Technical Report # 08-05

Instrument Development Procedures for Rapid Reading Rate Measures

Kimy Liu

Kristy Carling

Leanne Ketterlin Geller

Gerald Tindal

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

This research was supported by Project INFORM (H327B050013-07) from Office of Special Education Programs, U.S. Department of Education. BRT is affiliated with the College of Education, University of Oregon. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

Abstract

In this study, we describe the development of rapid reading measures, sentences presented to students in a nearly subliminal manner, with a literal comprehension question asked following their removal. After administering alternate forms of these measures to students, we present the results from three statistical analyses to ascertain their technical adequacy: classical test theory, 1-parameter Item Response Theory (IRT), and 2-parameter IRT. We conclude by noting considerable similarity of the results and recommend all analyses as offering valuable information.

Introduction

Perfetti (1999) links reading comprehension not only to the orthographic and phonological skills readers have mastered but also to their lexicon (the words in their vocabulary), underlying language processing mechanisms, their familiarity with the syntax of the language they are reading, and their general knowledge, independent of the activity of reading. "Successful word reading occurs," Perfetti (1999) explains, "when visual input from a string of letters, activating one or more word forms in the lexicon, results in the word corresponding to the letter string (rather than some other word) being identified" (p. 170-171). According to Perfetti's blueprint of the reader, reading comprehension involves the complementary processes of using the basic skills of alphabetic principle and phonological awareness to decipher printed text and then engaging language processing mechanisms (such as word form identification, parsing individual words into phrases and sentences, and constructing understanding of extended text) to create meaningful messages.

This process is iterative, with feedback and feed-forward loops, deficits in any of the sub-skills that comprise reading will be reflected in the general outcome of reading comprehension. Moreover, as readers move from deciphering single sentences to reading lengthier and more complicated text, their ability to form flexible mental representations of what they are reading becomes increasingly important. When reading narrative text, readers attend to five dimensions-time, space, protagonist, causality, and intentionality to create meaning (Zwaan, Langston, & Graesser, 1995). Readers begin constructing mental representations of the text they are reading with the first word they encounter. In skilled readers, these representations remain fluid, allowing readers to adapt their interpretation based on new information they encounter as they move through the passage. Less-skilled readers may be more prone to hold on to their initial

interpretations of text meaning even when they encounter contradictory information, thus reducing their overall comprehension, as early misinterpretations compound the difficulty of constructing meaning from later text. Clearly, this perspective of reading requires rapid recognition of letters, words, and sentences concurrent with associated vocabulary and associative meaning and the primacy of short-term memory as critical components in the process. Without any of these components, comprehension would not proceed in an orchestrated manner.

To build a measure that would be curriculum-based, we developed the Rapid Reading Measures and gave them to students in grades 3 through 8 to determine their reading accuracy and speed. The lengths of sentences ranged from 9 to 21 words. Each sentence appeared on the screen for a pre-determined amount of time and was immediately followed by a multiple-choice question. Students could not return to the sentence after it disappeared. The question was directly related to the information presented in the sentence. Once the student answered a question and advanced to the next sentence, they could no longer return to the previously answered question. The measure was designed so that students who complete reading sentences before the time elapsed should be able to correctly answer the question that follows.

An item writer created ten sentences with corresponding multiple-choice questions for each grade level, grades 3 through 8. The sentences were written at grade level according to the Flesch-Kincaid Readability index and ranged in length from 9 to 21 words (with shorter sentences for lower grades and longer sentences for higher grades). Following each sentence, a multiple-choice question was written that could be answered by reading the entire sentence. Four multiple-choice options, the correct answer and three distracters followed each question. The answer to the question was found in the end of the sentence for six of the 10 sentences and

in the middle of the sentence for four of the 10 sentences. Two of the three distracters were to be near distracters, and the remaining one was to be a far distracter.

Methods

Settings

Rapid Reading Rate measures were designed for administration to students in grades 3-8. After internal and external reviews, they were administered to students attending public schools in two mid-size towns of the Pacific Northwest: 239 Grade 3 students, 185 Grade 4 students, 273 Grade 5 students, 203 Grade 6 students, 206 Grade 7 students and 210 Grade 8 students participating in this study. The test was administered on the computer. Two trained research assistants administered the test. Students took the Rapid Reading Rate measures along with several other measures of reading and math proficiency. Testing took approximately 45 minutes. Students who need additional time were given additional time to log on the test website later, with the supervision of the classroom teachers.

Instrument Development

Prior to implementation of the computer-based rapid reading measures, an extensive process was used to ensure the sentences and questions would function. The first step in developing the silent reading measure was writing narrative text with the same readability. All sentences were designed to have a basic grammatical structure using direct statements. The second step was to analyze the passages with both an internal and external review.

