SPI SCHOLASTIC PHONICS INVENTORY™

# Technical Guide

## ASSESSING PHONICS AND SIGHT WORDS FOR OLDER STRUGGLING READERS

**Dr. Richard K. Wagner**

In partnership with the
Scholastic Research & Validation Department

TM & © Scholastic Inc. All rights reserved.

Copyright © 2009 by Scholastic Inc.

All rights reserved. Published by Scholastic Inc. Printed in the U.S.A.

PDF0016
ISBN-13: 978-0-545-11679-4 (paperback)
ISBN-10: 0-545-11679-1 (paperback)

SCHOLASTIC, SCHOLASTIC PHONICS INVENTORY, SCHOLASTIC READING INVENTORY, SYSTEM 44, READ 180, and associated logos and designs are trademarks and/or registered trademarks of Scholastic Inc. LEXILE and LEXILE FRAMEWORK are registered trademarks of MetaMetrics, Inc. Other company names, brand names, and product names are the property and/or trademarks of their respective owners.

# TABLE OF CONTENTS

Introduction

Administration and Scoring

Score Reporting and Interpretation

Development of the Scholastic Phonics Inventory

Validity of Scholastic Phonics Inventory Scores

TM & © Scholastic Inc. All rights reserved.

## LIST OF TABLES

## LIST OF FIGURES

TM & © Scholastic Inc. All rights reserved.

# INTRODUCTION

## Overview

The *Scholastic Phonics Inventory* (SPI) was designed to measure fluency for two word-level reading skills: phonological decoding and sight word reading. Phonological decoding fluency is assessed by the speed and accuracy with which pronounceable nonwords are decoded. Sight word fluency is assessed by the speed and accuracy with which high-frequency words are read. The SPI is administered individually via a personal computer in approximately 10 minutes.

## Uses

The SPI was developed to identify 3rd–12th grade students who are poor decoders and/or unable to recognize sight words with fluency, and to differentiate these students from those who are adequate decoders and able to recognize sight words with fluency. Within the poor decoder category, the SPI further places students who need instruction in foundational phonological decoding skills, starting with Series 1 of the *System 44* software, separate from those students who need instruction in basic phonological decoding skills, starting with Series 4 of the *System 44* software.

TM & © Scholastic Inc. All rights reserved.

## Rationale

Phonological decoding at the word level is a building block upon which fluent single-word reading and fluent reading of connected text for comprehension are based, and an important predictor of reading comprehension. The SPI uses nonword-reading fluency as an effective measure for evaluating phonological decoding. When presented with a nonword, readers must break it into parts, retrieve sounds associated with the parts, and string them together to pronounce the unfamiliar word. This process is assessed with the SPI by presenting examinees with pronounceable nonwords.

A related element that contributes to fluency is sight word knowledge. Skilled readers have a large vocabulary of sight words that can be recognized automatically. However, developing a large vocabulary of sight words is largely dependent on the reader's ability to decode efficiently. Skilled readers analyze unfamiliar words or nonwords more fully than poor readers do. For example, some poor readers tend to use initial consonant cues to guess at the rest of the word. A full analysis of unfamiliar words contributes to their becoming sight words over time. With repeated, accurate reading of the same word, the word eventually becomes stored in memory as a sight word—one that is identified automatically and without conscious thought.

The more accurate and automatic readers become with these word-level reading processes, the more cognitive resources become available for comprehending strings of text. In fact, for elementary-age students, word-level reading has been found to be a major determinant of reading comprehension (Jenkins et al., 2003; Stanovich, 1991).

Difficulties with word-level reading become increasingly problematic as students get older. Problems with phonological decoding and sight word fluency result in poor comprehension and lower motivation (Snow, Burns, & Griffin, 1998), and as texts become increasingly advanced with each grade, poor readers fall further and further behind. Recent studies of struggling adolescent readers in urban schools indicate that over half are deficient in word-level reading skills (Hock et al., in press).

TM & © Scholastic Inc. All rights reserved.

# ADMINISTRATION AND SCORING

## Administration

The SPI is administered individually via a personal computer in approximately 10 minutes. To log in, students are instructed to enter their name and password on the log-in screen and then click the **Go On** button to begin. Students will follow the audio directions to begin the first section of the SPI. During all sessions of the assessment, students can access the **Pause**, **Play**, and **Replay** buttons. Once a student answers the last SPI question, he/she will be asked to click on the **Go On** button to complete the test and exit.

## Scoring

With respect to scoring, both fluency (i.e., speed and accuracy) and accuracy are assessed for sight words and nonwords. Fluency is important because it frees the reader to attend to comprehension. If a student is accurate but slow, it is likely that reinforcement of basic skills along with ongoing practice and corrective feedback will increase word-level fluency. If a student is fluent with nonwords but not fluent with sight words, a plausible explanation is good phonological decoding skills but limited knowledge of the English vocabulary being assessed. On the other hand, if a student is fluent with sight words but not fluent or inaccurate with nonwords, the explanation may be an extensive sight word vocabulary along with a lack of basic decoding skills.

TM & © Scholastic Inc. All rights reserved.

# SCORE REPORTING AND INTERPRETATION

### Reporting

The SPI generates a Screening and Placement Report (see Figure 1) that includes the following information:

- SPI Test Date
- Percent Accurate and Fluent on SPI subtests
- Recommended Placement
- Decoding diagnosis (Pre-Decoder, Beginning Decoder, Developing Decoder, Advancing Decoder, and Proficient Decoder)

A Lexile score obtained from the Scholastic Reading Inventory (SRI) is also included in the report if it is available.

### Placement and Diagnosis

The recommended placement and decoding diagnosis are determined as follows:

- Criteria for a Fluent Response: Responses on the sight word and nonword items are labeled fluent if they are accurate AND if they are produced within a time limit, also known as the fluency threshold.
- Recommended Placement and Diagnosis
    - If a student's total fluency score is less than 37, the student is placed in **Series 1**[1] of the *System 44* software.

