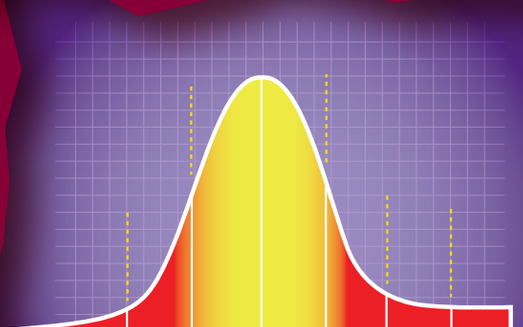
PAR™

Predictive Assessment of Reading

Pre-K to Grade 3 Edition

TECHNICAL MANUAL



Red-e Set Grow, LLC

Black line is trim line - does not print

Description: Par Manual Cover

Joel Shelton
517 Art & Design
joshelton@yahoo.com
336.414.6023

PAR™

Predictive Assessment of Reading

Pre-K to Grade 3 Edition

TECHNICAL MANUAL

Frank Wood, Ph.D.

Wake Forest School of Medicine

Winston-Salem, NC

Address all correspondence to:
Red-e Set Grow, LLC
PO Box 1674
Clemmons, NC 27012

Phone: (336) 777-0075

Printed in the United States of America.

123456789– 12 11 10 09 08 07 06 05 04 03 02 01

ISBN: 978-1-940616-04-9

Contents

Summary

1. The Predictive Assessment of Reading (PAR) was developed at Wake Forest School of Medicine in Winston-Salem, North Carolina, through longitudinal and field research, begun in 1986 and continuing to present. The research has been funded both by government (NIH) sponsored grants and by a private foundation grant from The Dyslexia Foundation.
2. The longitudinal research showed that highly reliable and valid predictions of future reading skill could be produced from four carefully chosen tests given in kindergarten through 3rd grade. These four tests also accurately monitor the progress of remediation over three years, yielding highly valid updated predictions of future improved reading due to remediation.
3. New versions of the four tests were developed and field tested — initially on N=500 K-3 students for cross validation and then N=5,000 students, from second semester Pre-K through 4th grade, for norm confirmation. These were drawn from across the USA and closely matched the racial and ethnic composition of the nation's early elementary school population. This cross validation and replication showed that PAR predicts end of grade tests as well as the concurrent and future Woodcock-Johnson III Broad Reading — with uncommonly high validity across the ability range and with equally high sensitivity and specificity for predicting reading impairment. The four subtests were also found to be highly reliable, making PAR suitable not only for screening, but also for showing profiles of reading-relevant skills and for progress monitoring.
4. PAR meets the standards of the federal Agency for Healthcare Research and Quality (2002) for language tests for children and adults, as shown in the chart below.

In addition to the AHRQ standards noted below, PAR meets other important methodological criteria: (1) Its normative samples are normally distributed, so the correlations are directly comparable to other correlation values on normal samples; (2) scores at each grade level have the same means and standard deviations, permitting appropriate calculations of change scores across time; and (3) validity is equally high before and after remediation — necessary for progress monitoring.

No other screening or progress monitoring test matches PAR's adherence to the standards noted in the chart below.

AHRQ Standard	**PAR Compliance**
Reliability in all forms (internal consistency and test-retest) exceeds .90.	Cronbach's internal consistency alpha >.90 for the three item based subtests. Retest reliability > .90 for all subtests as well as for the overall score.
Validity at least moderate (>.30)	Validity for concurrent WJ-III >=.91.
National norms are ethnically and geographically representative of the population served, and relevant subgroups have N>=100.	National geographically diverse calibration on N=500; norm validation on N=5,000. Validity equally strong in African-American and Hispanic-Latino subgroups of N=100 each. Norms replicate US school ethnicity.
Subtests *a-priori* theoretically predicted relations to each other.	Distinct predictive contributions of vocabulary and rapid naming fluency replicate prior published work.

Development of the Wake Forest School of Medicine Predictive Assessment of Reading (PAR)

Since 1986, research in reading, dyslexia, and related educational and clinical topics has been funded at Wake Forest School of Medicine (WFSM) through NIH grants (including the large program project on dyslexia, entitled "Genotypic and Phenotypic Heterogeneity in Dyslexia"), and through grants from the The Dyslexia Foundation, formerly the National Dyslexia Research Foundation. These have focused on longitudinal studies, in kindergarten and primary grades, on three major cohorts and many additional population sampling studies and intervention studies in schools across the USA and in South Africa. Genetic and neuroimaging studies have been explicitly included in many of the studies. These studies have to date exceeded N=14,000, including three formal randomized prospective intervention studies and numerous field studies.

Initial longitudinal studies for prediction of concurrent and future reading

Initial reports from our longitudinal studies include those by Felton and Wood, 1989; Felton and Wood, 1992; Wood and Felton, 1994; Meyer et. al., 1998a; Meyer et. al., 1998b; and Flowers et. al, 2001. These longitudinal studies assessed a very wide range of linguistic, perceptual, motor, mnestic, executive function, and intelligence skills, in a comprehensive test battery lasting six hours, in K, 1, 3, and 8. Meta-analysis of these longitudinal data showed that four simple constructs, which can be measured through alternative tests, suffice for very strong predictions of concurrent and future reading on the Woodcock-Johnson Broad Reading (Woodcock and Johnson, (1977) or Gates-MacGinitie Reading tests. No other measured skill, nor any demographic variable, accounted for any useful additional explanatory or predictive variance. See Table 1 for the strong multiple regression predictions of concurrent and future reading achievement. The four predictive constructs were:

1. **Phonemic Awareness** as measured either by standard phoneme deletion and same/different judgment tasks (Stanovich et. al., 1982) or by instructed manual manipulation of colored blocks to represent individual sounds (Lindamood Auditory Conceptualization Test);

Table 1.
Standard error, regression R^2 (total variance explained), and multiple R for concurrent and future predictions by the PAR skill battery of the Woodcock-Johnson Broad Reading (WJBR) and the Gates-MacGinite (GM).

		Outcome Criterion and Grade			
Predictor	**Statistic**	**WJBR 1st**	**WJBR 3rd**	**WJBR 8th**	**GM 8th**
First Grade WJBR:	Standard error	0	7.39	7.99	10.08
	Model R^2	1	.74	.68	.55
	Multiple R	1	.86	.82	.74
First Grade PAR:	Standard error	5.08	7.24	7.89	8.67
	Model R^2	.86	.75	.69	.67
	Multiple R	.93	.87	.83	.82
Third Grade PAR**:	Standard error		6.85	7.67	8.91
	Model R^2		.84	.78	.75
	Multiple R		.91	.88	.86

N=220, randomly sampled from a school district population with WJBR mean 102, STD 14, 50% male, 32% ethnic and racial minority. Note**: 3rd grade PAR was N=200, with no difference in sample properties. Two predictor test batteries, using different tests of the same four constructs, showed alternate forms r = .925. Note also: neither race, ethnicity, gender, nor age within grade were close to significant.

2. **Picture vocabulary** either by expressive confrontation picture naming (Boston Naming Test) or by receptive recognition (Peabody Picture Vocabulary);
3. **Rapid Naming** of letters and digits, or of colors and objects (Denckla and Rudel, 1977); and finally
4. **Single word calling** either 5th grade criterion-referenced (Decoding Skills Test, Richardson and DiBenedetto, 1985) or wide-range norm-referenced (WJBR).

The WJBR composite, as an outcome, does contain a single word reading subtest as does the set of predictor variables (but not the same single word reading subtest). When predictor and criterion share a subtest measuring the same domain (here single-word reading), then the high validity is partly due to the expected "autocorrelation" between the two similar subtests. The extreme case would be when the full WJBR predicts itself into the future; the autocorrelation is complete: predictor and criterion are identical. Table 1 includes this extreme case in its comparison, and shows that 1st grade PAR predicts the future 3rd and 8th grade WJBR actually somewhat better than the 1st grade WJBR itself does. By comparison to WJBR in 1st grade, PAR in 1st grade must then be measuring total variance that is at least as well related, likely better related, to the long term WJBR. Even if PAR and WJBR share one area of subtest content (single word reading), the total PAR (3 of whose subtests are not at all similar to WJBR subtests), is functioning as a better predictor than the total WJBR (2 of whose subtests are not all similar to the PAR subtests). Auto-correlation cannot fully explain PAR's high predictive validity.

