Predictive Assessment of Reading

Frank B. Wood, Deborah F. Hill, Marianne S. Meyer, and D. Lynn Flowers

Wake Forest University Health Sciences

Study 1 retrospectively analyzed neuropsychological and psychoeducational tests given to N = *220 first graders, with follow-up assessments in third and eighth grade. Four predictor constructs were derived: (1) Phonemic Awareness, (2) Picture Vocabulary, (3) Rapid Naming, and (4) Single Word Reading. Together, these accounted for 88%, 76%, 69%, and 69% of the variance, respectively, in first, third, and eighth grade Woodcock Johnson Broad Reading and eighth grade Gates-MacGinitie. When Single Word Reading was excluded from the predictors, the remaining predictors still accounted for 71%, 65%, 61%, and 65% of variance in the respective outcomes. Secondary analyses of risk of low outcome showed sensitivities/specificities of 93.0/91.0, and 86.4/84.9, respectively, for predicting which students would be in the bottom 15% and 30% of actual first grade WJBR. Sensitivities/specificities were 84.8/83.3 and 80.2/81.3, respectively, for predicting the bottom 15% and 30% of actual third grade WJBR outcomes; eighth grade outcomes had sensitivities/specificities of 80.0/80.0 and 85.7/83.1, respectively, for the bottom 15% and 30% of actual eighth grade WJBR scores. Study 2 cross-validated the concurrent predictive validities in an* N = *500 geographically diverse sample of late kindergartners through third graders, whose ethnic and racial composition closely approximated the national early elementary school population. New tests of the same four predictor domains were used, together taking only 15 minutes to administer by teachers; the new*

Annals of Dyslexia, Vol. 55, No.2, 2005

ISSN 0736-9387

Woodcock-Johnson III Broad Reading standard score was the concurrent criterion, whose testers were blind to the predictor results. This cross-validation showed 86% of the variance accounted for, using the same regression weights as used in Study 1. With these weights, sensitivity/specificity values for the 15% and 30% thresholds were, respectively, 91.3/88.0 and 94.1/89.1. These validities and accuracies are stronger than others reported for similar intervals in the literature.

Key Words: Fluency, phonemic awareness, prediction accuracy, reading, screening, vocabulary

INTRODUCTION

Early screening to predict future reading ability has become important in recent years. Testing in the early school years to predict concurrent and later reading achievement can identify children who need extra help in learning to read, and children so identified and helped can enjoy normal levels of success in the early to middles stages of their elementary school careers. As Torgesen (1998, p. 32, original italics) put it: "*The best solution to the problem of reading failure is to allocate resources for early identification and prevention*. It is a tragedy of the first order that while we know clearly the costs of waiting too long, few school districts have in place a mechanism to identify and help children before failure takes hold." Taking the matter to its logical conclusion, Torgesen urged including all children in the expectation of progress: "School-based preventive efforts should be engineered to maintain growth in critical word reading skills at *roughly normal* levels throughout the elementary school period."

This view represents a fundamental shift in the rationale of testing. Previously, testing was to measure a child's "potential," or else a child's progress in reaching that potential. In that view, many children's potential was too low to permit normal achievement. Here, by contrast, testing defines a child's need for help in reaching normal achievement, with the clear expectation that almost all children will indeed achieve normal levels of reading skill. While severe mental incapacity would undoubtedly still prevent some children from normal levels of reading, this new view unquestionably expects many more children to achieve at normal levels than previously thought.

Torgesen was not only revolutionary, but also prescient. The No Child Left Behind Act of 2001 proposed precisely what he advocated: an overall goal for all children to read normally.

Initially, the Act proposed major financial support for schools to institute scientifically valid procedures, both for early screening in kindergarten through third grade and for preventive and remedial efforts to those found to be at risk. Debates since the adoption of the Act have revolved around both the adequacy of the financial support that has been forthcoming and the scientific validity of the procedures required of schools supported under the Act, but these debates have not diminished the importance of the policy change that was attempted.

ACCURACY OF PREDICTION

In this new context, practical issues take first place: appropriate preventive and remedial efforts require accuracy in large-scale screening to predict concurrent and future reading skills in school settings. These predictions are of two kinds: (1) continuous, where the outcome is a score on an accepted measure of overall reading; or (2) dichotomous, where the outcome is simply whether a learner scores above or below an accepted threshold on that test.