[1] For placement purposes, students are placed in either Series 1 or Series 4 of *System 44* software. Altogether there are 25 series. In order to evaluate whether advanced skills are mastered, FastTrack assessments occur at the beginning of each series, starting at Series 4.

TM & © Scholastic Inc. All rights reserved.

  - If the student can identify fewer than 70 percent of the letter names[2] OR fluently decode fewer than 30 percent of the consonants and vowels in nonwords, then the student is diagnosed as a **Pre-Decoder**.
  - If the student can identify at least 70 percent of the letter names OR fluently decode at least 30 percent of the consonants and vowels in nonwords, then the student is diagnosed as a **Beginning Decoder**.

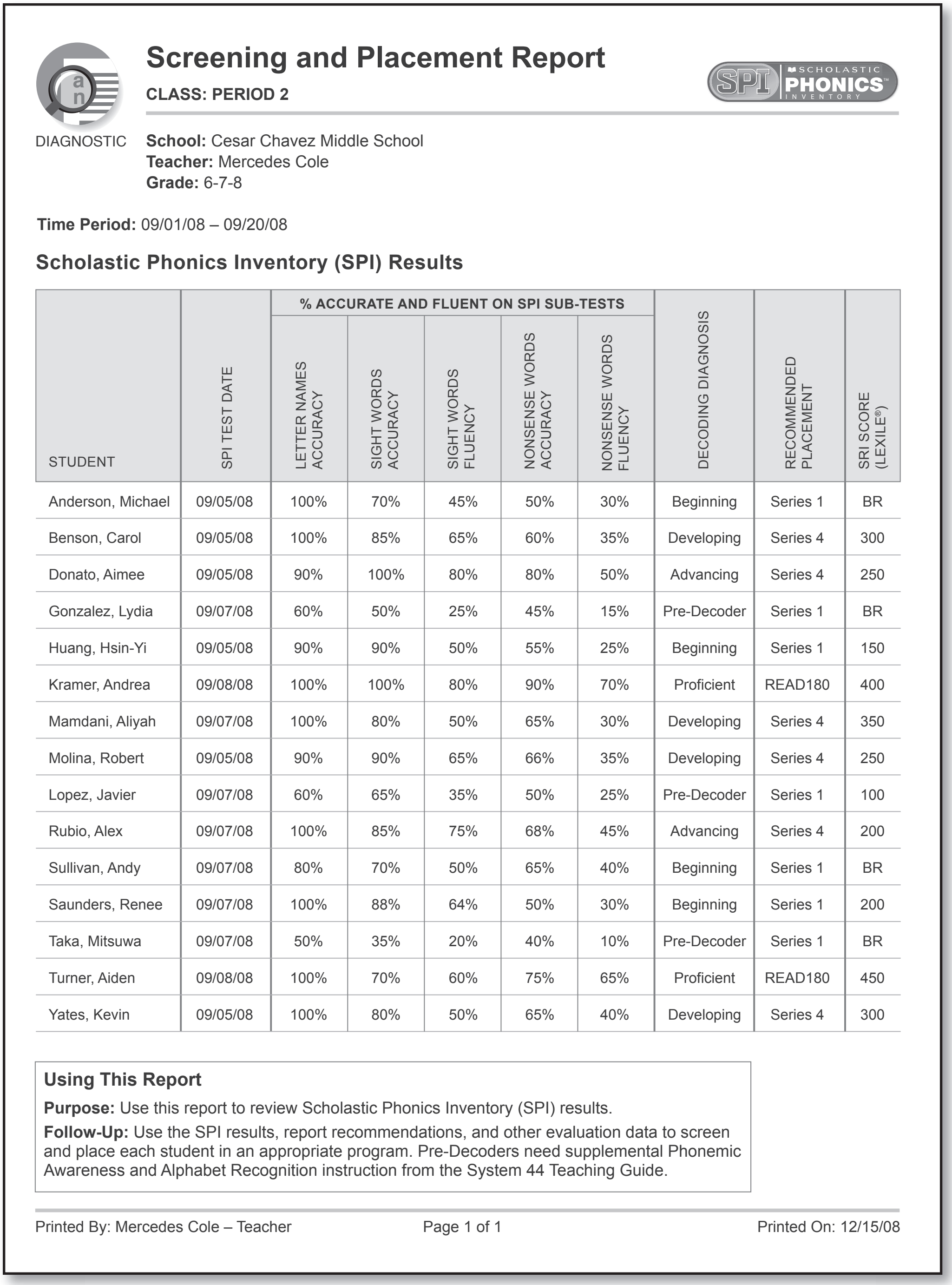
- If a student's total fluency score is between 38 and 67, the student is placed in **Series 4** of the *System 44* software.
  - If the student can fluently decode less than 70 percent of the blends and digraphs in nonwords, then the student is diagnosed as a **Developing Decoder**.
  - If the student can fluently decode at least 70 percent of the blends and digraphs in nonwords, then the student is diagnosed as an **Advancing Decoder**.

- If a student's total fluency score is 67 or higher, the student is placed in the *READ 180* software.
  - In this case the student is diagnosed as a **Proficient Decoder**, and it is not expected that a word-level intervention, such as *System 44*, is necessary. Instead such students will likely benefit from an intervention designed to improve oral reading fluency and comprehension, as is the case with *READ 180*. It is important to note that if taken, the *Scholastic Reading Inventory* (SRI) will provide further information about comprehension performance in reading.

[2] For the identification of letter names, a correct response takes into account accuracy only. Fluency is not measured.

TM & © Scholastic Inc. All rights reserved.

