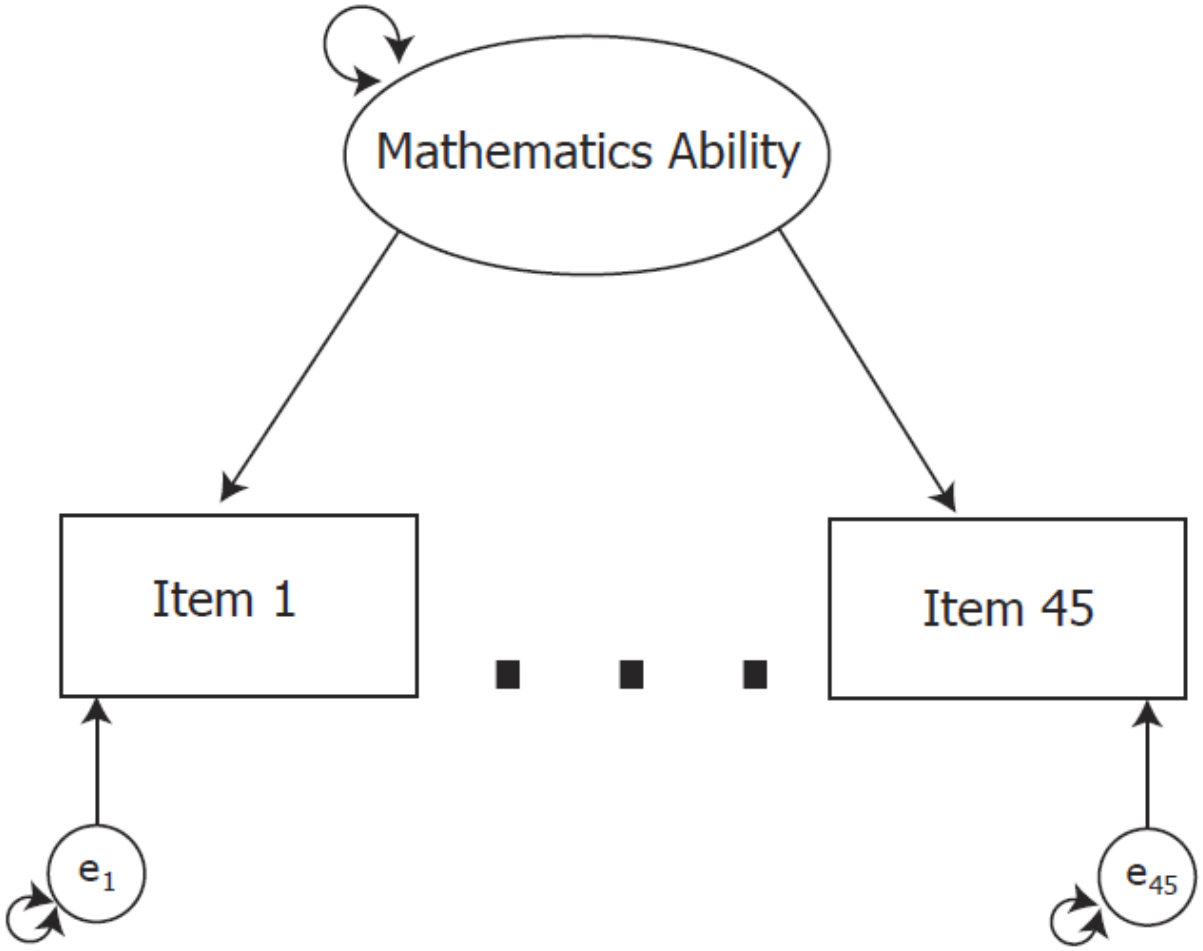


Figure 12.1. One factor CFA model.