Internal review procedures and results. After the test items were written and edited, the internal reviewer conducted an internal review to evaluate the measures before distributing for an external review. For each measure, the internal reviewer assessed for readability, grade-level appropriate language, grade-level appropriate concepts, length and flow of sentences, and

possible biases. The internal reviewer reported the range of grade-level readability using the Flesch-Kincaid readability formula as well as the sentence length. She also inspected whether the wording and topics of the sentences were appropriate for the indicated grade level. Next, the internal reviewer commented on any gender, cultural, or linguistic biases for students reading the measures. The internal reviewer made few suggestions for changes in the Rapid Reading Rate measures. All of the sentences and questions were within the appropriate grade-level specifications in terms of readability, sentence length, and appropriateness of language and concepts. She did not encounter bias in any of the sentences. However, she did make some suggestions based on the grammatical structure of the sentences.

Qualifications of the internal reviewer. The internal reviewer was a third year doctoral student in the area of Special Education. She was a general education teacher and finished the special education licensure program. The internal reviewer had experience in developing math curricula and ESL reading curricula. Her in-depth knowledge in reading, instructional design and her work experience with diverse student populations allowed her to provide constructive feedback on our instrument development, particularly in the issues that are related to content validity, clarity of direction, and bias against students with limited English proficiency and students with disabilities.

External review procedures and results. The teachers examined the language and vocabulary of the passages for grade-level appropriateness, the concepts described in the passages for grade-level appropriateness, the clarity of writing, and potential bias. Teachers were also asked to provide additional suggestions if needed. Teachers rated the Rapid Reading Rate sentences on a Likert scale of 1-4 for each criterion. A rating of 1 indicated that the criterion was not at all appropriate, a rating of 2 indicated that the criterion was somewhat

appropriate, and rating of 3 indicated that the criterion was appropriate, and a rating of 4 indicated that the criterion was extremely appropriate. The teachers provided feedback and suggestions to improve the quality of the measures, in addition to rating the measures using the Likert scale. In all, the reviewing teachers found the vocabulary and content were appropriate for the grade-level. The reviewing teachers did not find evidence of bias.

Qualifications of external reviewers. Six teachers working in local schools were selected to review the passages for grade-level appropriateness in terms of content and readability within the grade level in which they were currently teaching.

1. An eighth-grade teacher in a local middle school, who previously taught sixth grade. He had taught for four years and holds a Master of Education degree.

2. A special-education teacher in a local elementary school who had taught for 20 years. She holds a Master of Education degree and previously taught special education in middle school.

3. A fifth-grade teacher in a local elementary school who holds a Master of Education degree and had taught for two years.

4. A teacher who had taught for 28 years and was teaching seventh-grade at a local middle school at the time of the review. She holds a Master of Arts degree and previously taught sixth-grade and third-grade.

5. A reading specialist at a local elementary school who had taught for 23 years. She previously taught fourth- and fifth-grades and was pursuing a Master's Degree in Educational Leadership and Administrative Licensure at the time of the review.

6. A local middle school teacher in grades 6-7 who holds a Master of Education degree and was in his first year of teaching.

Results

We report the results in two sections. We describe the results from both classical and item response theory (IRT) analyses and then describe the final item selection process.

The examinee's responses to the multiple-choice questions were scored dichotomously with observed scores used to calculate estimates of item difficulties. The authors obtained the estimates of item difficulties through three different approaches.

First, we used Classical Test Theory (CTT) model to calculate *p*-values, the proportion of valid responses which were correct (Embretson & Reise, 2000). For example, in the first item of the Grade 3 measures, 142 of 216 valid responses were correctly answered. The *p*-value of this item was .66. All *p*-values of test items have been shown in Appendix A. In general, Grade 5 items had the narrowest range of *p*-values (87% - 97%) while Grade 3 had the widest range (66% - 93%).

Second, the authors obtained estimates of item difficulties using the Rasch Item Response Model (1PL). Examinees' likelihood of answering the questions correctly was determined by estimating item difficulty parameter and examinee's ability level on the same scale. When the person's ability level matched the level of item difficulty, the person was estimated to have a 50% probability of answering the item correctly. If the test item difficulty exceeded the person's ability level, the person's likelihood of answering the question correctly was estimated as less than 50%. The tables in Appendix B have been assembled to display the item difficulties, outfit mean squares and other relevant item characteristics. Items with higher values for item difficulty parameter were more challenging than items with lower values.

We also considered whether the items provided productive information on estimates of a respondent's ability level by examining the outfit mean squares. Items with outfit means squares

between .5 and 1.5 were considered productive test items, because they generated predictable response patterns in which the students who gave correct answers were more skillful than their peers who answered incorrectly. The values of outfit mean square below 0.5 suggested that estimates are "overly predictable" that would mislead test developers to overestimate the quality of the test item (attenuation paradox). The value of outfit mean squares exceeding 2.0 meant that the items did not conform to the expectations and potentially degraded the measure. The value of outfit mean squares between 1.5 and 2.0 reflected noticeable noise in the measure. Although it did not necessarily degrade the measurement, it nevertheless did not provide useful information on the measurement either.