In the first case, the accuracy (often called the strength) with which a set of predictor variables predicts a continuous criterion is often described by the squared coefficient of correlation (Pearson r^2 or multiple R^2), reflecting the percentage of variance in the outcome criterion that is being predicted. However, conventional notions of what constitutes a "strong" R^2 (as in Cohen's classical 1969 interpretation that any r^2 above .25 is "large") depend on a given sample being relatively normally distributed, with mean and variance that are closely similar to those of the population from which the sample is drawn. Sometimes, therefore, a better way to conceptualize accuracy of prediction is to consider the average size of the individual errors of prediction (differences between an individual learner's predicted outcome and that learner's actual outcome). The standard error of prediction is essentially the z-score of the errors around the prediction, so 1 standard error above or below the predicted score defines a margin within which 68% of the errors fall and 2 standard errors above or below the prediction defines the margin in which 95% of errors fall. The standard error can be an informative index for comparison across samples since the margin of error on the outcome measure speaks for itself in any given sample.

Given an outcome test that is sufficiently well predicted by the predictor variables, it is then sometimes also useful to describe the accuracy with which those predictors can predict an

educationally useful dichotomous cut score, differentiating typical from low performance on the outcome measure. Necessarily, the cut score is arbitrary, but is face-valid as a descriptor of the % of outcomes that it defines as low or impaired. See Meehl and Rosen (1955) for one of the classical discussions of the issues surrounding these types of categorical outcome predictions, the accuracy of which cannot be guaranteed solely by predictive validity coefficients such as R or R^2.

The conventional descriptors of accuracy in the prediction of dichotomous categorical outcomes (i.e., risk classification) are sensitivity and specificity. These refer to the criterion outcomes, specifically to how many of those outcomes are correctly predicted. In the present context in which a criterion threshold on the outcome measure distinguishes "low" from "typical" reading outcomes, sensitivity is that percentage of all the actual low outcome children who were correctly predicted by the screening test. Specificity is the percentage of all the actual normal outcome children who were correctly predicted by the screening test.

As a supplement to sensitivity and specificity, false positive and false negative rates are also sometimes used, although their definitions are not standard. Often (e.g., Torgesen, 1998; Catts, Fey, Zhang, & Tomblin, 2001), they are defined in terms of the predictions and how many of them are falsified by the actual outcome. The present report also adopts that definition: false positive rate is the percentage of predictions of "low" reading that turn out to be incorrect because the actual outcome was "typical." False negative rate is then the percentage of predictions of "typical" reading that turn out to be incorrect because the actual outcome was "low."

The ratio of sensitivity to specificity, and of false positive to false negative rates, is adjustable by raising or lowering the cut score for prediction. Often, the cut score for prediction is raised to a level higher than the actual outcome threshold being predicted, so that more cases are predicted to be "low" than will actually turn out "low." By definition, that elevates false positive rates, but also reduces false negative rates by minimizing the likelihood of predicted "typical" cases actually turning out "low." That is usually a favorable balance since the cost of false negative errors is high (missing the opportunity to help a child who would actually have a low outcome). False positive errors, on the other hand, might invoke additional remedial assistance to a student whose reading outcome would have been above the threshold of normalcy even without that assistance. How-

ever, that assistance would not be adverse and would almost certainly be helpful, particularly since the outcome—although higher than the cut score—would, in most cases, still be below average.

Extant studies have tended to report predictions that range from concurrent to three years forward, and have reported sensitivities ranging between 56% and 92%, with specificities somewhat better at 80% to 92% (see review by Snow, Burns, & Griffin, 1998). As a benchmark, the Committee on Children with Disabilities of the American Academy of Pediatrics (AAP) has issued a Policy Statement (2001) commenting that good developmental screening tests should have sensitivities and specificities that are both at least in the range of 70% to 80%, so against that standard, most available prediction paradigms would compare favorably, although it must still be remembered that an overall accuracy of 80% means that fully 20% of learners will be incorrectly classified. As to false positive and false negative rates, they must be separately considered since they cannot be directly calculated from specificity and sensitivity values alone. Torgesen's (1998) review concluded that false positive rates—as we have defined them above—tended to range from 20% to 60% with an average around 45%, whereas false negative rates tended to range from 10% to 50% with the average around 22%. Here, too, a threshold that generates high false positive rates is usually acceptable, but if that produces false negative rates that are on average no better than 22%, then the educational usefulness of these screening paradigms would still seem limited: more than one in five of the learners predicted to have satisfactory outcomes are destined for a low outcome and would miss the opportunity to be recognized and helped. That may not be altogether satisfactory in an educational setting where the goal is for almost all learners to succeed in reading.

PREDICTION STUDIES

Reviews of prediction paradigms (Scarborough, 1998; Torgesen, 1998; Snow et al., 1998) have noted that phonemic awareness and fluency variables seem essential to effective prediction, and Scarborough (1998) was insightful in noting that vocabulary is equally important.

It is generally accepted (Snow et al., 1998) that effective prediction paradigms require more than one predictor variable.