As a practical matter, single word reading is often included in prediction and screening tests, e.g. Texas Primary Reading Inventory (TPRI) or Dynamic Indicators of Basic Early Literacy (DIBELS), and this typically improves the strength of prediction over what it might be if based only on the underlying skills, excluding single word reading. On the other hand, if all one has from a predictor test is a "short form" of the criterion, with largely similar content, then the basic theoretical question remains pertinent: are PAR's three underlying skill subtests (phonemic awareness, picture vocabulary, and rapid naming) accounting for useful variance, by themselves?

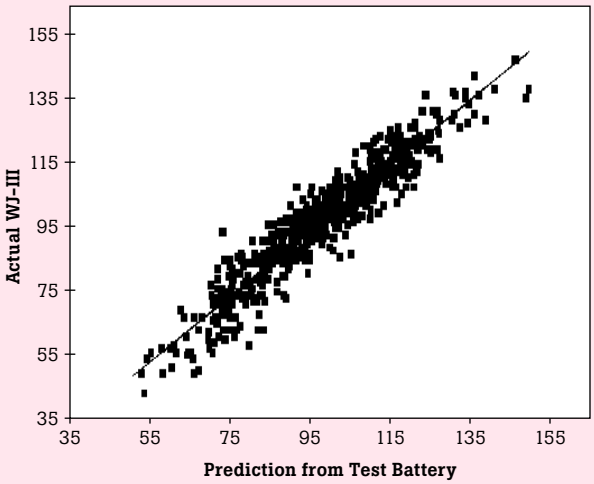
The three underlying skills themselves have high validity: when the set of PAR predictor variables is restricted to these **three underlying skills** of auditory-oral phonemic awareness, auditory-oral picture vocabulary, and visual-oral color-object rapid naming, then these three skills alone still account for **71% and 62% of variance**, respectively in the concurrent and predictive (1st to 3rd grade) predictions of WJBR. Even the 1st to 8th grade prediction from these 3 variables is strong, at 59% of 8th grade WJBR variance, and 65% of 8th grade GM variance.

PAR's three underlying skills are also differentially important at different longitudinal time points:

- From among 1st grade tests, phonemic awareness is the single strongest predictor of concurrent or future reading. The percentages of variance in 1st, 3rd, and 8th grade WJBR outcomes, accounted for by first grade **phonemic awareness** alone, are **54%, 44%, and 40%** respectively; whereas, for 1st grade **picture naming vocabulary** these are **40%, 39%, and 37%** respectively.
- From among 3rd grade tests, picture naming vocabulary is the single strongest predictor of concurrent or future reading. The percentages of variance in concurrent 3rd and future 8th grade WJBR accounted for by **picture naming vocabulary** alone, are **40% and 42%**, respectively; whereas, for third grade **phonemic awareness** alone, these are **31% and 30%,** respectively.
- This appears to confirm the common belief among educators that as reading becomes more content based (as early as 3rd grade), vocabulary (see Scarborough, 1995) accounts for more of its variance than phonemic awareness does. It is interesting to note that this is less an increase in the strength of the vocabulary than a decrease in the strength of phonemic awareness.

Of special note also is the quite small **correlation between picture naming and rapid naming** scores: **.12 in 1st and .05 in 3rd grade,** yet picture naming and rapid naming each have their own strong correlations with WJBR outcomes. In the case of rapid naming, it is partly but not entirely an interaction (Meyer et. al., 1998): within the group of 3rd grade very low readers (below the 10th percentile), the single strongest predictor of WJBR outcome in 8th

Figure 1.
Predicting WJ-III BR PAR: N=500. R=.91 using the regression



Weights from the previous WFUHS longitudinal study. Maximal R=.93.

grade is rapid naming, stronger even than verbal IQ. Taken together with the vocabulary findings above, this suggests that for long term prediction — not to mention prevention and remediation — we cannot focus only on phonemic awareness or phonological decoding; but must also address vocabulary and fluency (Scarborough and Leach, 2000, Wolfe, 1999 respectively). Before looking at intervention studies, though, let us first review the further PAR test development.

Development of the current Predictive Assessment of Reading (PAR).

New tests of the 4 skills were built from items separately field tested to assure that in each skill their overall relation to the criterion was independent of ethnicity or gender. N=500 took the new PAR, from North Carolina, New York, Minnesota, Colorado, Arizona, and California, randomly sampled in demographic strata matching ethnic percentages in early public school grades in the U.S. African American and Hispanic-Latino students each comprised 20% each; majority race 57%; and 3% other. There were at least 100 cases in each grade from K second semester through 3rd.

The validity criterion was the Woodcock-Johnson-III Broad Reading, which replaced the earlier word attack subtest (non-word reading, i.e. phonological decoding), with a text fluency subtest (rapid reading of sentences for true-false judgments). The sample **mean WJ-III BR was 100.1**; **STD of 15.3,** and no departure from normal distribution parameters, in total or by grade level. Scores on the subtests were standardized to the normative grade level group (for Rapid Naming, after a log normal transformation), enabling comparability across grades. The original regression weights gave validity **R=.91, with standard error of 5.80 (in standard score points);** the maximal solution was R=.93. Because of their excellent cross validation, the current PAR uses the original weights. For predicting a 30th percentile WJ-III BR cut score, **sensitivity = 91.7%, specificity = 89.7%** (these are percentage correct predictions of low and high cases, respectively).

See Table 2: Cronbach's internal consistency alpha reliabilities are calculable for the three item based subtests. These were statistically identical across grades; consequently, grade has no significant impact on the results. An N=5,000 ethnically stratified sample was then taken from schools in North Carolina, Florida, New Hampshire, New York, Pennsylvania, Maryland, Minnesota, Colorado, Arizona, and California — 50.5% male, 59% Caucasian, 18% African American, 17% Hispanic-Latino, 3.3% Asian, 1.1% Native American, and 1.6% other or unknown. The earlier norms were strongly confirmed; the N=5,000 mean predicted WJ-III BR was 101.3, with STD of 15.0**. Subsequently, these numbers have risen to over 20,000 children, with highly similar results.**

Table 2.
Internal consistency and parallel forms reliabilities for the four PAR subtests and the total composite PAR prediction.

	Type of Reliability	
Test	**Cronbach's Alpha**	**Parallel Forms**
Picture Naming	.90	.91
Phonemic Awareness	.92	.94
Single Word Calling	.93	.97
Rapid Naming*		.92
Total PAR Prediction		.93

The Cronbach's alpha calculations were based on the full N=500 sample; the parallel forms reliabilities on a later N=50 group.

PAR-referenced original intervention and progress monitoring study

To the original longitudinal study was also added a formal prospective randomized intervention study, throughout first and second grade, involving children who tested at-risk, on a variety of indicators, for low reading at the end of kindergarten (Felton and Brown, 1990). Recently, a separate analysis was done including only those whose risk status was specifically confirmed by the above original four-subtest PAR battery derived from the longitudinal studies. These were a one-year-later cohort, different children from the larger longitudinal study reported above.

The criterion for risk was a kindergarten PAR-predicted WJBR 8th grade outcome below the normative 30th percentile; actual outcome was measured on the WJBR composite. The interventions consisted of special small-group reading circles throughout the 1st and 2nd grade, within the regular classroom, for those children at risk. They were randomized either to an explicit systematic direct code condition (A), a contextually embedded code condition (B), or a passive instruction-as-usual condition (C). The results are presented as follows in Table 3, showing not only the predicted and actual outcomes, but the change in prediction after the two years of treatment (reflecting improvement in the skills that PAR tests), for the two treatment groups and the control group.

The code based instructional paradigm (Group A) did significantly better (indeed 9.1 standard score points better), in its long term WJBR outcomes, compared to the controls, who actually declined over time (itself a common finding).

The major improvement in subtest skills, for Group A, was in phonemic awareness scores, whose scores rose from a mean standard score of 78.4 to a mean standard score of 100.2. By contrast, vocabulary did not rise at all, changing only from 82.7 to 81.8 As a general matter, the fact that vocabulary didn't improve must imply that — if it had — the WJBR outcomes could have been even better. As already concluded previously, effective teaching must address not only phonemic awareness and phonological decoding, but also vocabulary and fluency. The data suggest that PAR can serve as a progress monitoring instrument, a suggestion that was more explicitly tested in the new PAR intervention field tests, reported below.

Table 3.
PAR Predicted and Actual 8th grade outcomes, from K and again from 3rd grade, by intervention conditions during 1st and 2nd grade.