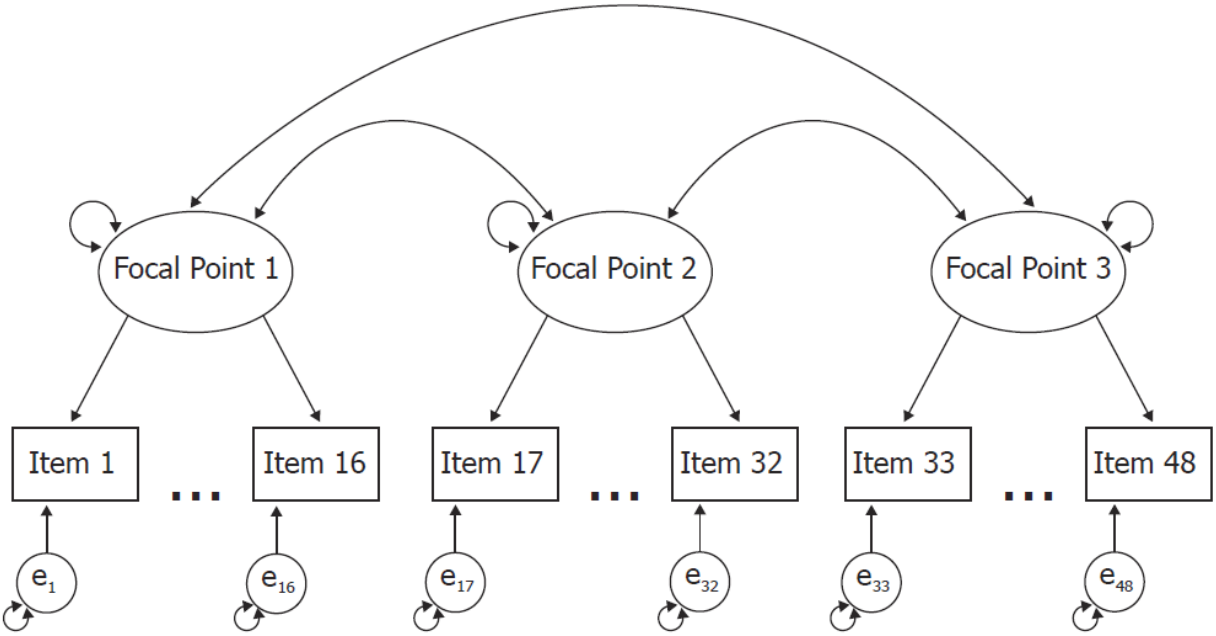


Figure 12.2. Three factor CFA model.

References

Alonzo, J., Lai, C. F., & Tindal, G. (2009a). The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 3 (technical report 0902). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Lai, C. F., & Tindal, G. (2009b). The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 4 (technical report 0903). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Lai, C. F., & Tindal, G. (2009c). The development of K-8 progress monitoring measures in mathematics for use with the 2% and the general education populations: Grade 2 (technical report 0920). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009a). The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 1 (technical report 0919). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009b). The development of K-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Kindergarten (technical report 0921). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2010a). Diagnostic efficiency of easyCBM math: Oregon (technical report 1009). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2010b). Diagnostic efficiency of easyCBM mathematics: Washington state (technical report 1008). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2010c). easyCBM mathematics criterion related validity evidence: Oregon state test (technical report 1011). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2010d). easyCBM mathematics criterion related validity evidence: Washington state test (technical report 1010). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2011a). A cross-validation of easyCBM mathematics cut scores in Oregon: 2009-2010 (technical report 1104). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2011b). A cross-validation of easyCBM mathematics cut scores in Washington state: 2009-2010 Test (technical report 1105). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Lai, C. F., Nese, J. F. T., Park, B. J., Sáez, L., Jamgochian, E. M., et al. (2010). Technical adequacy of the easyCBM primary-level mathematics measures (grades k-2), 2009-2010 version (technical report 1006). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012a). The alignment of the easyCBM grades 6-8 math measures to the Common Core Standards (technical report 1230). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012b). The alignment of the easyCBM Grades k-2 math measures to the Common Core Standards (technical report 1228). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Johnstone, C., Altman, J., & Thurlow, M. (2006). A state guide to the development of universally designed assessments. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Lai, C. F., Alonzo, J., & Tindal, G. (2009a). The development of k-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 5 (technical report 0901). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2009b). The development of k-8 progress monitoring measures in mathematics for use with the 2% and general education populations: Grade 8 (technical report 0904). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2009c). The development of k-8 progress monitoring measures in mathematics for use with the 2% and general populations: Grade 7 (technical report 0908). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Lai, C. F., Alonzo, J., & Tindal, G. (2009d). The development of k-8 progress monitoring measures in mathematics for use with the 2% and the general education populations: Grade 6 (technical report 0907). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Nese, J. F. T., Lai, C. F., Anderson, D., Jamgochian, E. M., Kamata, A., Saez, L., et al. (2010). Technical adequacy of the easyCBM mathematics measures: Grades 3-8: 2009-2010 version