With the exception of two items in the Grade 6 Rapid Reading Rate measure, all other items in Grades 3-8 were deemed productive items because their values of the outfit mean squares fell within the range of .5 and 1.5. In the Grade 6 measures, one item had an outfit mean square value below .5, while another item had an outfit mean square value exceeding 1.5. These items were flagged for removal.

Third, the authors obtained the estimates of item difficulty and item discrimination for each item using a two parameter (2PL) model with BILOG-MG software. The intercepts and item difficulty of the 2PL model were similar to the ones used in the Rasch model: The intercepts were defined as the probability that students with extremely low ability answered the question correctly. The difference of item difficulty and a person's ability level was used to calculate the probability of a person's answer an item correctly. Compared with the Rasch model (1PL), the two-parameter item response model added item discrimination as an additional item parameter, which was expressed by the steepness of the slope. The items that discriminated well were items with steep slopes. The two-parameter model did not constrain the slopes of the

items but allowed the slopes of the items to vary in reflecting that not all items were equally related to the latent abilities. For those items that discriminated well, the incremental changes in ability were smaller while still having the same unit increase in the odds ratio. These values of intercepts, slopes and item difficulty were used to describe the unique item characteristics curve for each item, which was used to calculate the probability of examinees of different skill levels answering the test item correctly. Tables in Appendix C have been organized to display the following item characteristics: intercepts, slopes and item difficulty for each test item as well as the standard errors of the intercepts, slopes and item difficulty, respectively. The slopes of test items in the Rapid Reading Rate measure varied but the variance of slopes was relatively small for almost all test items.

All items in grades 3-5 and grade 8 appropriately fit the 2PL model. One item in Grade 6 (Item 8) was excluded from calibration because the response pattern of this item was negatively correlated with the responses for other items in the same grade. We did not report the parameter scores and scores files for the Grade 7 measures because the data did not converge under 2PL models.

In Appendix D we have reported the scores for the Rapid Reading Rate measures under the 2PL model, including the minimum and maximum of the scale scores as well as the minimum and maximum of the standard errors of scale scores. The scale scores were the calibrated scores on the scale shared by the item difficulty and person's ability level. The probability of a person's likelihood of answering a test item correctly was determined by the item characteristics (i.e. the item difficulty and slope) as well as a person's ability level. The distribution of scores across different scores brackets also have been displayed. When there was only one case within a particular score bracket, the minimum and maximum of scale scores were

identical. The standard errors of the estimated extreme scale scores were larger than the standard errors of average scores. The distributions of all items in Grades 3-6 and 8 were negatively skewed: In each measure, overwhelming majority of examinees received perfect scores.

In summary, the CTT model and the IRT models (1PL model and 2PL model) yielded different representations in the range of item difficulties. Under all the three models, the Grade 5 measure had the narrowest range of item difficulty. However, the grade levels that had the widest range of item difficulty varied, depending on which model was used to obtain the results. In CTT model, Grade 3 had the widest range of *p*-value. In 1PL model and 2PL model, the Grade 6 measure had the widest range of item difficulty. Of note, although results varied, the ranking order of the estimates of item difficulties within each grade was consistent across the three methods.

Cautions should be applied in using the *p*-value to estimate item difficulty because *p*-value is "population dependent." In other words, given the same items, a more skillful group may have lower estimates of item difficulties; whereas, a less skillful group may have higher estimates of item difficulties. Using CTT model to estimate the item difficulty is appropriate given the goals of this project. Furthermore, in CTT model, the true scores only applies to a specific set of items or their equivalent. Item difficulty is influenced by the respondents' overall ability levels; the item difficulty and respondent's ability levels can not be estimated separately (Embretson & Reise, 2000). In contrast, IRT models do not have these same technical limitations.

At the same time, IRT models have several technical advantages over the CTT model. The Rasch model (1PL model) predicts probability of respondents' giving correct responses from

two independent variables (respondent's ability level and item difficulty). The 2PL model predicts probability of respondents' giving correct responses from two item parameters (item difficulty and slopes, controlling for guessing) and a person's ability level. Estimates of person ability levels can be obtained in a non-representative sample population (sample invariance). The comparison of item difficulties, therefore, is meaningful between two different subtests from the same item bank because IRT models (including Rasch model) place the ability level and item difficulty at the same scale.