## Figure 1. Screening and Placement Report.



DIAGNOSTIC

### Screening and Placement Report

**CLASS: PERIOD 2**

**School:** Cesar Chavez Middle School
**Teacher:** Mercedes Cole
**Grade:** 6-7-8

**Time Period:** 09/01/08 – 09/20/08

**Scholastic Phonics Inventory (SPI) Results**

| STUDENT | SPI TEST DATE | % ACCURATE AND FLUENT ON SPI SUB-TESTS | | | | | DECODING DIAGNOSIS | RECOMMENDED PLACEMENT | SRI SCORE (LEXILE®) |
|---|---|---|---|---|---|---|---|---|---|
| | | LETTER NAMES ACCURACY | SIGHT WORDS ACCURACY | SIGHT WORDS FLUENCY | NONSENSE WORDS ACCURACY | NONSENSE WORDS FLUENCY | | | |
| Anderson, Michael | 09/05/08 | 100% | 70% | 45% | 50% | 30% | Beginning | Series 1 | BR |
| Benson, Carol | 09/05/08 | 100% | 85% | 65% | 60% | 35% | Developing | Series 4 | 300 |
| Donato, Aimee | 09/05/08 | 90% | 100% | 80% | 80% | 50% | Advancing | Series 4 | 250 |
| Gonzalez, Lydia | 09/07/08 | 60% | 50% | 25% | 45% | 15% | Pre-Decoder | Series 1 | BR |
| Huang, Hsin-Yi | 09/05/08 | 90% | 90% | 50% | 55% | 25% | Beginning | Series 1 | 150 |
| Kramer, Andrea | 09/08/08 | 100% | 100% | 80% | 90% | 70% | Proficient | READ180 | 400 |
| Mamdani, Aliyah | 09/07/08 | 100% | 80% | 50% | 65% | 30% | Developing | Series 4 | 350 |
| Molina, Robert | 09/05/08 | 90% | 90% | 65% | 66% | 35% | Developing | Series 4 | 250 |
| Lopez, Javier | 09/07/08 | 60% | 65% | 35% | 50% | 25% | Pre-Decoder | Series 1 | 100 |
| Rubio, Alex | 09/07/08 | 100% | 85% | 75% | 68% | 45% | Advancing | Series 4 | 200 |
| Sullivan, Andy | 09/07/08 | 80% | 70% | 50% | 65% | 40% | Beginning | Series 1 | BR |
| Saunders, Renee | 09/07/08 | 100% | 88% | 64% | 50% | 30% | Beginning | Series 1 | 200 |
| Taka, Mitsuwa | 09/07/08 | 50% | 35% | 20% | 40% | 10% | Pre-Decoder | Series 1 | BR |
| Turner, Aiden | 09/08/08 | 100% | 70% | 60% | 75% | 65% | Proficient | READ180 | 450 |
| Yates, Kevin | 09/05/08 | 100% | 80% | 50% | 65% | 40% | Developing | Series 4 | 300 |

**Using This Report**

**Purpose:** Use this report to review Scholastic Phonics Inventory (SPI) results.

**Follow-Up:** Use the SPI results, report recommendations, and other evaluation data to screen and place each student in an appropriate program. Pre-Decoders need supplemental Phonemic Awareness and Alphabet Recognition instruction from the System 44 Teaching Guide.

Printed By: Mercedes Cole – Teacher    Page 1 of 1    Printed On: 12/15/08

TM & © Scholastic Inc. All rights reserved.

# DEVELOPMENT OF THE SCHOLASTIC PHONICS INVENTORY

## Development of the SPI Item Bank

*Nonword Items.* The SPI contains 92 nonword items. Each item consists of a target and three distractors. The items were chosen to represent the full range of decoding skills taught in *System 44*, with overrepresentation of the first half of the *System 44* scope and sequence. All targets and distractors are nonwords or obscure English words (e.g., kens) that are unlikely to be known. The targets and distractors were chosen to avoid Spanish words, slang, and nonwords that sounded like real words.

*Sight Word Items.* The SPI contains 37 sight word items. As is the case for nonword items, each sight word item consists of a target and three distractors. The targets were chosen from Fry's 300 Instant Sight Words. The distractors were relatively common words, orthographically similar to the target words.

## Scoring and Cross-Validation Samples

Two primary samples were used in the scoring, reliability, and validity analyses presented in this manual: a Southwestern (SW) sample, which was subdivided into a scoring and a cross-validation sample, and a Southeastern (SE) cross-validation sample.

*Southwestern Sample.* The SW sample consisted of a secondary-school sample of 192 poor readers who were nominated by their teachers as either (a) having sufficient decoding skills to participate successfully in *READ 180* (*N* = 89) or (b) lacking decoding skills necessary to participate in *READ 180* (*N* = 103). From here on forward, these groups are referred to as the "*READ 180* level decoders" and "*System 44* level decoders," respectively. Members of the sample ranged in age from 13 years 11 months to 17 years 7 months, with a median age of 14 years 7 months. The sample contained somewhat more males (54 percent) than females (41 percent), with the gender of the remaining sample (5 percent) unknown. The primary language spoken at home was predominantly Spanish (64 percent) or English (34 percent), with the remaining 2 percent having a variety of other primary languages spoken at home. The sample was largely Hispanic (84 percent), with lesser numbers of non-Hispanic African American (7 percent) and Caucasian (6 percent)

TM & © Scholastic Inc. All rights reserved.

students. The sample contained students who were identified as English Second Language learners (28 percent) and students who were eligible to receive special education services (47 percent).

In addition to the SPI, three decoding subtests were administered to the sample: the Sight Word Efficiency and the Phonetic Decoding Efficiency subtests from the Test of Word Reading Efficiency (TOWRE) (Torgesen, Wagner, & Rashotte, 1999), and the Word Analysis subtest from the Woodcock-Johnson III (Woodcock, McGrew, & Mather, 2001). Six students who were apparently misclassified, based on their test performance, were dropped from the sample. One student nominated as having decoding skills sufficient for *READ 180* had standard scores below 65 on the TOWRE; five students who were nominated as having insufficient decoding skills obtained TOWRE standard scores above the mean of 100. This resulted in a sample size of 186 for the SW sample.