Group, N	PAR prediction at end of K, before treatment	PAR prediction in 3rd Grade, after treatment	Actual WJBR in 8th grade mean, std
A, 21	83.7	90.7	92.8, 10.3
B, 16	84.4	83.8	88.1, 9.1
C, 20	87.7	82.1	83.7, 7.5

Progress monitoring with the new PAR, after field interventions in Florida and North Carolina

Before reporting of the intervention trials, we present a final note on the progress monitoring validity of PAR, demonstrated in the two separate intervention studies, in Sarasota, Florida, and Wake County, North Carolina. In each case, all classes within a specified grade at a single school were tested in late winter or spring of 2005 with the WJ-III BR as well as with the PAR.

> In both cases, the schools had received PAR testing beginning in the second semester of the 2003-2004 school year, and teachers had been given explicit feedback on the PAR results, with training on the use of those results to adapt and guide instruction at the whole class, small group, and individual levels. Retesting was done by PAR in the fall of 2004, and again in the second semester of 2005 when the Woodcock-Johnson III was also given. In both cases, the Woodcock-Johnson was given at least a week after the PAR testing, by examiners who came to the site from outside the county and who were blind to the results of the locally administered PAR test. The question was whether PAR, given repeatedly during intervention, would still accurately predict the Woodcock-Johnson III Broad Reading Composite.
>
> The North Carolina data were from 100 first graders at a single school; the Florida data from 78 second graders at a given school. The North Carolina students had on average shown a five point gain, since the preceding spring, in their PAR-predicted Woodcock-Johnson III Broad Reading Standard Score; those in the Florida school had shown no gain, since the previous spring, in their PAR-predicted WJ-III BR standard score.

Table 4 on the next page, shows the correlations and standard errors for the original N=500 standardization sample and for the two post-intervention schools.

The correlations themselves, while similar, are not really comparable since there is some reduction in variance and restriction in range, especially for the Florida school. The standard errors of prediction are

more directly comparable, and also similar. Given the limited sample sizes in the two schools, it is unlikely that a reliable basis for the possible difference could be inferred from the existing data. At present, though, the evidence suggests that significant improvement in reading skills is accurately tracked by PAR.

Large one-year field trials of PAR-based intervention programs have been completed in Sarasota County, Florida, and in Wake County (Raleigh), North Carolina, as follows.

The Sarasota Project

In May of 2004, after extensive consultation with the administrators of the Sarasota County Public Schools, ***Sarasota YesRead*** was launched with support from The Dyslexia Foundation. It was the first in a series of such projects, collectively titled as ***America YesRead***.

That May, students from kindergarten, first, and second grades from six schools in Sarasota County were tested with the Predictive Assessment of Reading (PAR). PAR is an instrument which, in a nationwide N=500 calibration sample has been validated against the Woodcock-Johnson III Broad Reading score, R=.93 and reliabilities at least .90 for each subtest. Further nationwide norming has been conducted on a total sample of N=5000, with total field tests involving more than 14,000 children.

In July of 2004, two full days of teacher and administrator training were provided to the participating schools, using materials jointly developed by the ***YesRead*** staff and the administrative leadership of the Sarasota Schools. Teachers were briefed on the May PAR test results for children on their upcoming student roster in August, and they were provided with detailed resources for addressing the specific needs each child in their classrooms — at the whole class, small group, and individual levels. Teachers responded with explicit lesson planning; and principals responded with equally explicit techniques for monitoring and supporting the advanced instruction efforts. Reports indicate that teachers and principals alike exhibited high enthusiasm. Further training on comprehension and vocabulary skills was provided for teachers in January of 2005 and in on-site coaching sessions during the year.

Classroom instruction by teachers, guided by the PAR test results along with the supplementary curriculum materials and techniques developed from the training in July, began when school started in early August. In the last week of September and the first two weeks of October, 2004, 1855 of these children were re-tested with PAR, and in early November the principals of the participating schools were provided an initial summary of the results.

Table 4.
Comparison of PAR predictive performance in the norming sample, and in the North Carolina and Florida schools.

	National Norm	North Carolina	Florida
Grade	K through 3	1	2
Sample size	500	78	58
PAR predicted WJ score (mean, std)	100.7 (13.4)	113.5 (13.8)	110.3 (9.1)
Actual WJ-III BR (mean, std)	100.9 (15.3)	112.5 (13.0)	110.9 (10.9)
Correlation: PAR with WJ-III BR	.92	.91	.83
Standard Error of Prediction	5.7	5.1	6.2

The January and May, 2005, rounds of testing were conducted to provide an approximate mid-year and end-year progress monitoring opportunity, and overall results from that testing have been reported to the principals of the schools involved. The Sarasota School system and Wake Forest University have archival copies of the full data from all four testing occasions. The following report summarizes the major finding for the first year of the ***YesRead*** project.

Reduction in Risk of Reading Failure

PAR is able specifically to predict future overall reading on the 3rd grade Florida Comprehensive Achievement Test in Reading (FCAT) at approximately R=.76; thereby, risk of failure was defined for students at each of the four testing times. The major gains were made by first graders, whereas second and third graders made only marginal gains. A particularly strong feature of the ***YesRead*** achievement has been the disproportionately strong gains made by ethnic and racial minority groups. These

gains for Hispanic–Latino and African–American students, in first grade, brought them to a level significantly better than that ordinarily achieved in Sarasota by these groups in first grade. Insofar as the particular ethnic or racial populations are concerned, the results for individual populations are also relatively immune from statistical regression artifact, since all members of the particular ethnic or racial populations were included in the results. Notably, African–American, Hispanic–Latino, and Mixed–Race subgroups made disproportionately strong gains, especially in first grade. The results, in terms of risk reduction are summarized in Table 5.

Table 5.
Percentage of students at risk for FCAT failure across time by grade and selected ethnic subgroups.

	May04	Oct04	Jan05	May05
Overall 1st Grade Percent at Risk:	**21.6%**	**17.4%**	**8.9%**	**8.2%**
Black	60.7%	46.4%	25.0%	21.4%
Hispanic–Latino	50.0%	37.5%	17.5%	15.0%
Mixed	31.0%	27.6%	17.2%	13.8%
White	16.2%	13.7%	6.9%	6.7%
Overall 2nd Grade Percent at Risk:	**8.2%**	**7.2%**	**5.6%**	**7.4%**
Black	40.0%	46.7%	46.7%	33.3%
Hispanic–Latino	19.5%	14.6%	9.8%	14.6%
Mixed	14.7%	20.6%	11.8%	17.6%
White	6.0%	4.8%	3.8%	5.6%
Overall 3rd Grade Percent at Risk:	**11.5%**	**13.3%**	**11.8%**	**11.3%**
Black	16.7%	30.0%	13.3%	20.0%
Hispanic–Latino	28.6%	31.0%	28.6%	28.6%
Mixed	11.4%	11.4%	14.3%	14.3%
White	9.5%	10.7%	10.1%	8.8%
Combined Grades Percent at Risk:	**13.7%**	**12.5%**	**8.8%**	**9.0%**

Note: the findings represent only the N=1803 students who were tested on each of the four test dates. Note also: the 2nd graders had a significantly higher beginning average reading score than either the 1st or 3rd graders. The May 04 average scores for full rising 1st, 2nd, and 3rd grades were 107.9, 114.9, and 111.4, respectively. By May 05, the finishing 1st, 2nd, and 3rd grade averages were 116.1, 111.7, and 108.2 respectively.

A context for Table 5's results is found in beginning and ending average estimated Woodcock–Johnson Broad Reading scores for each grade. The overall changes by grade were as follows.

For 1st graders finishing May 2005: their May 04 average was 107.9; May 05 was 116.1.

For 2nd graders finishing May 2005: their May 04 average was 114.9; May 05 was 111.7.