In the IRT models, the items provide most informative estimates on the respondents' ability level when an item was "on target," or the distance between the person's estimated ability level and item difficulty of the selected item is narrow. When the items were "on target," the measurement errors are relatively small. Therefore, the estimates are more accurate. Providing items with wide range of item difficulty increased the likelihood of having items that are on target for respondents' ability levels. Based on these considerations, we propose using IRT models (1PL and 2PL models) to obtain the estimates of item difficulties and respondents' ability levels, with the range of item difficulty used as a criterion to evaluate the effectiveness of the test bank.

Results from the 2PL model were used as the primary data source for selecting items for the final test bank. Item difficulty and discrimination parameters were evaluated when selecting the final items. The levels of item difficulty of all test items under 2PL models have been reported in the parameter files in Appendix C, along with the slopes, item difficulty as well as the standard errors of slopes and item difficulty. The slope of the item was determined by the amount of change in skill level needed to reach 50-50 probabilities (the level at which the person can answer the question correctly 50% of the time). Given the same amount of skill level

improvement, people show greater improvement (in terms of likelihood of answering the question correctly on the items) with steeper slopes than on the items with less steep slopes.

The response pattern of the second item was considered overly predictable because the gap of a person's ability levels between the correct responses and incorrect responses was large. This might lead to an overestimate in the quality of this item. The response pattern of the sixth item was troublesome because all but one person gave correct answers. However, the authors had no evidences to believe these two aberrant items negatively influenced the overall quality of the items.

Next, the authors inspected the scale scores and item characteristics of each items. The Rasch model constrained the slopes to be identical (slope =1) for all items. Under the 2PL model, the slopes were allowed to be varied rather than forcing the data to fit the model. The review of parameter models in Appendix D indicated that Item 8 of Grade 6 was excluded because the response pattern was negatively correlated with the rest of items in the same grade. All the other items were included for calibration; they all had similar, but not identical slopes. The item difficulty levels were all below 0.00 on the scale between -4 to +4, which indicated there are no difficult items. The distribution of examines was negatively skewed.

Standard errors of item difficulty and slopes of the test items varied. The score files provided estimates of respondents' ability levels (calibrated scale scores) based on their raw scores from the 2PL. The scale scores varied within each raw score bracket. For example, there were 216 students responding to the third grade measures. Seventy-four students answered 9 of 10 multiple-choice questions correctly, however, they received different scale scores ranging from -0.33 to +0.15. The standard errors of the scale scores fell within the range of .63 and .72. The estimated scale scores and standard error of scale score under the 2PL model varied

depending on which answer the examinee answered correctly. Because there were no anchoring items or anchoring persons in the design of the test items and the administration of the tests, it was not appropriate to compare the item difficulties and examinees' skills levels across the grades (Embretson & Reise, 2000).

These results were used to select the final set of Rapid Reading Rate items that contributed to the IEP Decision-Making framework for Project INFORM with two considerations addressed: range of item difficulty and timing of administration. The range of item difficulty was restricted in these piloted items due to the extended timing of administration. When administered in the IEP Decision-Making framework, the Rapid Reading Rate measure needs to be timed, as it relates to students' proficiency. For example, the teacher may choose to administer the Rapid Reading Rate measure at the 50^{th} percentile or 75^{th} percentile rank (based on National Norms) depending on the student's level of skills. By restricting the timing of administration, we hypothesize that the difficulty of the item increases. However, for the piloting of these items, we chose to administer the items at the 20th percentile; in other words, the sentences were displayed for an extended amount of time to make sure that the students had ample opportunity to read the sentence and respond to the question. Because of the extended timing of this administration, we have hypothesized that the item difficult was influenced, thereby making the items too easy.

Six items were selected to serve as the final items for each grade level by considering item difficulty and placement of correct answer in the question stem. Items were selected to represent a range of difficulty levels. Four of the six items were selected that had the answer to the multiple-choice question at the end of the question stem. The remaining two items had the answers in the middle of the question stem.

Discussion

The Rapid Reading Rate measures were designed to assess student's speed and accuracy of reading. Speed is one of the pre-requisite task for answering the follow-up question correctly, because all the answers are embedded either in the middle or at the end of the sentences.

We compared three models of analyzing students' response patterns. The CTT model did not provide technically adequate estimates of item difficulty, because p-value (proportion correct) under the CTT model was population dependent. In theory, the Rasch model was not technically appropriate for Rapid Reading Rate measures, because speed was not accounted for in the 1PL model. The 2PL model, however, was thought to be a viable option, because the item discrimination reflected the influences of time elapsed on the response patterns.