(technical report 1007). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Nese, J. F. T., Lai, C. F., Anderson, D., Park, B. J., Tindal, G., & Alonzo, J. (2010). The alignment of easyCBM math measures to curriculum standards (technical report 1002). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Park, B. J., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). The Alignment of the easyCBM Grades 3-5 math measures to the Common Core Standards (technical report 1229). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

U.S. Department of Education. (2005). *Title 1--Improving the Academic Achievement of the Disadvantaged; Individuals With Disabilities Education Act (IDEA)--Assistance to States for the Education of Children with Disabilities*. Washington, DC: Office of Elementary and Secondary Education, Office of Special Education and REhabilitative Services, U.S. Department of Education.

Webb, N. L. (1997). *Research monograph No. 6: Criteria for alignment of expectations and assessments in mathematics and science education*. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers.

Chapter 13: CCSS Math Measures

The technical adequacy evidence gathered for the easyCBM© CCSS Math measures to date suggests they are functioning largely as intended. An advanced statistical technique for scaling items (Rasch modeling) was used during the developmental process, providing information on the difficulty and functioning of each item. These results then guided test form creation to help ensure all test forms within each grade level were essentially equivalent in terms of difficulty and the distribution of easy and difficult items within each test form. Anderson, Irvin, Alonzo, and Tindal (2012) showed that the items aligned quite strongly with the Common Core State Standards (CCSS) of the respective grade level. Wray, Alonzo, and Tindal (2014) found the internal consistency of the measures to be quite strong across all grades ($\alpha \geq .80$). Finally, Anderson, Rowley, Alonzo, and Tindal (2014) showed the measures had a strong relation with the Stanford Achievement Test – 10th edition, in Grades 6-8.

In what follows, we summarize the technical adequacy evidence for the easyCBM© NCTM Math measures. We begin with the developmental process used, including creating alternate test forms of equivalent difficulty, as well as the alignment between the measures and the NCTM Focal Point Standards. We also report the results of an exploratory study of the alignment between the NCTM measures and the Common Core State Standards. We then summarize the reliability evidence collected to date, and conclude by discussing criterion and construct validity evidence. For a conceptual discussion of the theoretical purpose for any analysis conducted, please see Chapter 3. Note that this document is intended to provide a summary of evidence, not to explain in detail the different studies from which the evidence was gleaned. For a complete account of each study, the full reports are available at www.brtprojects.org/publications/technical-reports.

Measure Development

The easyCBM© Common Core State Standards (CCSS) Math measures were developed in two separate cycles based on grade band. Middle school measures (Grades 6-8) were initially developed in 2011-2012 and later refined in 2012-2013, while elementary measures (Grades K-5) were developed in 2012-2013. Measure development for the two grade bands, though unique in some ways, was broadly similar, following the same specifications for item writing/review and scaling. In what follows, we give a broad view of the development of the CCSS Math measures in Grades K-8, with important differences between grade bands highlighted.