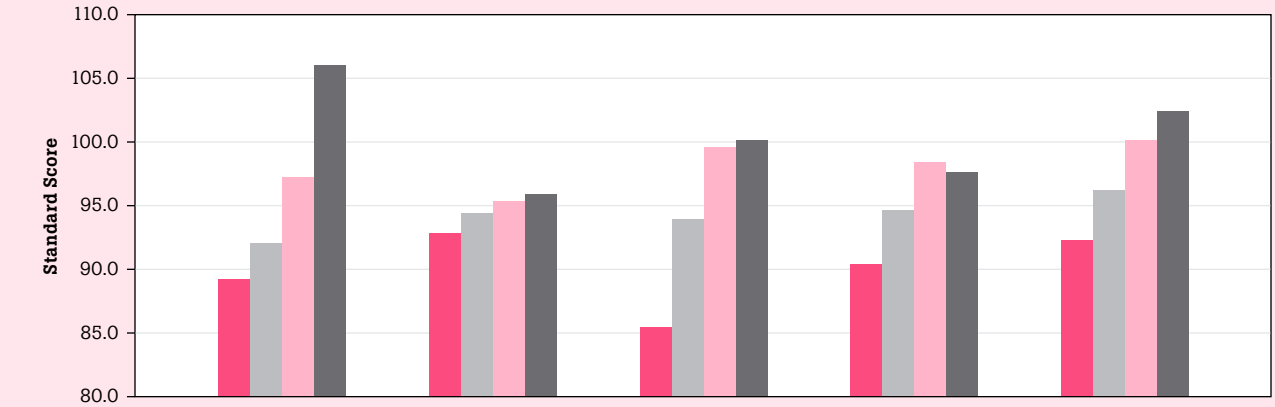
For 3rd graders finishing May 2005: their May 04 average was 111.4; May 05 was 108.2.

Thus, the overall averages of the second and third graders actually showed a marginal decline; whereas the first grade average showed a remarkable gain of almost 10 standard score points. In the context of the PAR's continuing ability to predict the Woodcock–Johnson accurately, it must be concluded that gains shown by the first graders are genuine, representing highly significant change across the year.

Given the above findings, it is appropriate to consider the changes in the 126 first graders who finished kindergarten at risk for future FCAT failure. The changes in their skill profiles are shown in Figure 2 found on the next page.

The "overall" column in the figure represents the composite reading achievement level of the children. It shows a steady gain in actual reading skill — and these gains on PAR were verified in February of 2005 by selective testing in two schools on the Woodcock–Johnson Psychoeducational Battery. That testing showed that there was no reduction whatever in the accuracy, or in slope or intercept, for the prediction by PAR of the Woodcock–Johnson Broad Reading score, even after the PAR had been given three times.

Note especially the vocabulary column: though showing a smaller gain, it still represents a steady and significant achievement, all the more important because the literature generally shows that vocabulary gains, above those by typically reading children, have been particularly difficult to achieve, even in the early grades. (The gains displayed exceed the average gains that typically reading children would make, since normative gains would be represented on the graphs as a horizontal line across time; the



	Word	Vocab	Phon	Fluency	Overall
May 04	89.21	92.80	85.38	90.51	92.34
Oct 04	92.04	94.40	93.92	94.74	96.24
Jan 04	97.28	95.29	99.60	98.47	100.44
May 05	106.00	95.94	100.05	97.74	102.38

Figure 2.
Growth in reading related skills, from May of kindergarten through May of first grade, in N=126 learners at risk for failure on FCAT.

degree of upward slope is the degree of gain in excess of normative gains.) Vocabulary gains are considered particularly important for Hispanic-Latino and African American children, if they are to achieve parity with their majority race peers.

Additional Comments

1. Informal, but extensive, evidence — some of it from within the Sarasota system, but most of it from outside — suggests that training for second and third grade teachers must squarely address the fact that children at risk cannot improve unless they are taught systematically in all relevant skill areas, including phonemic awareness and single word decoding as well as fluency and vocabulary. For some teachers that represents an overt philosophical shift. Achieving that shift almost certainly requires sustained commitments from local and district leadership.
2. An enlargement and further detailing of the Toolkit resource for teachers is appropriate and feasible. It could include pooling of resources from other sites engaged in similar curriculum supplement development. In addition to the Wake County, NC, materials, already available to the Sarasota staff, other materials under development at WFSM are also becoming available.
3. Training of teachers, and the Toolkits themselves, could well focus much more explicitly on actual lesson plan development. Walk-through and other fidelity monitoring approaches are likely to help assure their actual use.
4. The evidence indicates that the entire year of curriculum supplementation, supported by the initial test profiles for students, achieves the best impact. Students transferring into the schools that are using the supplementary materials tend to perform slightly less well than those who have been in that school for longer time periods. New students may need some extra monitoring.
5. Maintaining and extending the already strong gains for ethnic minorities will require not only a continuing effort in phonemic awareness and decoding, but also will need: (1) intentional and explicit vocabulary stimulation; and (2) a sustained repeated readings approach.

The Wake County Project

Randomization by schools within strata

In February of 2004, 14 schools in the Wake County NC School System were randomly divided, within strata of free and reduced lunch percentage, into two groups — control and experimental.

Schedule of testing

Kindergartners in both groups were tested by their teachers using the Predictive Assessment of Reading (PAR), developed at Wake Forest School of Medicine (WFSM) and published by Child's Mind Publishing. Testing was done again for both groups of kindergartners in May, 2004, and again in late April–early May of 2005 when they were in first grade.

Experimental vs. control conditions

Teachers in both groups were trained to administer the test in sessions provided by personnel from WFSM and WCPS. In February of 2004, both groups of teachers also received a brief one-hour training session on general principles of teaching from the PAR test.

Teachers in both groups administered the PAR to their learners. Both groups of teachers received the test results in real time, by online computer scoring, and got a "starter kit" of general principles on how to use the test results to guide instruction.

In addition, the experimental group teachers received an additional full day of teacher training in early March of 2004, for kindergarten teachers, and again in August and September of 2005 for first grade teachers.

Thus, the control group itself received a benefit beyond usual practice — the provision of the PAR test and its results, along with specific suggestions for using these results to guide instruction. The design does not permit a full interpretation of the benefit of this control condition standing alone. Nonetheless, their quantitative gains are apparent in the following results and can be evaluated against conventional expectations for progress in these particular schools. In checklists provided by the central staff, teachers in both conditions, teachers from the control and experimental conditions, provided similar and high levels of endorsement of the usefulness to them of the PAR test results.

Table 6.
Means and standard deviations for the experimental and control groups, on PAR subtests, across three testing occasions.

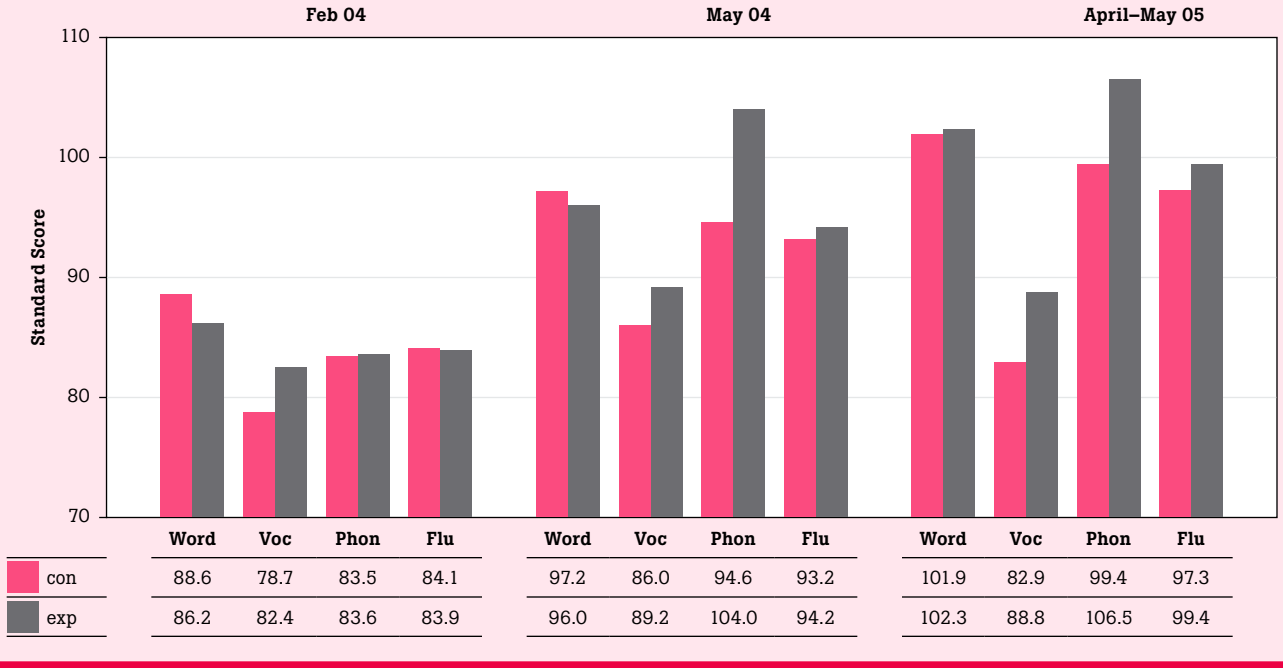
		Feb 04				May 04				Apr-May 05			
		Word	Voc	Phon	Flu	Word	Voc	Phon	Flu	Word	Voc	Phon	Flu
Control	Mean	100.0	100.5	99.5	98.4	113.0	104.0	110.3	105.1	120.8	104.6	108.9	104.9
	STD	16.9	14.7	14.2	13.6	19.7	14.0	12.7	12.3	18.4	16.0	9.5	15.8
Experimental	Mean	99.7	102.7	100.0	97.7	111.0	107.8	113.7	104.2	119.4	107.6	111.8	106.4
	STD	17.7	14.7	16.1	13.9	19.3	15.2	10.6	12.1	18.2	16.3	7.7	15.0

Word=letter-word calling; Voc=vocabulary; Phon=phonemic awareness; Flu=rapid naming fluency.