These IRT models, however, should be taken with some cautions. First, the assumption of local independence often is not met (Embretson & Reise, 2000). Also, speed is not accounted for in the model. Second, there are no anchoring items or persons in the design and administration of the test, so the readers should not compare the item difficulties across the grade levels. Last, the current reading measures are relatively easy. For the students who receive perfect scores, the estimates of their skill levels are not as trustworthy as the estimates of students who have predictable response patterns that include both correct and incorrect responses. The authors recommend embedding some off-grade-level items in the test banks for the future item writing. For instance, in the fourth grade measures, two items from Grades 5 and one item from Grade 6 measures should be included in the test bank. One of the dual-purposes of embedded off-grade-level items is to expand the range of item difficulty so that students with diverse skill levels can all receive "on-target" estimates of their abilities. The other purpose is to create anchoring items, so that the items and students can be compared across the grade levels.

The strength of the current design is utilizing the computer adaptive testing platform. The tests are short, easy and quick to score. They can be administered in groups, which require less time for administering in comparison with individually administered tests. The Rapid Reading Rate measures can be used as initial screening measures in schools. When the appropriately calibrated alternate forms are created, this measure can also be used for screening, progress monitoring as well as end-of-the year tests.

References

Embretson, S. E. & Reise, S. R. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 167-208). Oxford University Press.

Zwaan, R. Langston, M., & Graesser, A. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6,* 292-297.

Appendix A: Estimates of item difficulty under the Classical Test Theory (CTT) Model

Table A1.
Grade 3 Data.

	1	2	3	4	5	6	7	8	9	10
Correct	142	201	199	190	197	199	200	163	192	199
Incorrect	74	15	15	25	18	16	15	52	23	15
Valid	216	216	215	215	215	215	215	215	215	214
Missing	23	23	24	24	24	24	24	24	24	25
***Valid Percent**	.66	.93	.93	.88	.92	.93	.93	.76	.89	.93

Table A2.
Grade 4 Data.

	1	2	3	4	5	6	7	8	9	10
Correct	136	151	153	160	128	162	162	159	155	155
Incorrect	33	18	6	9	41	7	7	10	13	12
Valid	169	169	169	169	169	169	169	169	168	167
Missing	16	16	16	16	16	16	16	16	17	18
Valid Percent	.80	.89	.91	.95	.76	.96	.96	.94	.92	.93

Table A3.
Grade 5 Data.

	1	2	3	4	5	6	7	8	9	10
Correct	219	236	239	242	228	246	244	242	238	242
Incorrect	32	15	12	9	23	5	7	9	13	9
Valid	251	251	251	251	251	251	251	251	251	251
Missing	22	22	22	22	22	22	22	22	22	22
Valid Percent	.87	.94	.95	.96	.91	.98	.97	.96	.95	.96

Table A4.
Grade 6 Data.

	1	2	3	4	5	6	7	8	9	10
Correct	128	161	164	160	164	165	158	160	153	162
Incorrect	38	5	2	6	2	1	8	8	13	4
Valid	166	166	166	166	166	166	166	166	166	166
Missing	37	37	37	37	37	37	37	37	37	37
Valid Percent	.77	.97	.99	.96	.99	.99	.95	.96	.92	.98

Table A5.
Grade 7 Data.

	1	2	3	4	5	6	7	8	9	10
Correct	150	181	184	174	175	170	178	184	179	164
Incorrect	36	5	2	12	11	16	8	2	7	21
Valid	186	186	186	186	186	186	186	186	186	185
Missing	20	20	20	20	20	20	20	20	20	21
Valid Percent	.81	.97	.99	.94	.94	.91	.97	.96	.96	.89

Table A6.
Grade 8 Data.

	1	2	3	4	5	6	7	8	9	10
Correct	151	156	174	168	177	171	166	157	168	153
Incorrect	30	25	7	13	4	10	13	22	11	25
Valid	181	181	181	181	181	181	179	179	179	178
Missing	29	29	29	29	29	29	31	31	31	32
Valid Percent	.83	.86	.96	.93	.98	.94	.93	.88	.94	.86

Appendix B: Item Parameters Using the Rasch Item Response Model

Table B1.
Grade 3 Rapid Reading.

ENTRY	Difficulty	COUNT	SCORE	OUT. MSQ	OUT. ZSTD	PTME	OBS MATCH	EXP MATCH
1	72.16	132	58	1.19	1.46	0.59	62.9	67.1
2	43.46	132	117	0.72	-0.78	0.45	89.4	89.3
3	42.52	132	118	0.85	-0.32	0.43	90.2	89.9
4	50.94	132	107	0.86	-0.57	0.55	85.6	83.9
5	46.02	132	114	1.23	0.81	0.38	86.4	87.7
6	44.36	132	116	0.71	-0.86	0.46	90.2	88.7
7	43.46	132	117	0.81	-0.46	0.47	90.9	89.3
8	63.93	132	80	1.24	2.08	0.52	65.2	72.3
9	49.65	132	109	0.84	-0.64	0.54	88.6	84.9
10	43.50	131	116	0.75	-0.68	0.44	89.3	89.3

Table B2.
Grade 4 Rapid Reading.