The SW sample was divided into an SW Scoring sample and an SW Cross-Validation sample. A block randomization procedure was used to ensure that half of the *READ 180* level decoders and half of the System 44 level decoders ended up in each of the two samples.

*Southeastern Sample.* The second primary sample used in the validity analyses was an SE sample of 217 fifth-, seventh-, and ninth-grade students who represented a random sample of readers.

## SPI Scoring Algorithm

*Item Level Fluency Thresholds.* Fluency thresholds were determined empirically for each item. The data were provided by the SW scoring sample. Descriptive statistics for the SW scoring sample are presented in Table 1.

TM & © Scholastic Inc. All rights reserved.

### Table 1. Descriptive Statistics of Standard Scores (SS) for SW Scoring Sample on Word-Level Criterion Measures

| ***SYSTEM 44*** **LEVEL DECODERS (*N* = 98)** | | |
|---|---|---|
| **Measure** | **Mean** | **Standard Deviation** |
| TOWRE Sight Word Efficiency (SS) | 75.1 | 12.2 |
| TOWRE Phonetic Decoding Efficiency (SS) | 73.3 | 19.5 |
| Woodcock-Johnson Word Analysis (SS) | 15.7 | 16.9 |
| ***READ 180*** **LEVEL DECODERS (*N* = 88)** | | |
| **Measure** | **Mean** | **Standard Deviation** |
| TOWRE Sight Word Efficiency (SS) | 89.8 | 9.4 |
| TOWRE Phonetic Decoding Efficiency (SS) | 94.8 | 13.4 |
| Woodcock-Johnson Word Analysis (SS) | 40.1 | 21.2 |

These results indicate that both groups were below average in decoding, with the *System 44* level decoders scoring about approximately one standard deviation below the *READ 180* level decoders.

The item fluency thresholds were set so as to differentiate *System 44* and *READ 180* level decoders. For each item, a receiver operating characteristic (ROC) curve was generated. ROC curves are plots of sensitivity versus 1 minus specificity for all potential fluency threshold values. In the present example, sensitivity is the proportion of *System 44* level decoders who are correctly categorized as inadequate decoders by the SPI. Specificity refers to the proportion of *READ 180* level decoders who are correctly categorized by the SPI as adequate decoders.

An example is presented in Table 2 for the purpose of illustrating sensitivity and specificity calculations.

### Table 2. Example of Sensitivity and Specificity Calculations

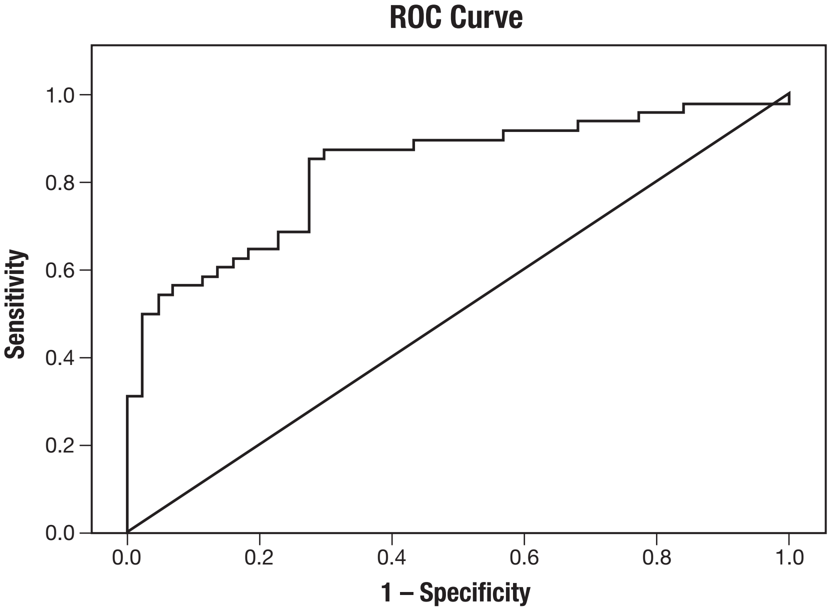
| | **ACTUAL LEVEL OF DECODING** | |
|---|---|---|
| **SPI Performance** | ***System 44*** | ***READ 180*** |
| Inadequate Decoders | 4 | 2 |
| Adequate Decoders | 1 | 8 |

TM & © Scholastic Inc. All rights reserved.

For this example, *sensitivity* (i.e., the proportion of *System 44* level decoders who are correctly categorized by the SPI as inadequate decoders) is 4 (i.e., number of *System 44* level decoders correctly categorized) divided by 5 (i.e., total number of *System 44* level decoders), or .80. *Specificity* (i.e., the proportion of *READ 180* level decoders who are correctly categorized by the SPI as adequate decoders) is 8 (i.e., number of *READ 180* level decoders correctly categorized) divided by 10 (i.e., total number of *READ 180* level decoders), or .80.

An ROC curve for an SPI nonword item is presented in Figure 2.

**Figure 2. Receiver Operating Characteristic (ROC) Curve for an SPI Nonword Item.**



The strategy is to pick the threshold value that represents the point on the curve that is as close to the upper left-hand corner as possible. This maximizes sensitivity and specificity (i.e., minimizes 1 minus specificity). In practice, a table that provides the ROC data in the form of values of sensitivity and specificity for all possible threshold values is used to identify the optimal item fluency threshold. This process was used to identify optimal individual threshold values for each individual sight word and nonword items.

TM & © Scholastic Inc. All rights reserved.

*Combining Accuracy and Latency Into Fluency Scores.* A fluent response must be accurate as well as sufficiently fast. To get credit for a fluent response to an item, the response had to be accurate and the total response time (latency) could not exceed the threshold time. This method of scoring is represented in Table 3.