Table 7.
Predicted third grade end-of-grade (EOG) scores, as means (standard deviations), and percent failure risks, for experimental and control groups across three testing occasions.

Condition	Prediction	Feb04	May04	Apr-May05
Control	3rd EOG	100.0 (11.6)	108.4 (11.8)	116.6 (13.1)
Experimental	3rd EOG	100.6 (12.9)	109.8 (16.8)	117.2 (12.9)*
Control	Risk if Threshold @ 90	18,4%	7.2%	6.4%
Experimental	Risk if Threshold @ 90	19.7%	4.5%*	2.8%**
Control	Risk if Threshold @ 93	25.4%	9.7%	8.5%
Experimental	Risk if Threshold @ 93	28.5%	7.9%	4.5%*

* experimental > control @ p <= .05; ** experimental > control @ p<.01



	Word	Voc	Phon	Flu	Word	Voc	Phon	Flu	Word	Voc	Phon	Flu
con	88.6	78.7	83.5	84.1	97.2	86.0	94.6	93.2	101.9	82.9	99.4	97.3
exp	86.2	82.4	83.6	83.9	96.0	89.2	104.0	94.2	102.3	88.8	106.5	99.4

Figure 3.
Changes in the PAR skill profile for N=157 at risk learners, over 15 months.

Results from the comparison of February and May, 2004, have previously been presented for the 1252 children in both groups who were tested on both occasions. These results showed substantial gains for both groups, with an additional gain for the experimental group in phonemic awareness, picture naming vocabulary, and in the overall predicted reading score.

One school voluntarily discontinued the program in August of 2005. That, together with natural attrition, left 827 children in the study with data at all three time points.

Results

Table 6 (see the previous page), shows the subtest scores for the total experimental and control groups. These scores are standardized to the national PAR norms, with mean 100 and standard deviation 15. A typical sample would therefore be expected to score 100, with standard deviation 15, on each subtest.

By last testing, the experimental group's gain in phonemic awareness exceeded that of the control group at $p<.0001$; in fluency at $p<.05$, and in picture naming at a single tailed $p=.05$.

PAR also produces an overall composite prediction, calibrated either to the individually administered Woodcock-Johnson III Broad Reading or to 3rd grade end-of-grade group-administered achievement tests in reading. Though quite similar, the 3rd grade end-of-grade test score predictions tend to run marginally lower, reflecting the additional impact of attentional and other factors that are less well controlled in a group administered setting. Risks of 3rd grade end-of-grade test failure are calculated and tracked across testing sessions for both groups. Table 7 (see the previous page), shows each group's 3rd grade end-of-grade predictions, and risks of failure at two thresholds.

Changes in the skill profiles of those at risk in the two groups

The risk threshold of 90, yielding near 20 percent at risk, was judged appropriate for this group of schools; their prior end of grade test history is of no more than a 10% risk of failure. A cut score of 90 would ordinarily indicate 25% at risk, yet that same threshold in the present sample identified only about 19% at risk. That appears attributable to a degree of leptokurtosis in the distribution, notwithstanding that the mean and median of the distribution are both 100. Since the February testing occurred at the earliest end of the normative group's testing to which it was compared, the scores for the February testing may also somewhat underestimate the "true" performance at that time. Changes in the skill profile of the risk group are shown in Figure 3 above.

Across the time range, the experimental group's gains in both vocabulary and fluency were statistically greater than the gains of the control group, at $p < .05$.

Changes in the performance of high functioning learners in the two group in Wake County

Both groups showed substantial and statistically equal increases in the number of children scoring above 113, which is the level that would represent an average second grader at that point in the year. The percent scoring at least one grade level ahead were **11.5%, 34.8%, and 48.1%,** respectively for Feb04, May04, and Apr–May05.

(Note: Even if the Feb04 performance of the kindergartners was slightly better than represented by the norms, this caution would not apply to the strong gains made by the first graders between May04 and Apr–May05 because the latter two are referenced to norms for those specific months. Those gains are not explainable by norming artifact.)

Comments on the Wake County Study

The gains shown by the learners in this study are of two kinds: (1) those due specifically to the experimental treatment, which delivered substantially more teacher training than was delivered in the control condition; and (2) those attributable to the PAR testing itself — these latter gains were shared equally by the control and experimental groups.

1. The teacher training has been unquestionably helpful. As is familiar, phonemic awareness was the first and strongest gain (by the experimental group compared to the control group). The vocabulary gain, however, while of smaller magnitude, is in some ways the most important, since vocabulary gains have in the literature always been more difficult to achieve: that they were achieved at all is a strong tribute to the teacher training program. Consistent with an emphasis in the program, the experimental advantage over controls was most pronounced in the progress made by learners at risk, not so much by high functioning learners.
2. There appears to be a major effect of the PAR testing itself, including its feedback to teachers, even in the control group. Some gains are no doubt due to the fact that Wake's kindergartners receive more and better reading readiness instruction than the national average of kindergartners, but a part of the explanation likely includes the fact that the information about skill profiles from PAR, together with the starter kit guidance, was somewhat helpful in its own right. Stated another way, we could not expect gains of the magnitude shown, between February and May of the kindergarten year, in untreated Wake County schools, on average.
3. Individual school summaries are provided below. Their overall interpretation might well take into account the starting points in the February testing; many with the same starting predicted EOG diverged considerably in the outcome. Of note also are the differences between gains in risk (a specific emphasis of the PAR and the Tool Kit supplementary program developed by Wake County) and other gains. As a general rule, an ending risk of less than 5% represents a major achievement that is near the average limit expected in the literature. Five of six experimental schools achieved that; only two of seven control schools did.
4. Because these are real gains, they can be expected to fade unless actively maintained in second and third grade. Evidence elsewhere confirms that such gains will not automatically persist; they must be actively maintained and reinforced in the second and third grades.

Individual School Results

The results shown on the next page, are grouped by schools for the 2004–2005 school year, again reporting only those who were tested on all three dates: February 2004, May 2004, and April–May 2005.

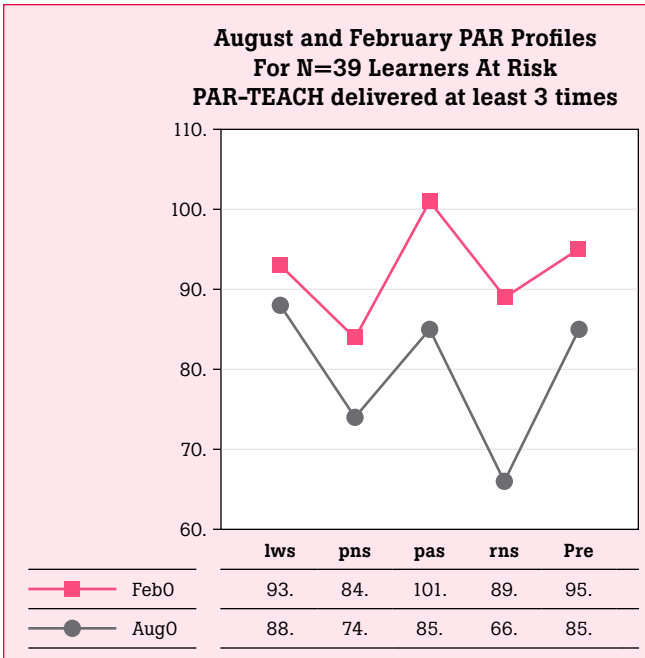
Table 8.
Predicted 3rd grade end-of-grade (EOG), failure risk percent, and percent high functioning learners, by schools, across testing occasions.