ENTRY	MEASURE	COUNT	SCORE	OUT. MSQ	OUT. ZSTD	PTME	OBS MATCH	EXP MATCH
1	61.67	95	62	1.17	1.37	0.47	64.2	71.2
2	52.04	95	77	0.90	-0.36	0.47	83.2	82.6
3	60.00	95	65	0.95	-0.33	0.56	73.7	72.6
4	42.75	95	86	0.44	-1.66	0.49	92.6	91.1
5	65.81	95	54	1.26	2.27	0.51	58.9	67.1
6	39.63	95	88	0.63	-0.74	0.44	93.7	92.9
7	39.63	95	88	0.70	-0.55	0.35	93.7	92.9
8	44.09	95	85	0.96	0.00	0.38	91.6	90.2
9	47.70	94	81	0.82	-0.53	0.44	86.2	87.2
10	46.69	93	81	1.25	0.81	0.32	87.1	88.0

Table B3.
Grade 5 Rapid Reading.

ENTRY	MEASURE	COUNT	SCORE	OUT. MSQ	OUT. ZSTD	PTME	OBS MATCH	EXP MATCH
1	63.99	91	59	0.99	-0.04	0.57	64.80	68.80
2	52.92	91	76	1.23	1.02	0.34	83.50	84.00
3	50.08	91	79	1.06	0.32	0.35	84.60	87.00
4	46.57	91	82	0.69	-0.93	0.44	91.20	90.10
5	58.82	91	68	1.03	0.28	0.47	72.50	76.30
6	39.80	91	86	1.01	0.19	0.26	94.50	94.50
7	43.62	91	84	1.00	0.13	0.30	92.30	92.30
8	46.57	91	82	1.10	0.41	0.33	91.20	90.10
9	51.08	91	78	0.87	-0.44	0.45	85.70	86.00
10	46.57	91	82	0.93	-0.11	0.37	89.00	90.10

Table B4.
Grade 6 Rapid Reading.

ENTRY	MEASURE	COUNT	SCORE	OUT. MSQ	OUT. ZSTD	PTME	OBS MATCH	EXP MATCH
1	79.07	65	27	1	0.01	0.69	58.5	58.4
2	48.8	65	60	0.44	-1.39	0.42	92.3	92.3
3	38.82	65	63	0.9	0.14	0.23	96.9	96.9
4	50.9	65	59	1.01	0.16	0.3	90.8	90.7
5	38.82	65	63	1.1	0.39	0.15	96.9	96.9
6	31.6	65	64	1.66	0.86	0.07	98.5	98.5
7	54.3	65	57	0.87	-0.29	0.38	87.7	87.7
8	50.9	65	59	1.38	0.96	0.2	90.8	90.7
9	60.48	65	52	1	0.07	0.43	81.5	80.4
10	46.3	65	61	1.08	0.32	0.22	93.8	93.8

Table B5.
Grade 7 Rapid Reading.

ENTRY	MEASURE	COUNT	SCORE	OUT. MSQ	OUT. ZSTD	PTME	OBS MATCH	EXP MATCH
1	71.49	78	42	1.45	3.83	0.53	55.1	64.9
2	42.66	78	73	0.96	0.11	0.36	94.9	93.8
3	32.13	78	76	1.18	0.5	0.26	97.4	97.4
4	53.79	78	66	0.81	-0.63	0.49	85.9	85.8
5	52.62	78	67	0.87	-0.34	0.45	87.2	86.9
6	57.83	78	62	0.78	-1.04	0.57	84.6	81.5
7	32.13	78	76	1.24	0.55	0.2	97.4	97.4
8	48.47	78	70	0.91	-0.11	0.42	91	90.3
9	46.78	78	71	0.72	-0.58	0.42	92.3	91.4
10	62.09	77	56	1.29	1.67	0.42	70.1	76.1

Table B6.
Grade 8 Rapid Reading.

ENTRY	MEASURE	COUNT	SCORE	OUT. MSQ	OUT. ZSTD	PTME	OBS MATCH	EXP MATCH
1	60.89	94	64	1.13	1.07	0.46	66	72.5
2	57.96	94	69	1.06	0.45	0.47	76.6	76.2
3	41.2	94	87	0.79	-0.38	0.33	92.6	92.5
4	48.77	94	81	0.86	-0.42	0.43	87.2	86.3
5	34.89	94	90	0.9	0.03	0.27	95.7	95.7
6	45.47	94	84	0.92	-0.11	0.34	88.3	89.4
7	49.09	93	80	0.91	-0.24	0.43	87.1	86.2
8	56.35	93	71	0.88	-0.7	0.53	80.6	78.4
9	46.96	93	82	1.01	0.14	0.37	89.2	88.2
10	58.42	92	67	1.12	0.8	0.43	72.8	75.7