**Table 3. Combining Accuracy and Latency into Fluency Scores: Four Possible Response Patterns**

| PATTERN | RESPONSE ACCURATE? | LATENCY BELOW THRESHOLD? | FLUENCY SCORE |
|---|---|---|---|
| 1. | No | No | 0 |
| 2. | No | Yes | 0 |
| 3. | Yes | No | 0 |
| 4. | Yes | Yes | 1 |

There are a number of advantages to this kind of scoring. First, this method of scoring produces "hybrid" scores that combine accuracy and speed of responding.  Hybrid scores have proven to be effective on other reading measures such as the TOWRE and the Test of Silent Reading Efficiency and Comprehension (TOSREC) (Wagner, Torgesen, Rashotte, & Pearson, in press). One reason that hybrid scores are effective is that individual and developmental differences in underlying reading skill affect both accuracy and speed of response.

A second advantage of this method of scoring is that outlying response times are handled implicitly. If performance on an assessment is measured in terms of average response time, a practical problem that must be dealt with is what to do about outlying response times. For example, an outlying response time of 20 seconds will have a large impact on the average response time for a set of responses that typically fall in the range of 1 to 2 seconds. The scoring method used on the SPI handles this potential problem in that a response that exceeds the threshold value gets an item fluency score of 0 regardless of how slow the response is.

A third advantage of this method of scoring is that it handles a practical problem that arises in the SPI. Because the mouse must be moved to select the correct response in a list of distractors, the amount of mouse movement

TM & © Scholastic Inc. All rights reserved.

required varies across items depending on the position of the target item in the list of distractors. This presumably affects response times. This potential unwanted source of variability is handled implicitly by the fact that item thresholds are determined empirically for each individual item. Differences in response time associated with differences in amount of mouse movement required are reflected in the empirical distribution of response times that are the basis of the ROC curves used to identify the optimal item threshold.

A final advantage of this method of scoring is that it facilitates maximal use of the information gained from responses to all items, ranging from easy sight word items to difficult nonword items, for the task of differentiating adequate and inadequate decoders. Consider the following example of accuracy and fluency scores obtained for the easy sight word item YOU and the difficult nonword item TABINATE. The mean *accuracy* scores for these two items are presented in Table 4 for the entire SW scoring sample and for the System 44 level decoders and *READ 180* level decoders separately.

**Table 4. Mean Accuracy Scores for SPI Items YOU and TABINATE**

| AVERAGE ITEM DIFFICULTY (ACCURACY ONLY) | | | |
|---|---|---|---|
| **Item** | **Entire Sample** | ***SYSTEM 44* Level** | ***READ 180* Level** |
| YOU | 1.00 | 1.00 | 1.00 |
| TABINATE | 0.54 | 0.42 | 0.68 |

As expected, everyone is perfectly accurate for YOU, as indicated by the item difficulties of 1.00 for the entire sample, for *System 44* level decoders, and for *READ 180* level decoders. This item is not useful for differentiating *System 44* and *READ 180* level decoders if we look at accuracy alone. For the much more difficult TABINATE, only a little more than half of the entire sample gets it correct (0.54), and performance is worse for *System 44* level decoders (0.42) than for *READ 180* level decoders.

TM & © Scholastic Inc. All rights reserved.

Now consider the mean *fluency* scores for these two items, which are presented in Table 5.

**Table 5. Mean Fluency Scores for SPI Items YOU and TABINATE**

| AVERAGE ITEM DIFFICULTY (FLUENCY) | | | |
|---|---|---|---|
| **Item** | **Entire Sample** | ***SYSTEM 44* Level** | ***READ 180* Level** |
| YOU | 0.37 | 0.25 | 0.50 |
| TABINATE | 0.30 | 0.17 | 0.43 |

These results are quite different. The YOU item now is helping out in differentiating *System 44* and *READ 180* level decoders, as indicated by the average difficulties of .25 for *System 44* level decoders and .50 for the *READ 180* level decoders. It helps because to get credit for the item, the student needs to respond accurately and quickly. The TABINATE item works in a similar fashion.

TM & © Scholastic Inc. All rights reserved.

## RELIABILITY OF SCHOLASTIC PHONICS INVENTORY SCORES

Internal consistency reliability coefficients were calculated for Total Fluency, Sight Word Fluency, and Nonword Fluency scores using the data from the SW cross-validation sample. These reliability coefficients, presented in Table 6, support the internal consistency reliability of the SPI fluency scores for secondary-school-age poor readers.

**Table 6. Internal Consistency Reliability Coefficients (Coefficient Alpha) from SW Cross-Validation Sample (*N* = 93) of Secondary-School-Age Poor Readers**

| COEFFICIENT ALPHA | |
|---|---|
| Total Fluency Score | .975 |
| Sight Word Fluency Score | .934 |
| Nonword Fluency Score | .965 |

The SE sample of fifth-, seventh-, and ninth-grade students who represented a random sample of readers was also available to evaluate the reliability of SPI scores. Internal consistency reliability coefficients from this sample are presented by grade in Table 7.

**Table 7. Internal Consistency Reliability Coefficients (Coefficient Alpha) from SE Sample**

| MEASURE | GRADE 5<br>(*N* = 88)[A] | GRADE 7<br>(*N* = 54) | GRADE 9<br>(*N* = 75) |
|---|---|---|---|
| SPI Total Fluency Score | .931 | .973 | .955 |
| SPI Sight Word Fluency Score | .840 | .912 | .886 |
| SPI Nonword Fluency Score | .906 | .964 | .942 |

[A] For SPI Scores, *N* = 58 for Grade 5

TM & © Scholastic Inc. All rights reserved.

In general, the reliabilities mirrored the number of items on the SPI that were included in the score. The SPI Sight Word Fluency subscale contained fewer items than the SPI Nonword Fluency subscale, and the SPI Total Fluency score contained the most items. The reliability of the SPI Total Fluency score was the highest, followed by the SPI Nonword Fluency score and the SPI Sight Word Fluency score.