Schools		Predicted EOG			Risk of EOG Failure			Percent High Functioning		
		Feb04	May04	Apr05	Feb04	May04	Apr05	Feb04	May04	Apr05
Control:	A	103.9	111.8	114.5	8.5%	2.8%	0.9%	15.1%	44.3%	56.6%
	B	100.6	107.3	109.2	13.8%	10.3%	10.3%	10.3%	31.0%	43.1%
	C	92.5	98.8	102.2	43.8%	16.7%	14.6%	2.1%	6.3%	16.7%
	D	99.7	110.0	110.0	16.4%	6.6%	6.6%	11.5%	41.0%	47.5%
	E	106.8	117.1	123.7	5.9%	0.0%	0.0%	23.5%	58.8%	82.4%
	F	100.4	108.8	110.7	21.5%	9.2%	7.7%	9.2%	36.9%	55.4%
	G	98.0	107.6	109.5	20.5%	6.0%	6.0%	6.0%	26.5%	41.9%
Experimental:	T	100.2	109.1	111.5	23.2%	7.3%	4.9%	17.1%	45.1%	52.4%
	V	92.2	102.7	107.2	41.7%	4.2%	0.0%	0.0%	14.6%	27.1%
	W	97.4	107.1	105.1	23.8%	14.3%	9.5%	7.1%	38.1%	31.0%
	X	106.0	114.6	118.2	10.7%	0.0%	0.0%	25.0%	46.4%	71.4%
	Y	101.1	109.0	110.2	15.8%	2.6%	2.6%	7.9%	28.9%	42.1%
	Z	103.5	112.8	115.2	10.1%	1.1%	1.1%	15.7%	37.1%	58.4%

Note: High functioning are those whose estimated reading skills are equivalent to those of a learner who is one full grade level ahead, so in April of 2005, in the first grade, the percent high functioning is the percent whose scores are those of an average ending second grade learner.

Note also: Control school G and experimental school Y both implemented the Wilson Fundations program in the regular classroom. (Fundations is a briefer version of the classical Wilson Reading System remedial program, addressed to the main stream classroom and taking 30 to 40 minutes a day of programmed whole-class instruction from the teachers.) Since Fundations does not explicitly address the needs of low functioning learners, PAR's guidance for small group work on specific skill challenges was still used by school Y. Experimental school Y did achieve a distinctly low level of risk; all of the risk reduction was achieved in the kindergarten second semester, however.

Note especially that monitoring of the control schools' fidelity to instruction, coordinated by central office staff in the literacy department of the Wake County Schools, generated an independent prediction that school V achieved significantly higher fidelity, which was rated as excellent; compared to the other five experimental schools whose fidelity was rated adequate.



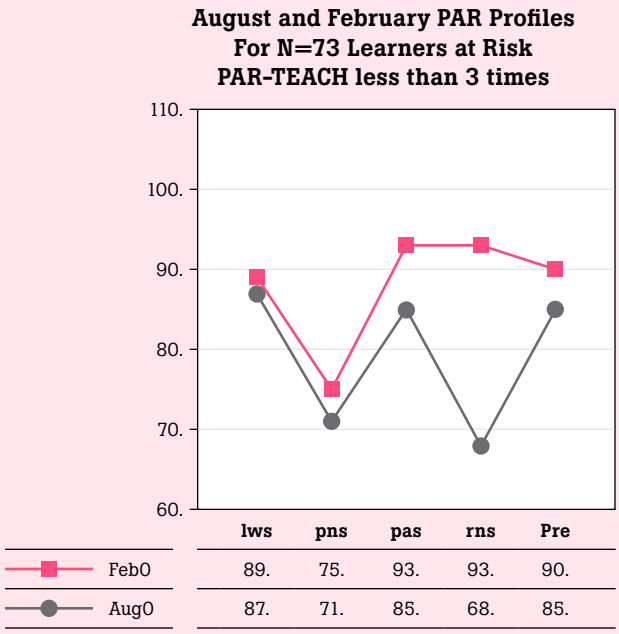


Figure 4.
Gains from August to September in PAR subtest and overall performance, as a function of the frequency of PAR-TEACH supplementary instruction. Note: lwss, pnss, pass, and rnss are the respective standard scores for letter-word calling, picture naming vocabulary, phonemic awareness, and rapid naming; Pre is the overall predicted reading standard score.

The Winston-Salem Forsyth County Study

In the 2005–2006 school year, Winston-Salem Forsyth County Schools invited WFSM to provide PAR-TEST and PAR-TEACH in several volunteering schools. PAR-TEST was the standard PAR, but was supplemented for some ELL students with the PAR Spanish Supplement. PAR-TEACH was a fully scripted set of lesson plans for the year, involving Teacher Talk Time, Text Time, and Small Group phonological decoding instruction for those needing it.

Guided Reading was in general use, and the new specific phonological package from Fountas and Pinel was also used. Both Fountas and Pinel and PAR-TEACH were offered to teachers for use at their discretion, so there was variable frequency in use of either curriculum supplements.

The results in Figure 4 show quite encouraging gains for the children at risk (with predicted EOG testing. At $p<,005$, the level of PAR made an independent contribution to the overall reading score, regardless of the level of Fountas and Pinel instruction. The level of Fountas and Pinel instruction makes a contribution at $p<.05$, independent of the level of PAR instruction.

The above gains are highly typical of the other field trials, showing even more vocabulary and fluency gain. For overall reading (Pred), there was about a 5 standard score point advantage when PAR-TEACH was done at least three times a week (we recommend four times). These gains are typical of our other field trials, but these show even more vocabulary and fluency gain. For overall reading (Pred), there was about a 5 standard score point advantage when PAR-TEACH was done at least three times a week (we recommend four times).

Ongoing Studies

Additional large-sample evidence, now exceeding N=20,000 is available from current PAR testing in a very large city in Pennsylvania, a small semi-rural county in North Carolina, and individual public and private schools in scattered locations across the US. This evidence continues to show not only the same reliability and validity results as before, but also the same polled teacher satisfaction.

Comparisons between PAR and Other Nationally Used Tests

Before comparing PAR to other tests, it may be useful to review some consensual definitions.

Reliability. (1) Test-Retest: the extent to which a test yields the same score if, in a short time span, a group is re-tested or given an alternate form. The Pearson correlation coefficient between the two administrations is the index. (2) Internal consistency: the degree to which the items measure the same thing, often indexed by Cronbach's Alpha.

Validity. The degree to which a test predicts another test, either concurrently, i.e. in a similar time frame, or predictively, i.e. for a test that is given at a significantly future time. The square of the Pearson correlation between the two tests is the consensual index.

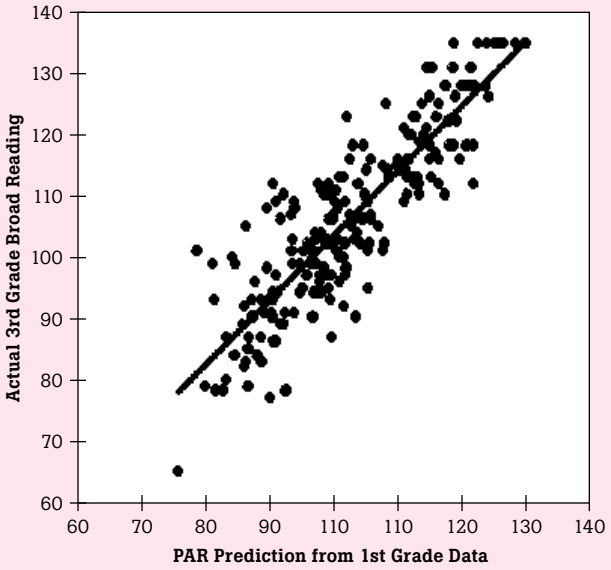
Scatter Plot. A two-dimensional graph with a horizontal (x) and a vertical (y) axis. In the space bounded by these two axes, a single point is plotted for each child, representing that child's score on the Screening Test (the x axis) and the Criterion Test (y axis). Such a plot displays the **correlation** between the two tests, which is **"r,"** the Pearson Product-Moment Correlation. Consensually the strength of the prediction is represented as the square of the correlation, hence **"r2 "** — the square of the correlation, which describes the actual proportion of the Criterion variance that is predictable from the Screen. Correlations can be considered moderately strong when the **r2** is .50 or higher, so that at least half the variance in the criterion is predictable from the screen. An r2 value of .50 requires an r (correlation) of .71. Figure 5 is a scatter plot of data from our laboratory (Wood et al, 2005).