The K-8 CCSS are divided into four (Grades 1 and 2) or five (Grades K, 3-8) content domains that are further subdivided into 21-29 individual standards. Though some grade-level standards are further divided into more granular sub-standards, all such sub-standards were treated as part of the parent standard for the purpose of item development. As an initial step in developing the elementary grade CCSS math tests, a total of 3,000 new math items were written: 500 at each of grades K-5. These items were added to an item bank of items previously written to align to the National Council of Teachers of Mathematics (NCTM) Focal Point Standards, which had been identified through previous alignment research as addressing the CCSS (see Irvin, Park, Alonzo, & Tindal, 2012b; Park, Irvin, Alonzo, & Tindal, 2012). For the middle-school grades, a total of 2,700 new items were written: 900 at each of grades 6, 7, and 8, stratified across all CCSS in the respective grades. In other words, for Grades K-5, item development focused only on specific standards not addressed by items in previously operationalized measures (those written to the NCTM Focal Point Standards – see Chapter 11), while in Grades 6-8 item development was designed so that an even number of items would be written to each standard.

For both development efforts, groups of content experts were recruited to write and

review K-8 CCSS math items: *teacher-lead item reviewers* and *item writers*. Teacher-leads and item writers were recruited and assigned duties commensurate with their experience and expertise in mathematics. In general, teacher-leads had more experience in math teaching and assessment, often serving in positions of leadership at the school or district level, while item writers were typically classroom teachers with at least 5 years of math teaching experience. Teacher-leads served as main contacts for grade-level item writers—collecting, reviewing, editing and then transferring newly written math items to the lead researcher for further in-depth review and revision.

Teacher-leads and item writers were trained on specific guidelines on how to approach the item writing and review process through in-person trainings or online webinars. Two major principles of item/test development were emphasized including: (a) recommendations for quality multiple-choice item writing given by Haladyna (2004) and Downing (2006a, 2006b); and (b) principles of Universal Design for Assessment, as outlined by Thompson, Johnstone, and Thurlow (2002). For each new math item, item writers were required to write a stem (question or prompt) and three response options (one correct and two meaningful distractors) that were similar in length and complexity level and that aligned with a targeted standard. Throughout trainings, a variety of math items were shown to teacher-leads and item writers, followed by discussion of why the items met or did not meet the principles of quality item development or why they were (or were not) aligned to the CCSS. The lead researcher and study participants then brainstormed and discussed specific ways in which items could be altered to meet item design principles or better align to a targeted CCSS. Trainings included time to allow participants to practice writing items. In this manner, trainings were designed to mimic expectations around item writing and review required by participants in the studies.

Specifically, teacher-leads and item writers worked concurrently to write, review and revise math items across all grades, with item writers submitting items in batches (for example, 25 items at a time) to teacher-leads for review of mathematical correctness and for the features discussed during the training (Universal Design for Assessment features, alignment to standards, etc.). A professional graphic artist produced original graphics for items using Adobe Illustrator. Consistent with Universal Design, all graphics were line drawings with no shading. Items only required a graphic if the item writer/teacher-lead had explicitly requested and described a desired graphic for the item.

Once submitted to the lead researcher (and subsequent to graphics development), math items were loaded into a secure online system and reviewed internally by University of Oregon faculty and research assistants (undergraduate and graduate students employed by the research institute developing the assessments, Behavioral Research and Teaching) with experience in test and item development. Through these internal reviews, test items, including associated graphics, were again edited for mathematical correctness and features of quality item/test development as emphasized in trainings, and additionally to ensure that items were consistently formatted. In the easyCBM© online assessment system, students have the option to click on a speaker button next to the stem and/or distractor answer, which results in having the text-based portions of the item "read-aloud" to them. Thus, subsequent to internal reviews, we also contracted people to produce audio recordings (in both English and Spanish) of text-based portions in all math items. The person who recorded the English language recordings was a native English speaker. The person who recorded the Spanish language recordings was a native Spanish speaker. For Grades K-5 we also commissioned a separate external review of math items. We contracted with three individuals who did not have item writing or mathematics teaching/assessment experience to

review math items (through the secure online system) for mathematical accuracy, consistent formatting, and audio accuracy.