*Summary of the Reliability Analyses.* The reliability analyses supported the internal consistency reliability of SPI scores. As expected, the SPI Total Fluency score was more reliable than the SPI Sight Word Fluency score or SPI Nonword Fluency score.

TM & © Scholastic Inc. All rights reserved.

# VALIDITY OF SCHOLASTIC PHONICS INVENTORY SCORES

## Content-Description (Content) Validity

Content description validity refers to the examination of the content of the test to determine whether it is a representative sample of the behavior domain that is being assessed (Anastasi & Urbina, 1997). The traditional term for this kind of validity is content validity.

The behavior domain that is assessed by the SPI is fluent decoding of sight words and nonwords. The sight word items on the SPI were sampled from Fry's 300 Instant Sight Words. The nonword items on the SPI were constructed to sample commonly taught phonics skills, which also are the skills addressed in *System 44*, including consonants, short vowels, double consonants, blends, digraphs, and *r*-controlled vowels.

## Criterion-Prediction (Criterion-Related) Validity

Criterion-prediction validity refers to the extent to which a test predicts performance that the test is intended to predict (Anastasi & Urbina, 1997). The traditional term for this kind of validity is criterion-related validity.

Four relevant criteria were available to be predicted in the SW cross-validation sample. These were the Sight Word Efficiency and the Phonetic Decoding Efficiency subtests from the Test of Word Reading Efficiency (TOWRE) (Torgesen, Wagner, & Rashotte, 1999), the Word Analysis subtest from the Woodcock-Johnson III test (Woodcock, McGrew, & Mather, 2001), and the Scholastic Reading Inventory (SRI). Descriptive statistics for these measures are presented in Table 8.

TM & © Scholastic Inc. All rights reserved.

**Table 8. Descriptive Statistics for SW Cross-Validation Sample**

| ***SYSTEM 44* LEVEL DECODERS (*N* = 49)** | | |
|---|---|---|
| **Measure** | **Mean** | **Standard Deviation** |
| **Word-Level Reading Skills** | | |
| TOWRE Sight Word Efficiency (SS) | 73.5 | 14.0 |
| TOWRE Phonetic Decoding Efficiency (SS) | 72.4 | 20.2 |
| Woodcock-Johnson Word Analysis (SS) | 17.7 | 19.5 |
| **Reading Comprehension** | | |
| Scholastic Reading Inventory | 348.1 | 281.0 |
| ***READ 180* LEVEL DECODERS (*N* = 44)** | | |
| **Measure** | **Mean** | **Standard Deviation** |
| **Word-Level Reading Skills** | | |
| TOWRE Sight Word Efficiency (SS) | 89.1 | 9.3 |
| TOWRE Phonetic Decoding Efficiency (SS) | 94.3 | 13.3 |
| Woodcock-Johnson Word Analysis (SS) | 39.9 | 21.4 |
| **Reading Comprehension** | | |
| Scholastic Reading Inventory | 641.1 | 154.2 |

Predictive validity coefficients were calculated by using the SPI fluency scores as predictors of the four criterion variables. The correlations between SPI fluency scores and the four criterion variables serve as validity coefficients, and are presented in Table 9.

**Table 9. Validity Coefficients for Predicting Reading Criterion Scores for the SW Cross-Validation Sample (*N* = 93)**

| **MEASURE** | **SPI TOTAL FLUENCY** | **SPI SIGHT WORD FLUENCY** | **SPI NONWORD FLUENCY** |
|---|---|---|---|
| **Word-Level Reading Skills** | | | |
| TOWRE Sight Word Efficiency | .77 | .74 | .77 |
| TOWRE Phonetic Decoding Efficiency | .68 | .68 | .70 |
| Woodcock-Johnson Word Analysis | .79 | .77 | .77 |
| **Reading Comprehension** | | | |
| Scholastic Reading Inventory | .56 | .52 | .55 |

Note: All coefficients are significant at $p < .001$. Fluency scores were missing for 4 students in the sample.

TM & © Scholastic Inc. All rights reserved.

These validity coefficients were large in magnitude and support the criterion-prediction validity of the SPI scores. As expected, given the fact that the SPI assesses word-level decoding, the coefficients are higher for the word-level reading measures (TOWRE and Woodcock-Johnson) than for the comprehension measure (Scholastic Reading Inventory).

Additional evidence of the criterion-prediction validity of SPI scores comes from the SE sample of 251 students who were sampled from fifth-, seventh-, and ninth-grade classrooms. The students were given the SPI and several reading criterion measures. Descriptive statistics for the SE sample are presented in Table 10.

**Table 10. Descriptive Statistics of Standard Scores (SS) for SE Validation Sample on Word-Level Criterion Measures (*N* = 217)**

| MEASURE | MEAN | STANDARD DEVIATION |
|---|---|---|
| TOWRE Sight Word Efficiency (SS) | 94.1 | 11.9 |
| TOWRE Phonetic Decoding Efficiency (SS) | 92.0 | 17.5 |
| Woodcock-Johnson Word Analysis (SS) | 95.0 | 13.5 |
| Woodcock-Johnson Letter Word Identification (SS) | 95.4 | 14.9 |

The descriptive statistics indicate that the decoding skills of the sample were about a third of a standard deviation below average in general on the word-level reading measures.

Predictive validity coefficients based on the SE sample are presented in Table 11.

**Table 11. Validity Coefficients Predicting Reading Criterion Scores for the SE Sample**

| MEASURE | SPI TOTAL FLUENCY | SPI SIGHT WORD FLUENCY | SPI NONWORD FLUENCY |
|---|---|---|---|
| TOWRE Sight Word Efficiency | .65 | .60 | .64 |
| TOWRE Phonetic Decoding Efficiency | .67 | .53 | .70 |
| Woodcock-Johnson Word Analysis | .62 | .48 | .65 |
| Woodcock-Johnson Letter Word Identification | .55 | .43 | .57 |

Note: All validity coefficients significant at $p < .01$.