Several issues require consideration in the interpretation of available prediction studies. For example, Scarborough (1989)

included family history of reading problems in her predictor model: that is of major theoretical importance since family history variables turned out to be the strongest single predictors, but it is arguably impractical to ask all parents in a large-scale school screening setting about their family reading histories. In a similar vein, Sénéchal, LeFevre, Thomas, and Daley (1998) showed that children's oral (but not written) language development was enhanced by their parents' reading storybooks to them, whereas the informal real-time teaching that parents often do during such reading to their children favorably influenced their written (but not oral) language development. Again, that result is of major theoretical significance, but assessment of parents' storybook reading and personal teaching to their children is unlikely to be practical in large-scale school-based screening situations. Finally, the study of Catts et al., (2001) raises a similar, if milder, question: one of their effective predictors was biological mothers' education. While this may be easier to assess than family history or parental involvement in storybook reading and teaching, it is not free of difficulties including privacy (asking about maternal education when some parents would not wish to answer) or practicality (some biological mothers are present or their educational histories are not available, as in adoption). Clearly, the most convenient large-scale screening would be that which involves the children themselves and does not rely on family variables; a major empirical question that emerges is whether useful predictive formulae are available when the predictor variables are confined to tests given to children.

Similarly, to be educationally relevant, the criterion outcomes need to include comprehension as well as single word or nonword decoding. For example, the Elbro, Borstrøm, and Petersen (1998) study in Denmark is of considerable theoretical importance because it includes children at genetic risk, and its sensitivity and specificity values (78% and 79%, respectively) are at the high end of the range expected by the AAP Policy Statement (Committee on Children with Disabilities, 2001). However, since the outcome criterion is limited to a composite of nonword reading and pseudo-homophone detection, teachers would likely find it somewhat removed from the global text comprehension skill that is measured in most standardized achievement tests (not to mention the now ubiquitous high stakes, end-of-grade accountability tests). Studies using standardized reading achievement tests that include comprehension components are, therefore, of more direct practical use.

Among studies limiting their variables to child test scores, and using outcomes that include text comprehension, one consideration remains important to the interpretation of prediction studies: are the findings generalizable to situations and populations where the screening paradigm would be used? Two somewhat different issues apply.

1. Unless the study in question appears to represent a normally distributed population sample, representing the major ethnic and geographic diversity of the population in question, then the strengths or accuracies of prediction will not necessarily generalize to future normally distributed samples. Catts et al. (2001) successfully addressed the issue by differentially weighting the sample so as to approximate mathematically a normal distribution.
2. Cross-validation is quite helpful in ensuring that a given prediction formula works in subsequent trials. There are, however, no formal reports of attempted cross-validation assessing the generalizability of prediction formulae.

Three studies appear to illustrate the current state of the art in school-age prediction. A fourth study by Badian (1994) deserves special note: it screened prekindergartners and assessed first grade outcomes on standard achievement tests, with sensitivity 80.0% and specificity 87.4%, also false positive 52%, and false negative only 3.2%. Prekindergarten screening is important in its own right, and Badian's results are encouraging, but not directly comparable to school-age predictions in kindergarten and later.

Consider Flynn (2000), who examined $N = 210$ kindergartners with a battery of classroom administered paper and pencil tests, employing standard achievement test outcomes in first and third grades. For predicting a normative 40th percentile outcome (comprising 23% of her slightly above average sample), Flynn's battery alone yielded sensitivity, specificity, false positive, and false negative percentages of 80, 72, 31, and 20, respectively. Flynn considered these only marginally successful because of the relatively high false positive rate, but was able to improve the outcomes when teacher ratings were added to the predictors: the percentages were then 88, 57, 39, and 12, respectively, a reasonably good prediction that would be of educational utility. Flynn noted that teacher ratings depended to some extent on teacher training, and she also advocated multiple

screenings across the kindergarten and first grade period in order to improve the accuracy of the results.

The Texas Primary Reading Inventory (Foorman, Fletcher, & Francis, 1998) has had wide use, and there is much field experience presented in their technical report. As one example, for predicting the spring outcomes of the Woodcock-Johnson-Revised Broad Reading (Woodcock & Johnson, 1989) from the fall screening, they show sensitivity, specificity, false positive, and false negative rates of 93.3, 63.5, 61.2, and 2.6, respectively. (It must be noted that they report false positive and negative values using a definition different from ours, so we recalculated those particular values from their data to obtain the above results). Foorman and colleagues also provided reliability information, showing a median internal consistency reliability of .875, across subtests on two testing occasions. This battery is brief and includes a single word reading subtest, which tends to improve the predictive validity since single word reading is also a component of the outcome measure. This is a reasonable way to improve the predictive power, but see also the section on autocorrelation artifact in the method section for Study 1, below.