Piloting was conducted in winter 2012 for middle school items and spring 2013 for elementary items using a national sample of convenience based on teachers signing up through the easyCBM assessment system. Data were collected via a secure online piloting system that presented each student with a pilot test form comprised of a series of items (32 items per pilot form in Grades K-1, 41 items per pilot form in Grades 2-5, and 50 items per pilot form in Grades 6-8). For all Grades K-8, pilot test forms included 10 horizontal anchor items within grade, 5 horizontal anchor items linking to the previous pilot form and 5 horizontal anchor items linking to the subsequent pilot form. Including horizontal anchor items allowed items within grade to be calibrated to the same scale. Further, in the case of the elementary test development, horizontal anchoring allowed us to more than double the item pool from which we would select items to create the operational CCSS math test forms by calibrating the two original item pools (older elementary NCTM Math items deemed aligned to the CCSS, and the newly-written elementary CCSS Math items) to a common scale.

One difference in pilot form creation and scaling that is important to note was the inclusion of vertical anchor items in the pilot forms for Grades 6-8. Sixth and eighth grade forms included an additional 10 vertical anchors: 5 from within the grade and 5 from seventh grade. However, the seventh grade forms included 15 vertical anchors: 5 sixth grade items, 5 seventh grade items, and 5 eighth grade items. The vertically-anchored piloting design was thus balanced, with students in both grades taking the same vertically-anchored items (e.g., sixth graders taking sixth- and seventh-grade items and seventh graders taking the same sixth- and seventh-grade items). Again, vertical anchor items were not included in the K-5 piloting. Rather, we included

prior- and subsequent-grade math items in the operationalized test forms for Grades 1-5 (and subsequent-grade items for Grade K) in order to calibrate to a common vertical scale in future years.

All CCSS Math items were scaled with a Rasch model (see Chapter 2 for a conceptual overview of this methodology) with common items linked across all test takers used to scale items concurrently within each grade (and vertically across grades for Grade 6-8). The difficulty and model fit of each item are reported for the full item banks and by operational benchmark and progress monitoring test forms in a series of grade-specific technical reports (Anderson, Irvin, Patarapichayatham, Alonzo, & Tindal, 2012; Irvin et al., 2013a, 2013b, 2013c; Saven et al., 2013a; Saven et al., 2013b, 2013c). Results from the Rasch analyses helped ensure test forms had adequate range (from easy to difficult items) to sufficiently classify students into risk categories, along with an adequate number of items on the lower tail of the distribution to detect small changes for students being progress monitored over time. The Rasch analyses also allowed us to construct the forms so the average difficulty was essentially equivalent across grade-level test forms.

In the first year of operational use, Anderson, Alonzo, and Tindal (2013) examined the reliability of the CCSS Math middle school measures and found the test forms to be operating at less-than-ideal levels (i.e., Cronbach's alpha < .70). Based on these reliability results, an additional pilot was conducted in winter 2013, with Rasch scaling again used to calibrate items to a common vertical scale. During the summer of 2013, operational test forms in Grades 6-8 were revised to remove the five least-technically-adequate items from each test form, replacing them with items that met acceptable Rasch model fit criteria (see Chapter 2). Additionally, we added five additional items to each test form from the NCTM item pool deemed aligned through

research by Irvin, Park, Alonzo, and Tindal (2012a) to improve the accessibility of the tests for use within RTI (Anderson et al., 2013). Rasch analyses were again used in the revision of the CCSS Math middle school measures to ensure adequate range from easy to difficult items on all test forms, and to maintain approximately the same average difficulty across grade-level test forms.

Alignment to Standards

The alignment of content between the middle school CCSS Math measures and targeted CCSS, and the degree to which items deemed *not-aligned* with a given standard target a requisite skill to the standard is reported in Anderson, Irvin, Alonzo, et al. (2012). Fifteen middle school math teachers were recruited from across the United States to participate in the alignment study based on their knowledge of the CCSS, and knowledge of the mathematical content. Participating teachers were trained on study details and required tasks through an online webinar, and independently rated the alignment of a total of 270 items using a secure online tool (Distributed Item Review; DIR) designed to distribute test items across a broad geographic range to examine them for bias, sensitivity and alignment to standards.