TM & © Scholastic Inc. All rights reserved.

The predictive validity coefficients were moderate to large in magnitude, with somewhat larger coefficients for the SPI Total Fluency score and the SPI Nonword Fluency score than for the SPI Sight Word Fluency score. The magnitudes of these validity coefficients are impressive in light of the fact that the scoring system was optimized for differentiating poor readers who had serious decoding problems (i.e., *System 44* level decoders) from poor readers who were adequate in decoding (i.e., *READ 180* level decoders).

### Construct-Identification (Construct) Validity

Construct-identification validity refers to the extent to which a test measures the target theoretical construct or trait (Anastasi & Urbina, 1997). The previous term for this type of validity is construct validity. Construct-identification validity is a global form of validity that encompasses evidence provided about the content-description validity and criterion-prediction validity of a test, but includes other evidence as well. For the SPI, construct-identification validity is supported if groups that are known to differ in levels of decoding can be shown to differ in performance on the SPI.

Using the SW cross-validation sample, the SPI scores of the *System 44* level decoders were compared to those of the *READ 180* level decoders. These results are presented in Table 12.

**Table 12. Descriptive Statistics, Tests of Group Differences, and Effect Sizes for SW Cross-Validation Sample**

| MEASURE | *SYSTEM 44* LEVEL DECODERS (*N* = 49) MEAN (SD) | *READ 180* LEVEL DECODERS (*N* = 49) MEAN (SD) | *t* | COHEN'S *d* |
|---|---|---|---|---|
| TOWRE Sight Word Efficiency (SS) | 73.5 (14.1) | 89.1 (9.3) | 6.22 | 1.31 |
| TOWRE Phonetic Decoding Efficiency (SS) | 72.4 (14.1) | 94.3 (13.3) | 6.11 | 1.60 |
| Woodcock-Johnson Word Analysis (SS) | 17.7 (19.5) | 39.9 (22.4) | 5.14 | 1.06 |
| SPI Sight Word Fluency | 8.4 (6.2) | 21.9 (7.6) | 9.35 | 1.95 |
| SPI Nonword Fluency | 16.4 (11.3) | 43.1 (18.6) | 8.39 | 1.73 |
| SPI Total Fluency | 24.7 (16.3) | 65.0 (24.6) | 9.27 | 1.93 |

Note: All *t*-test values significant at $p < .001$ level.

TM & © Scholastic Inc. All rights reserved.

The groups differed substantially and significantly on all three SPI scores. The magnitude of these differences exceeded those of the differences on TOWRE and Woodcock-Johnson scores, as evidenced by larger effect sizes for the SPI scores compared to those for the TOWRE and Woodcock-Johnson. The magnitude of the group differences in SPI scores supports the construct-identification validity of the SPI scores.

Another way of examining the ability of SPI scores to differentiate *System 44* level decoders and *READ 180* level decoders is to examine validity coefficients in the form of point-biserial correlations between SPI scores and a group membership (i.e., *System 44* level versus *READ 180* level decoders). These results are presented in Table 13.

**Table 13. Validity Coefficients (Point-Biserial Correlation Coefficients) for SPI Scores as Predictors of *System 44* Level Versus *READ 180* Level Decoding Group Membership for the SW Cross-Validation Sample**

| VALIDITY COEFFICIENT | |
|---|---|
| Total Fluency Score | .70 |
| Sight Word Fluency Score | .70 |
| Nonword Fluency Score | .66 |

Note: All validity coefficients significant at $p < .001$.

Because the SPI was constructed to be a measure of word-level reading skills rather than a measure of perceptual motor speed, a second test of construct-identification validity is provided. The average difference in response time between *System 44* level decoders and *READ 180* level decoders for the initial matching items that did not require word-level reading skills and the sight word and nonword items that did require word-level reading skills is compared. The difference in average response times for *System 44* and *READ 180* level decoders was an order of magnitude greater for the items that required word-level reading skills (approximately 500 milliseconds) compared to the matching items that did not require word-level reading skills (approximately 50 milliseconds). This confirms that performance on the SPI is primarily determined by fluency at word-level reading as opposed to simple perceptual motor speed.

TM & © Scholastic Inc. All rights reserved.

*Classification Analyses.* The most stringent test of the construct-identification validity of the SPI is provided by classification analyses. A classification study was carried out in which SPI Total Fluency scores were used to predict group membership (i.e., *System 44* level decoders versus *READ 180* level decoders) for the SW cross-validation sample. For classification studies, four statistics are of importance:

1. *Sensitivity*. Sensitivity refers to the proportion of *System 44* level decoders who are correctly categorized by the SPI.
2. *Specificity*. Specificity refers to the proportion of *READ 180* level decoders who are correctly categorized by the SPI.
3. *Positive Predictive Value*. Positive predictive value refers to the proportion of students the SPI categorized as poor decoders who actually were *System 44* level decoders.
4. *Negative Predictive Value*. Negative predictive value refers to the proportion of students the SPI categorized as adequate decoders who actually were *READ 180* level decoders.

The previous example used to illustrate calculation of sensitivity and specificity (Table 2) is used here again to extend the calculations to positive predictive value and negative predictive value in Table 14.

**Table 14. Example of Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value Calculations**

| | ACTUAL LEVEL OF DECODING | |
|---|---|---|
| **SPI Performance** | **System 44** | ***READ 180*** |
| Inadequate Decoders | 4 (TP) | 2 (FP) |
| Adequate Decoders | 1 (FN) | 8 (TN) |

Note: TP = true positive. TN = true negative. FP = false positive. FN = false negative.

As illustrated previously, sensitivity (i.e., the proportion of *System 44* level decoders who are correctly categorized by the SPI as inadequate decoders) is 4 (true positives, which is the number of *System 44* level decoders correctly categorized as inadequate decoders by the SPI) divided by 5 (true positives and false negatives, which is the number of *System 44* level decoders incorrectly categorized as adequate decoders by the SPI), or .80.