Appendix C: Item Parameters Using the 2PL Model

Table C1.
Grade 3 Rapid Reading.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	0.45	0.10	0.54	0.14	-0.84	0.28
Item 02	2.37	0.44	1.13	0.33	-2.10	0.36
Item 03	2.17	0.38	1.03	0.28	-2.12	0.38
Item 04	1.70	0.23	0.84	0.21	-2.02	0.36
Item 05	1.73	0.22	0.72	0.18	-2.38	0.49
Item 06	2.27	0.42	1.14	0.32	-1.99	0.32
Item 07	2.31	0.37	1.03	0.26	-2.26	0.34
Item 08	0.79	0.11	0.52	0.13	-1.53	0.40
Item 09	1.84	0.29	0.99	0.24	-1.87	0.30
Item 10	2.17	0.37	1.03	0.28	-2.10	0.35

Table C2.
Grade 4 Rapid Reading.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	0.95	0.12	0.31	0.10	-3.10	1.02
Item 02	1.73	0.21	0.52	0.17	-3.33	0.98
Item 03	1.26	0.19	0.72	0.21	-1.74	0.41
Item 04	11.09	5.70	7.93	3.72	-1.40	0.20
Item 05	0.76	0.13	0.48	0.13	-1.58	0.45
Item 06	3.04	0.53	0.98	0.32	-3.09	0.67
Item 07	2.55	0.40	0.88	0.32	-2.90	0.82
Item 08	1.90	0.24	0.54	0.17	-3.51	0.93
Item 09	2.45	0.45	1.19	0.36	-2.06	0.38
Item 10	1.85	0.24	0.64	0.21	-2.88	0.83

Table C3.
Grade 5 Rapid Reading.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	1.42	0.17	0.61	0.19	-2.33	0.59
Item 02	1.82	0.19	0.56	0.18	-3.24	0.93
Item 03	2.08	0.25	0.64	0.21	-3.25	0.89
Item 04	3.12	0.80	1.34	0.52	-2.32	0.40
Item 05	1.56	0.17	0.55	0.17	-2.83	0.75
Item 06	2.90	0.45	0.83	0.32	-3.48	1.01
Item 07	2.47	0.33	0.73	0.24	-3.37	0.89
Item 08	2.36	0.28	0.66	0.23	-3.56	1.02
Item 09	2.27	0.36	0.96	0.30	-2.37	0.48
Item 10	2.52	0.38	0.84	0.28	-2.99	0.70

Table C4.
Grade 6 Rapid Reading.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	0.90	0.18	0.78	0.25	-1.16	0.29
Item 02	3.42	0.97	1.13	0.54	-3.02	0.71
Item 03	3.45	0.81	1.13	0.49	-3.05	0.95
Item 04	2.24	0.27	0.55	0.20	-4.07	1.34
Item 05	3.13	0.50	0.73	0.31	-4.30	1.62
Item 06	3.53	0.63	0.75	0.34	-4.69	2.06
Item 07	2.09	0.33	0.76	0.29	-2.77	0.78
Item 08	0.00	0.00	0.00	0.00	0.00	0.00
Item 09	1.95	0.26	0.76	0.26	-2.58	0.71
Item 10	2.49	0.35	0.66	0.26	-3.78	1.28

Table C5.
Grade 8 Rapid Reading.

	Intercept	Intercept SE	Slope	Slope SE	Difficulty	Difficulty SE
Item 01	1.16	0.15	0.54	0.16	-2.13	0.59
Item 02	1.28	0.19	0.76	0.21	-1.69	0.39
Item 03	2.26	0.29	0.63	0.22	-3.58	0.97
Item 04	1.94	0.27	0.68	0.22	-2.85	0.71
Item 05	2.68	0.39	0.71	0.26	-3.78	1.14
Item 06	2.10	0.31	0.76	0.26	-2.75	0.76
Item 07	1.90	0.30	0.87	0.25	-2.18	0.45
Item 08	1.55	0.21	0.68	0.19	-2.28	0.47
Item 09	1.80	0.27	0.69	0.24	-2.60	0.76
Item 10	1.27	0.14	0.41	0.12	-3.12	0.87

Appendix D: Item Scale Scores Using the 2PL Model

Table D1.
Grade 3 Rapid Reading Score Files.

Total Items	Number of Correct Items	Proportion Correct	Number of cases	Scale Scores		Standard Errors of Scale Scores	
				Minimum	Maximum	Minimum	Maximum
10	10	100	83	0.67	0.67	0.81	0.81
10	9	90	74	-0.33	0.15	0.63	0.72
10	8	80	24	-0.93	0.11	0.53	0.73
10	7	70	15	-1.33	-0.86	0.4	0.54
10	6	60	7	-1.46	-1.2	0.4	0.43
10	5	50	4	-1.82	-1.5	0.43	0.47
10	4	40	3	-1.8	-1.77	0.47	0.47
10	3	30	5	-2.3	-2.13	0.33	0.36
10	2	20	1	0.35	0.35	0.92	0.92
10	1	10	0	--	--	--	--
10	0	0	0	--	--	--	--

Table D2.
Grade 4 Rapid Reading Score Files.