Approximately 50% of the total middle school math item bank was selected for the alignment review. We used a matrix-sampling plan to select items that stratified across all selected items by CCSS and item writers. Items were grouped into grade-level sets comprised of 90 items stratified as evenly as possible by CCSS domain and standard to provide equal representation of the CCSS within and across each set. Teachers rated the alignment of individual math items in three assigned sets to a single target CCSS and then requisite skills (when an item was deemed not to align with a given standard). Items were rated on a 4-point ordinal alignment scale, as follows:

0 = no alignment,

1 = vague alignment,

2 = somewhat aligned,

3 = directly aligned.

The four-point scale allowed raters to indicate the degree to which an item was or was not aligned to a paired standard that in many instances required students to demonstrate multiple skillsets to demonstrate content/skill mastery. Ratings for each item from the 4-point scale were later collapsed into dichotomous *aligned* (score of 2 or 3) or *not aligned* (score of 0 or 1) categories for the purpose of creating an item pool comprised of *aligned* items from which we developed operational test forms.

We used the many-facets Rasch model (MFRM) to estimate and control for the effect of the severity/leniency of individual teacher raters. The MFRM provided an adjusted-average rating based on the severity/leniency of each teacher rating a particular item – it is this adjusted average rating, controlling for rater severity/leniency that was used in all cases to determine whether an item aligned or did not align with the target standard. Overall, 1,180/1,345 math items (87.73%) had an adjusted MFRM rating above 2.0, suggesting they were aligned to their corresponding standard. Of the remaining 165 items that were rated as not aligned, 160/165, or 97.00%, were rated as addressing a requisite skill to the standard by consensus. In other words, teachers rated 99.6% of items sampled as aligned with a grade-level CCSS or requisite skill to the standard.

The MFRM also provided a rater fit statistic as an indication of the consistency with which teacher raters made their alignment decisions. That is, given the estimated severity of the rater, as determined by his or her rating of all other items, and the estimated endorsability of the

item (the ease with which raters endorsed the item as aligned with the standard), does he or she consistently rate items as would be expected? In our study, we found that all raters fit the model expectations quite well, with mean square outfit statistics ranging from 0.76 to 1.16, where 1.0 is considered an "ideal" fit to the model, and mild deviations are expected with little cause for concern (Myford & Wolfe, 2000). Detailed alignment results are reported in a technical report by Anderson, Irvin, Alonzo, et al. (2012).

Reliability

Wray, Alonzo, and Tindal (2014) investigated the reliability of the easyCBM© fall and winter CCSS math benchmarks in Grades K-8. Internal consistency was documented with Cronbach's alpha and split-half reliability. Results suggested high internal consistency across grades, as displayed in Table 13.1 below. The authors also examined top/bottom reliability, and across grade levels all but one item (item 9 of the Grade 6 fall benchmark) functioned as expected. See the full report for full top/bottom reliability results.

Table 13.1

Internal Reliability: CCSS Math

Grade/Time	Cronbach's Alpha	Split-half Reliability		
		1st Half	2nd Half	Correlation
K/F	.84	.68	.81	.52
K/W	.81	.70	.73	.53
1/F	.81	.65	.77	.53
1/W	.84	.72	.76	.62
2/F	.87	.77	.81	.67
2/W	.88	.78	.81	.66
3/F	.87	.71	.84	.64
3/W	.88	.73	.86	.65
4/F	.90	.81	.86	.67
4/W	.90	.82	.86	.68
5/F	.92	.79	.90	.68
5/W	.91	.80	.89	.69
6/F	.92	.80	.92	.65
6/W	.95	.86	.95	.69
7/F	.93	.82	.94	.62
7/W	.95	.87	.94	.70
8/F	.93	.83	.92	.71
8/W	.93	.83	.92	.73

Validity

Anderson et al. (2014) examined the concurrent validity of the easyCBM© CCSS Math measures in Grades 6-8 by exploring the relation between students' scores on the winter benchmark and the Stanford Achievement Test, Tenth Edition (SAT-10). The authors used a relatively small sample from one district in the Pacific Northwest, ranging from 63-67 students per grade. The bivariate correlation between the measures ranged from .75 to .82, while simple linear regression analyses indicated that the easyCBM© winter benchmark accounted for 56%-67% of the variance in students' SAT-10 scores. The easyCBM© winter benchmark, therefore, had a high relation with the SAT-10, suggesting the tests were likely measuring the same underlying construct.