TM & © Scholastic Inc. All rights reserved.

Specificity (i.e., the proportion of *READ 180* level decoders who are correctly categorized by the SPI as adequate decoders) is 8 true negatives, which is the number of *READ 180* level decoders correctly categorized as adequate decoders by the SPI) divided by 10 (true negatives plus false positives, which is the number of *READ 180* level decoders incorrectly categorized as inadequate decoders by the SPI), or .80.

Positive predictive value (i.e., the proportion of students the SPI categorized as inadequate decoders who actually were *System 44* level decoders) is 8 (true positives) divided by 12 (true positives plus false positives, which is the number of *READ 180* level decoders incorrectly categorized as inadequate decoders by the SPI), or .75.

Negative predictive value (i.e., the proportion of students the SPI categorized as adequate decoders who actually were *READ 180* level decoders) is 4 (true negatives) divided by 8 (true negatives plus false negatives, which is *System 44* level decoders incorrectly categorized as adequate by the SPI), or .50.

Different authorities have proposed different standards for what constitutes acceptable values for classification statistics (see Hammill, Wiederholt, & Allen, 2006, for a review). Wood, Flowers, Meyer and Hill (2002) and Jansky (1978) proposed values of .70 as a standard for acceptable values of sensitivity and specificity. Wood et al. (2002) proposed accepting lower values for positive predictive value, whereas Jansky advocated that a standard that required positive predictive value should also achieve a value of .70. Gredler (1997) and Kingslake (1983) proposed that sensitivity, specificity, and positive predictive values should meet a higher standard of achieving values of .75 or better. Hammill et al. (2006) proposed a system of three levels of acceptability for classification statistics that is presented in Table 15.

**Table 15. Levels of Acceptability for Classification Statistics Proposed by Hammill et al. (2006)**

| |
|---|
| Level 1. Sensitivity and Specificity, or Sensitivity and Positive Predictive Value >= .70. |
| Level 2. Sensitivity, Specificity, and Positive Predictive Value >= .70. |
| Level 3. Sensitivity, Specificity, and Positive Predictive Value >= .75. |

TM & © Scholastic Inc. All rights reserved.

The results of the classification analyses for the SW cross-validation sample are presented in Table 16. The results from the analysis of the original complete cross-validation sample are presented in the left column in the table. These values achieve the highest level of acceptability in the Hammill et al. system. Because the SPI was designed to differentiate readers who genuinely had decoding problems sufficient to preclude their successful participation in *READ 180*, a second analysis was carried out after imposing two constraints: The *System 44* level decoders had to score at or below a standard score value of 70 on the TOWRE Phonetic Decoding Efficiency subtest and the *READ 180* level decoders had to score at or above a value of 80 on the same subtest. The results of this analysis are presented in the right column in the table. The results of this second analysis were that the values of specificity and positive predictive value approached their maximum possible value of 1.00.

**Table 16. Classification Statistics for Predicting Decoding Status using SPI Total Fluency Score**

| CLASSIFICATION | ORIGINAL SAMPLE | CORRECTED SAMPLE AFTER DROPPING MISCLASSIFIED CASES |
|---|---|---|
| Sensitivity | .83 | .85 |
| Specificity | .81 | .96 |
| Positive Predictive Value | .83 | .97 |
| Negative Predictive Value | .81 | .81 |

*Summary of the Validity Analyses.* The content-description validity of the SPI was demonstrated by examining the extent to which the items represented the target domains of sight word and nonword decoding. The criterion-prediction validity of the SPI was demonstrated by the magnitudes of the predictive validity coefficients generated when SPI scores were used to predict reading criteria in two studies. The construct-identification validity of the SPI was supported by the magnitude of group differences in SPI scores for *System 44* level decoders and *READ 180* level decoders, and by the success of the SPI in predicting group membership in a classification study. All classification statistics met the highest standard of acceptability. The construct-identification validity also was supported indirectly by the results of the investigation of the content-description validity and criterion-prediction validity of the measure mentioned previously.

TM & © Scholastic Inc. All rights reserved.

# REFERENCES

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Gredler, G. R. (1997). Issues in early childhood screening and assessment. *Psychology in the Schools*, *34*, 99–106.

Hammill, D. D., Wiederholt, J. L., & Allen, E. A. (2006). *Test of Silent Contextual Reading Fluency*. Austin, TX: PRO-Ed.

Hock, M. F., Brasseur, I. F., Deshler, D. D., Mark, C. A., Stribling, J. W., Catts, H. W., & Marquis, J. G. (in press). What is the nature of struggling adolescent readers in urban schools? *Learning Disability Quarterly*.

Jansky, J. J. (1978). A critical review of some developmental and predictor precursors of reading disabilities. In A. Benton & D. Pearl (Eds.), *Dyslexia: An appraisal of current knowledge* (pp. 331–347). New York: Oxford University Press.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, *95*, 719–729.

Kingslake, B. J. (1983). The predictive (in)accuracy of on-entry to school screening procedures when used to anticipate learning difficulties. *British Journal of Special Education*, *10*, 24–26.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.) (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Stanovich, K. E. (1991). Word recognition: Changing perspectives. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Volume 2, pp. 418–452). New York: Longman.

Torgesen, J. K., Wagner, R. K. & Rashotte, C. A. (1999). *Test of Word Reading Efficiency*. Austin, TX: PRO-Ed.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. (in press). *Test of Silent Reading Efficiency and Comprehension*. Austin, TX: PRO-Ed.

Wood, F., Flowers, L., Meyer, M., & Hill, D. (2002, November). *How to evaluate and compare screening tests: Principles of science and good sense*. Paper presented at the annual meeting of the International Dyslexia Association, Atlanta.

Woodcock, R. W., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson Tests of Cognitive Ability (3rd Ed.)*. Itasca, IL: Riverside Publishing.

TM & © Scholastic Inc. All rights reserved.