Cut Score for Prediction. The one score on a Screening Test below which a child is predicted to fail a subsequent Criterion Test. A child scoring at or above the Cut Score for is predicted to pass the criterion test. A Cut Score can be displayed as a vertical line on the scatter plot.

Pass Fail Line or Score. The one score on a Criterion Test below which a child is judged to have failed — i.e., not to have reached the required achievement level. A pass/fail line can be displayed as a horizontal line on the scatter plot.

Contingency Table. The consensual method of summarizing a scatter plot in relation to the Cut Score for Prediction and the Criterion Pass/Fail Line. It reports the number of cases, or percentages in the total sample, in each of the four quadrants defined by vertical and horizontal lines on the scatter plot, respectively representing the Cut Score of Prediction

Figure 5.
Prediction of 3rd grade Woodcock-Johnson Broad Reading from 1st Grade PAR. N = 220, r = .87, r2 = .76. From longitudinal sample described in Wood et al, 2005.



Note: in conjunction with the discussion on Contingency Tables: the horizontal dashed line is a Pass/Fail line on the Criterion, here approximating the 20th percentile of the sample. The vertical line is the optimal Cut Score of Prediction from the PAR test. The diagonal line is the linear regression best fit.

Comment: The prediction is uncommonly strong and the Table 9 accuracy indices are the highest yet seen for such a long (2 year) interval between prediction and outcome.

Table 9.
1st grade PAR-predicted pass/fail vs. 3rd Grade actual WJ Broad Reading pass/fail

	Predicted To Fail	Predicted To Pass	Total
Actual Pass:	6.8%	73.6%	80.5%
Actual Fail:	16.4%	3.2%	19.5%
Total:	23.2%	76.8%	100.0%
Accuracy Indicators			
False Positive %: 29.4%		Sensitivity: 83.7%	
False Negative %: 4.2%		Specificity: 91.5%	
Overall Accuracy: 90.0%			

and the Pass/Fail line on the Criterion Test. The cases in each quadrant are then defined as follows.

True Positive (TP). Lower left. "Positive" means predicted to fail. So TP means Screening score less than (left of) the Cut Score *AND* Criterion score less than (below) the Pass/Fail line.

True Negative (TN). Upper right. "Negative" means predicted to pass. So **True Negative** % means Screening score greater than or equal to (right of) the Cut Score *AND* Criterion score greater than or equal to (above) Pass/Fail line.

False Positive (FP). Upper left. Screening score less than the Cut Score, *BUT* Criterion score above the Pass/Fail line. **False Positive** % is the proportion of predicted failures that actually passed the Criterion, hence FP/(FP+TP). ***Note Well: False positive is perhaps the most potentially misleading of the various accuracy definitions. The above definition is the one we use, by consensus (see Torgeson, 1998; Catts et al, 2001; Wood et al, 2005). Others, notably the authors of the Texas Preschool and Primary Reading Inventory (TPRI — see below) have defined false positive as 1-sensitivity (see below), a value which is usually rather lower than the value obtained from the definition we use and has the disadvantage of being completely determined by the sensitivity value and therefore adding no additional information. Before tests are compared on false positive values, therefore, it is particularly necessary to be sure the same definitions are being used.***

False Negative (FN). Lower right. Screening score greater than or equal to Cut Score, *BUT* Criterion Score below Pass/Fail line. False Negative % is the proportion of predicted passes that actually failed the criterion: FN/(FN+TN).

Sensitivity. Proportion of actual failures that were correctly predicted to fail: TP/(TP+FN).

Specificity. Proportion of actual passes correctly predicted to pass: TN/(TN+FP)

Overall Accuracy. Proportion of all cases that were correctly predicted: (TN+TP)/total cases.

DIBELS. The acronym for the most widely used predictive screening test in the USA, "Dynamic Indicators of Basic Early Learning Skills." DIBELS contains several brief tests, different ones of which are used across kindergarten through third grade. It was developed at the University of Oregon. The website containing its technical information, including reliability and validity studies, is: https://dibels.uoregon.edu/techreports/index.php.

TPRI. The acronym for another common predictive screening test: "Texas Primary Reading Inventory." TPRI contains several subtests, differentially used across kindergarten through third grade. It was developed at the University of Texas–Houston and the University of Houston. Foorman, B. R., Fletcher, J. M., & Francis, **D.** J. (1998). *Texas primary reading inventory.* Texas Education Agency and University of Texas System. www.tpri.org

The adequacy of these screening tests is ultimately measured by their reliabilities and their validity: (1) by reference to concurrent or predicted future individually administered tests of reading achievement, such as the Woodcock–Johnson Psycho–educational Broad Reading (WJBR) score, and (2) by reference to High Stakes Testing, typically at the end of third grade. The former provides the user with assurance that the screening test can effectively predict a test, like the WJBR, that is already taken as a standard measurement of reading skill. The latter provides assurance that the outcome of a High Stakes Test can also be predicted in advance. The r2 provides an index suitable for predictions, concurrent or future, of standard achievement such as WJBR, where there is often interest in accuracy across the whole range. For High Stakes Tests, however, pass/fail accuracies are required — since the relevant outcome is whether

the child passed. In that case, accuracy is summarized by the conventional indices: sensitivity, specificity, false positive and false negative percentage, or their derivatives.

Standards for evaluating test performance. By consensus, high standards apply to tests that will be used to make decisions about individual children, as in screening. Norms should be based on at least N=100 for each relevant group, with evidence of the national representativeness of the sample. See Agency for Healthcare Research and Quality, January, 2002. *Criteria for Determining Disability in Speech-Language Disorders*. Summary, Evidence Report/ Technology Assessment: Number 52. AHRQ Publication No. 02-E009, January 2002. Rockville, MD. http://www.ahrq.gov/clinic/epcsums/spdissum.htm We interpret the requirement for norms to be at least N=100 for each grade level to which the screening test is applicable.

Test-retest and internal consistency reliabilities (Cronbach's Alpha) at least r=.90. See again Agency for Healthcare Research and Quality, January, 2002. *Criteria for Determining Disability in Speech-Language Disorders*. Summary, Evidence Report/Technology Assessment: Number 52. AHRQ Publication No. 02-E009, January 2002. Rockville, MD.

Sensitivities and specificities should both be at least 70%. See Committee on Children with Disabilities, American Academy of Pediatrics. (2001). Developmental surveillance and screening of infants & young children. *Pediatrics,* 108, 192-5.

False negative and false positive rates can properly vary with the particular needs of a school. As pointed out by Torgeson (1998) and Wood et al, (2005), educators can tolerate higher false positive than false negative percentages. To the extent of the overall r2, moreover, outcomes for a considerable number of false positive cases will still be below average. Such cases will still benefit from being identified and given additional help. Figure 5 illustrates that: about half of the cases in the upper left quadrant, i.e. the false positives, are still falling at or below the average (100). False negatives are also less serious if repeated wholesale screening is done, but approximately 10% or less is a commonly used goal for a false negative percentage.

Comparisons across tests

Tables 10 and 11 show reported reliability and validity/accuracy for PAR, DIBELS, and TPRI. Figure 6 shows ratings of PAR by the Technical Review Committee of the National Center on Response to Intervention. PAR was the only screening test with top rating on all five standards: classification accuracy, generalizability, reliability, validity, and disaggregated data for diverse populations. See http://www.rti4success.org/chart/screeningTools/screeningtoolschart.html.

Figure 6.
Ratings of PAR by the Technical Review Committee of the National Center on Response to Intervention.

Tools	Area	Classroom Accuracy	Generalizability	Reliability	Validity	Disaggregated Reliability, Validity, and Classification Data for Diverse populations	Efficiency			
							Administration Format	Administration and Scoring time	Scoring key	Benchmarks/Norms
Predictive Assessment of Reading	*Reading	●	Broad	●	●	●	Individual	16 minutes	No	Yes

● = Convincing Evidence (maximum rating)
Broad = (highest rating)

Table 10.
Reliability estimates for PAR, DIBELS, and TPRI.