References

Anderson, D., Alonzo, J., & Tindal, G. (2013). Study of the reliability of CCSS-aligned math measures (2012 research version): Grades 6-8 (technical report 1312). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). The alignment of the easyCBM middle school mathematics CCSS measures to the Common Core State Standards (technical report 1208). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Irvin, P. S., Patarapichayatham, C., Alonzo, J., & Tindal, G. (2012). The development and scaling of the easyCBM CCSS middle school mathematics measures (technical report 1207). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Rowley, B., Alonzo, J., & Tindal, G. (2014). Critierion Validity Evidence for the easyCBM CCSS Math Measures: Grades 6-8 (Technical Report No. 1402). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Downing, S. M. (2006a). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-301). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Downing, S. M. (2006b). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Haladyna, T. (2004). *Developing and validating multiple-choice test items* (Third ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012a). The alignment of the easyCBM grades 6-8 math measures to the Common Core Standards (technical report 1230). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012b). The alignment of the easyCBM Grades k-2 math measures to the Common Core Standards (technical report 1228). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Saven, J. L., Alonzo, J., Park, B. J., Anderson, D., & Tindal, G. (2013a). The development and scaling of the easyCBM CCSS elementary mathematics measures: Grade 2 (technical report 16). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Saven, J. L., Alonzo, J., Park, B. J., Anderson, D., & Tindal, G. (2013b). The development and scaling of the easyCBM CCSS elementary mathematics measures: Grade 4 (technical report 18). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Irvin, P. S., Saven, J. L., Alonzo, J., Park, B. J., Anderson, D., & Tindal, G. (2013c). The development and scaling of the easyCBM CCSS elementary mathematics measures: Grade K (technical report 1314). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Myford, C. M., & Wolfe, E. W. (2000). *Strengthening the ties that bind: Improving the linking network in sparsely connected rating designs.* TOEFL Technical Report No. 15. Princeton, NJ: Educational Testing Service.

Park, B. J., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). The Alignment of the easyCBM Grades 3-5 math measures to the Common Core Standards (technical report 1229). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Saven, J. L., Irvin, P. S., Park, B. J., Alonzo, J., Anderson, D., & Tindal, G. (2013a). The development and scaling of the easyCBM CCSS elementary mathematics measures: Grade 1 (technical report 1315). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Saven, J. L., Irvin, P. S., Park, B. J., Alonzo, J., Anderson, D., & Tindal, G. (2013b). The development and scaling of the easyCBM CCSS elementary mathematics measures: Grade 3 (technical report 17). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Saven, J. L., Irvin, P. S., Park, B. J., Alonzo, J., Anderson, D., & Tindal, G. (2013c). The development and scaling of the easyCBM CCSS elementary mathematics measures: Grade 5 (technical report 3019). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Wray, K., Alonzo, J., & Tindal, G. (2014). *Internal consistency of the easyCBM CCSS Math Measures: Grades K-8* (Technical Report No. 1405). Eugene, OR: University of Oregon, Behavorial Research and Teaching.

Chapter 14: Spanish Measures

The easyCBM© Spanish measures were added to the system during the 2013-14 school year. As such, few studies on the technical adequacy of the measures have been conducted, relative to other measures in the system. Below, we primarily document the development of the measures, as the foundation of an evolving validity argument for their use with RTI frameworks.