Total Items	Number of Correct Items	Proportion Correct	Number of cases	Scale Scores		Standard Errors of Scale Scores	
				Minimum	Maximum	Minimum	Maximum
10	10	100	74	.57	.57	.85	.85
10	9	90	50	-.56	.22	.64	.80
10	8	80	21	-1.44	.01	.32	.77
10	7	70	14	-1.70	-.64	.41	.62
10	6	60	6	-1.98	-.50	.28	.68
10	5	50	2	-2.06	-1.29	.16	.42
10	4	40	0	--	--	--	--
10	3	30	1	-2.49	-2.49	.46	.46
10	2	20	1	-2.60	-2.60	.49	.49
10	1	10	0	--	--	--	--
10	0	0	0	--	--	--	--

Table D3.
Grade 5 Rapid Reading Score Files.

Total Items	Number of Correct Items	Proportion Correct	Number of cases	Scale Scores		Standard Errors of Scale Scores	
				Minimum	Maximum	Minimum	Maximum
10	10	100	160	.38	.38	.89	.89
10	9	90	65	-1.00	-.28	.66	.80
10	8	80	16	-1.40	-.85	.59	.69
10	7	70	6	-1.89	-1.43	.53	.59
10	6	60	2	-2.21	-1.95	.45	.51
10	5	50	1	-2.12	-2.12	.47	.47
10	4	40	1	-2.67	-2.67	.49	.49
10	3	30	0	--	--	--	--
10	2	20	0	--	--	--	--
10	1	10	0	--	--	--	--
10	0	0	0	--	--	--	--

Table D4.
Grade 6 Rapid Reading Score Files.

Total Items	Number of Correct Items	Proportion Correct	Number of cases	Scale Scores		Standard Errors of Scale Scores	
				Minimum	Maximum	Minimum	Maximum
9	9	100	106	0.4	0.4	0.88	0.88
9	8	88.89	45	-0.85	-0.27	0.73	0.81
9	7	77.78	12	-1.47	-1	0.65	0.71
9	6	66.67	2	-1.95	-1.59	0.57	0.63
9	5	55.56	1	-2.51	-2.51	0.53	0.53
9	4	44.44	0	--	--	--	--
9	3	33.33	0	--	--	--	--
9	2	22.22	0	--	--	--	--
9	1	11.11	0	--	--	--	--
9	0	0	0	--	--	--	--

Table D5.
Grade 8 Rapid Reading Score Files.

Total Items	Number of Correct Items	Proportion Correct	Number of cases	Scale Scores		Standard Errors of Scale Scores	
				Minimum	Maximum	Minimum	Maximum
10	10	100	86	.55	.55	.86	.86
10	9	90	56	-.38	.08	.73	.79
10	8	80	22	-.92	.26	.65	.75
10	7	70	7	-1.40	-1.12	.58	.61
10	6	60	3	-1.58	.39	.57	.90
10	5	50	5	-2.13	-1.92	.52	.55
10	4	40	1	-2.22	-2.22	.51	.51
10	3	30	0	--	--	--	--
10	2	20	0	--	--	--	--
10	1	10	0	--	--	--	--
10	0	0	0	--	--	--	--
*6	6	100	1	.39	.39	.90	.90
*6	4	66.67	1	-1.21	-1.21	.72	.72

Appendix E: Final Item Order

Notes for interpretation:

1. Poorly fitting items were omitted from the item pool
2. Items with redundant item difficulties were removed
3. A range of difficulties were kept
4. Items were organized to follow the pattern: E, M, E, E, M, E
5. Difficulty of items in final mapping was mostly random; most difficult item never appeared first

Table E1.

Grade	Order 1	Order 2	Order 3	Order 4	Order 5	Order 6
3	Original Item 3	Original Item 2	Original Item 1	Original Item 4	Original Item 8	Original Item 9
4	Original Item 7	Original Item 2	Original Item 9	Original Item 5	Original Item 3	Original Item 1
5	Original Item 8	Original Item 4	Original Item 3	Original Item 7	Original Item 1	Original Item 2
6	Original Item 10	Original Item 9	Original Item 4	Original Item 8	Original Item 1	Original Item 3
7	Original Item 5	Original Item 10	Original Item 7	Original Item 6	Original Item 1	Original Item 9
8	Original Item 4	Original Item 6	Original Item 5	Original Item 1	Original Item 2	Original Item 8