TEST/ Grade level	r	Citations and Notes
PAR		**Wood et al, 2005**
Reliability Type is Alternate Forms Grade 1:	.90–.94	(Study 1, longitudinal: p 201–208). Grade 1 testing across 2nd semester and subsequent summer vacation (p. 201). The two forms had different methods of testing each of the four domains (picture vocabulary, phonemic awareness, rapid naming fluency, and single letter-word reading (p. 207). Each form yielded a predictions for each of the 4 Criterion outcomes: 1st, 3rd, or 8th Grade WJBR, or 8th Grade GM, and these predictions had somewhat different weightings for the four subtests of PAR . (p. 208). Each of those four Alternate Form correlations was >= .90, with median .925, (p. 207).
Reliability Type is Cronbach's Alpha Grades K,1,2,3:	.90–.93	(Study 2, nationwide one-time, stratified by grades, p. 208–211). Cronbach's alpha coefficients for each of the three item-based subtests, independent of grade level, were: picture vocabulary =.90; phonemic awareness =.92; letter-word reading =.93 (p. 208). See Mehrens and Lehman (1987) for alpha coefficient use in education. Note: Study 2 used the currently published forms of the PAR test.
DIBELS		**Good et al, 2002** ***Reliability Type is Alternate Forms*** (Note: DIBELS uses different screening tests at different grades.)
Midyear K:	.72	(p. 6) Initial Sound Fluency
Mid K, mid 1:	.88	(p. 7) Phoneme Segmentation Fluency
Mid Grade 1:	.83	(p. 8) Nonsense Word Fluency
Mid K, Mid 1:	.88	(p. 8) Letter Naming Fluency
Grade 1 to 3:	.94	(p. 9) Oral Reading Fluency (text passages) — replaced earlier form

TEST/ Grade level	r	Citations and Notes
TPRI Screen		**Foorman et al., 1998** ***Reliability Type is Test-Retest***
End K:	.66–.95	(p. 43) Letter-name =.95; Letter-sound=.87; Blending=.67
Begin Grade 1:	.70–.87	(p. 50) Letter-name =.80; Letter-sound=.76; Word Reading= .87; Blending =.70
End Grade 1:	.95, .58	(p. 59) Respectively Word Reading, Blending Phonemes.
Begin Grade 2:	.90	(p. 64) Word Reading
TPRI Screen		**Center for Academic and Reading Skills, 1999** ***Reliability Type is Cronbach's Alpha***
Mid & End K:	88–.91	(p. 16) Letter-Sound (mid & end K), Blending Onset-Phonemes (mid K); Blending Onset –Rhymes and Phonemes (end K).
Grade 1:	77–.88	(p. 19) Letter-sound (only begin year). Word Reading Task and Blending Phonemes (begin and end year).
Begin Grade 2:	.85	(p.22) Word Reading Task.

Table 11.
Predictive validity of PAR, DIBELS, and TPRI, for future 3rd grade Individually Administered Reading Achievement Tests (WJBR)

	PAR		DIBELS	TPRI
Reference:	**Wood et al, 2005**		**Assessment Committee (2001)**	**Foorman et al, 1998**
Time Span of prediction: Criterion Test:	Grade 1 to 3 WJBR		Grade 1 to End 2 WJBR	Grade 1 to 2 WJBR
Sample:	North Carolina normally distributed		Statewide Oregon, plausibly assumed to be normal	Texas regular education classes; plausibly normal
Predictive Strength, as r2:	**.76** (p. 206)		**.45** NWF is the predictor (p. 8: the r is squared to r2)	Not Reported
Criterion Pass/Fail Line:	15th %ile	30th %ile	Not Reported	20th %ile
Sensitivity:	81.4%	80.5%		92.2%
Specificity:	81.4%	82.0%		76.6%
False Positive:	48.5%	25.5%		49.6%
False Negative:	5.3%	13.5%		2.6%
Overall Accuracy:	81.4%	81.4%		79.8%

References

Agency for Healthcare Research and Quality, January, 2002… *Criteria for Determining Disability in Speech–Language Disorders.* Summary, Evidence Report/ Technology Assessment: Number 52. AHRQ Publication No. 02–E009, January 2002. Rockville, MD. http://www.ahrq.gov/clinic/epcsums/spdissum.htm

Catts, H. W., Fey, M. **E.,** Zhang, S., & Tomblin, J. B. (2001). Estimating risk for future

reading difficulties in kindergarten children: A research based model and its clinical

implications. *Language, Speech, and Hearing Services in Schools,* 31, 38–50.

Denckla, M. & Rudel, R. G. (1976). Rapid automatized naming (RAN): Dyslexia differentiated from other learning disabilities. *Neuropsychologia, 14,* 471–479.

Dunn, Ll. M, & Dunn, L. M. (1997). Peabody Picture Vocabulary Scale — Third Edition (PPVT–III). Circle Pines, MN: American Guidance Service.

Felton, R.H., & Brown, I.S. (1990). Phonological processes as predictors of specific reading skills in children at risk for reading failure. Reading and Writing: An Interdisciplinary Journal, 2, 39–59.

Felton, R. H. & Wood, F. B. (1989). Cognitive deficits in reading disability and attention deficit disorder. *Journal of Learning Disabilities, 22,* 3–13.

Felton, R. H. & Wood, F. (1992). A reading level match study of nonword reading skills in poor readers with varying IQ. *Journal of Learning Disabilities, 25,* 318–326.

Flowers, D. L., Meyer, M. S., Lovato, J., Felton, R. H., & Wood, F. B. (2001). Does third grade discrepancy status predict the course of reading development? *Annals of Dyslexia, 51,* 49–71.

Foorman, B. R., Fletcher, J. M., & Francis, D. J. (1998). Texas Primary Reading Inventory. Texas Education Agency and University of Texas System.

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test.* Baltimore, MD: Lippincott, Williams & Wilkins.

Lindamood, C. H. & Lindamood, P. C. (1979). *Lindamood Auditory Conceptualization Test.* Boston, MA: Teaching Resources Corporation.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates–MacGinitie Reading Tests 4th Edition.* (4th ed.) Itasca, IL: Riverside Publishing Co.

Meyer, M. S., Wood, F. B., Hart, L. A., & Felton, R. H. (1998). Selective predictive value of rapid automatized naming within poor readers. *Journal of Learning Disabilities, 31,* 106–117.

Meyer, M. S., Wood, F. B., Hart, L. A., & Felton, R. H. (1998). Longitudinal course of rapid naming in disabled and nondisabled readers. *Annals of Dyslexia, 48,* 91–114.

Richardson, E. & DiBenedetto, B. (1985). *Decoding Skills Test.* Parkton, MD: York Press.

Scarborough, H. S. & Leach, J. M. (2000). *Late–emerging reading disabilities: New research on the "fourth grade slump."* Washington, D.C.: International Dyslexia Association.

Stanovich, –. K.–E. (1982). "Individual differences in the cognitive processes of reading: II. Text–level processes." Journal of Learning Disabilities 15(9): 549–554.

Torgesen, J. K. (1998). Catch them before they fall. Identification and assessment to prevent reading failure in young children. *American Educator,* 22(1, 2), 32–39.

University of Oregon. (2000) *Dynamic Indicators of Basic Early Literacy (DIBELS).* Internet. http://dibels.uoregon.edu

Wolf, M., & Bowers, P. (1999). The "Double–Deficit Hypothesis" for the developmental dyslexias. Journal of Educational Psychology, 91(3), 1–24.

Wood, F.B., & Felton, R.H. (1994). Separate linguistic and attentional factors in the development of reading. Topics in Language Disorders, 14(4), 42–57.

Woodcock, R. W. & Johnson M.G. (1977). *Woodcock Johnson Psycho–Educational Battery.* Allen, TX: DLM.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson Tests of Achievement–III.* (3rd ed.) Itasca, IL: Riverside Publishing Co.

Red-e Set Grow consists of a team of leading educators in early childhood development, software developers, teachers and administrators who specialize in the development of software to support screening and assessment instruments.

Our goal is to help early childhood professionals track student progress with powerful, easy to use technology tools that assist in the development of quality educational programs for young children and their families. We do this through partnerships with content providers that provide researched based, reliable and valid assessment tools designed to assist teachers by identifying a child's strengths and those areas in which the child needs additional support. Each assessment is supported by an array of software platforms including handheld devices which provide teachers with the convenience of entering data and documenting a child's work through pictures, videos, and/or sound while working face-to-face with children or entering data online through an easy to use web page. Each software product includes easy to read and understand reports for the child, classroom, school, district, etc. to help teachers and administrators plan, make effective decisions, and implement strategies that can be used with their existing curriculum.

Red-e Set Grow

Measuring success one child at a time.

Red-e Set Grow, LLC • Attn: PAR Publications
P.O. Box 1674 • Clemmons, NC 27012
888-386-3822 • 336-777-0096 • email: PAR@redesetgrow.com

OnlinePAR.net

WF 1001

ISBN 978-1-940616-00-1