Measurement Development

We developed the Spanish literacy measures (Syllable Sounds, Syllable Segmenting, Word Reading Fluency, and Sentence Reading) for Grades K-2 during the 2007-2008 school year. As a basis for developing alternate test forms of these measures, we used correlation and multiple regression techniques to examine the appropriateness of different types of Spanish language phonological awareness and early reading assessments for inclusion in the easyCBM® assessment system. Examination of various literacy measures occurred over two studies with distinct student populations in order to test the robustness of findings across different samples with different language backgrounds: primarily native English speakers enrolled in a Spanish language immersion program (Study 1), and native Spanish speakers enrolled in a dual language immersion program at a large urban school (follow-up Study 2).

First- and second-grade students participating in Study 1 attended a small suburban school and were enrolled in a Spanish language immersion program. Students in the initial study were primarily native English speakers, but were instructed entirely in Spanish. First-grade students (n = 48) were administered five different early literacy measures: Letter Sounds, Syllable Sounds, Phoneme Segmenting, Syllable Segmenting, and Word Reading Fluency, while second-grade students (n = 50) were administered those five measures along with a Sentence Reading measure over two days in spring 2007. We analyzed the results with correlation and

regression analyses to test the strength of the relations between the different early literacy measures (Letter Sounds, Phoneme Segmenting, Syllable Sounds, Syllable Segmenting, and Word Reading Fluency) and the measure used to assess student ability to read orally in Spanish (Word Reading in Grade 1 and Sentence Reading in Grade 2).

We conducted Study 2 in early 2008 as a follow-up, with 72 students (35 first-graders and 37 second-graders) who were native Spanish speakers enrolled in an urban dual language (English/Spanish) bilingual immersion program. As with the initial study, all literacy instruction for these students had been provided in Spanish. Aside from two minor changes in test administration, by which the timing of the Phoneme and Syllable Segmenting measures was reduced to 30 seconds to avoid potential ceiling effects and the Sentence Reading measure was added to the five measures taken by first-grade students, all other methodological features of the two studies were identical.

Detailed results for the correlation and regression studies associated with the development of Spanish literacy measures are reported in a technical report by Alonzo, Gonzalez, and Tindal (2013). Across the two test development studies, findings suggested that the Syllable Sounds test was a good initial measure to use to track the progress of students receiving literacy instruction in Spanish regardless of grade and native language grouping. In both studies, students' performance on the Syllable Sounds measure was strongly correlated with Word Reading. Performance on Syllable Sounds was also strongly predictive of performance on both the Word Reading and Sentence Reading measures. Syllable Segmenting was also found to be an indicator of reading ability for students receiving instruction in Spanish. Performance on the Syllable Segmenting test was correlated with scores on Word Reading for second-grade native Spanish speakers. Similarly, for second-grade students who spoke English as their native

language, a combination of Syllable Sounds and Syllable Segmenting was the strongest predictor of performance on the Sentence Reading assessment.

Together, findings across the two development studies, along with anecdotal statements by participating teachers, suggests that syllable knowledge is an important pre-reading skill for students learning to read in Spanish. Along these lines, measures assessing smaller units of language, specifically the Letter Sounds and Phoneme Segmenting assessments, did not correlate as strongly with measures of reading fluency as the corresponding syllable-level measures. These findings suggest that such skills may be less important for students receiving literacy instruction in Spanish. Additional research in this area is, however, warranted given that evidence of some relation between the Letter Sounds and Phoneme Segmenting measures was observed across the two participating student populations.

We used the results of these two studies to inform the development of alternate forms (three seasonal benchmark tests and 10 progress monitoring tests) of two measure types at the Kindergarten level (Syllable Segmenting and Syllable Reading), four measure types at the first-grade level (Syllable Segmenting, Syllable Reading, Word Reading, and Sentence Reading) and two measures at the second-grade level (Word Reading and Sentence Reading). Future studies will be conducted to provide further evidence of the measures' technical adequacy.

References

Alonzo, J., Gonzalez, M., & Tindal, G. (2013). The development of easyCBM Spanish literacy assessments for Use in Grades K-2 (technical report 1301). Eugene, OR: Behavioral Research and Teaching, University of Oregon.