Shaw and Shaw (2002) showed useful prediction of April third grade reading achievement scores in a Colorado school, from April third grade testing on the Dynamic Indicators of Basic Early Literacy Screening, Oral Reading Fluency subtest. Sensitivity, specificity, false positive, and negative percentages were 73.3, 90.7, 26.7, and 26.7, respectively. As in Flynn (2000) above, the relatively high false negative rate would be of concern: a fairly substantial number of children who would perform below threshold on the outcome were not identified as such on the screening.

In general, the above studies show sensitivities and specificities in the 80% range so that whenever one value was high, as in the 93.3% sensitivity reported by Foorman et al. (1998), then the other tends rather lower (63.5% specificity in Foorman et al.). Interestingly, of the three school age studies, only Flynn's (2000) included vocabulary. The question left by the current state of prediction studies is then simply this: would the inclusion of all four constructs, already identified as necessary by separate reviewers (Scarborough, 1998; Torgesen, 1998; Snow et al., 1998) yield stronger and more efficient predictions?

The two studies reported here constitute, respectively, an initial demonstration and a cross-validation. In Study 1, four predictor variables, measured in first grade, were assessed for their joint ability—in a linear regression model—accurately to

predict concurrent (first grade) and future (third, eighth, and 12th grade) reading outcomes in a local sample from public schools. Study 2 was then conducted as a formal cross-validation of these predictive relations to test whether the same concurrent predictive relations could be found with newly built tests measuring the four predictor constructs, a new revision of the outcome reading criterion, and a new geographically diverse sample from six states across the United States.

STUDY 1: PREDICTION OF CONCURRENT AND FUTURE READING ACHIEVEMENT SCORES IN A POPULATION-BASED LONGITUDINAL SAMPLE

PARTICIPANTS

By ethnically and socioeconomically based stratified random sampling, 485 children were recruited with parental consent from the N = 3,011 first graders in the Winston-Salem/Forsyth County school system under protocols approved by the Wake Forest Health Sciences Internal Review Board. Testing of these 485 children began in November, but in order to reduce variance that might be related to the initial date of testing, the present sample for this report was restricted to those children who were tested during the second semester and the subsequent summer break. Some members of this cohort were retested in third grade and eighth grade, and the present longitudinal sample for Study 1 consists of 220 children who were tested at all three grade levels on the battery of cognitive tests reported below. Below, we refer to this sample as Cohort 1 to distinguish it from the N = 500 cross-validation sample reported in Study 2.

Intensive recruiting procedures were undertaken to ensure that Cohort 1 preserved the full diversity of the original N = 485 sample, both ethnically and in terms of ability levels. Cohort 1 was 51% male and 71% majority race. Except for three individual students—two Asian-American and one Hispanic-Latino—the minority ethnicity was entirely African American. The Peabody Picture Vocabulary Test–Revised (PPVT-R) (Dunn & Dunn, 1981) was administered in first grade and was adopted as the initial criterion for estimating the degree to which the cohort was matched to national norms on a measure of verbal ability. The cohort obtained a mean score of 100.9 on the PPVT-R, with standard deviation of 14.6 and range from 49 to 133,

thereby closely resembling the national norms and ensuring the expected number of participants at all ability levels.

ASSESSMENT INSTRUMENTS

1. Reading Criterion Tests. The major criterion test of reading was the Woodcock-Johnson Psycho-Educational Battery Broad Reading Standard Score (WJBR) in its original version (Woodcock & Johnson, 1977). It was administered in first, third, and eighth grade—and for a subset of children—in 12th grade. The composite broad reading score on the 1977 version of the WJBR combined the scores from separate measures of letter-word identification, word attack, and passage comprehension.

Since findings on the major WJBR criterion could conceivably be specific to that particular measure, prediction of the Gates-MacGinitie (GM) (MacGinitie, 1978) Reading Test, administered to all subjects in eighth grade, was also examined to see if the predictive relations obtained on the WJBR would generalize to a very different type of reading criterion test. Unlike the individually administered WJBR, the Gates-MacGinitie is a paper and pencil test, and is timed, but otherwise requires minimal interaction with the examiner. In the present study, children were left alone in a quiet room to complete the GM within the stated time limits. The GM total score combines multiple choice tests of reading vocabulary (from among synonyms for a target word) and reading comprehension (involving choosing the correct answers to questions about a previously presented narrative or explanatory text).

2. Predictive Constructs. Although a large number of tests was given in first grade (see Felton & Wood, 1989, for a comprehensive presentation and analysis of their concurrent validities or lack thereof), we selected only four constructs, each defined as a composite of two different tests that measured that construct in somewhat different ways. For each of the four constructs, the two tests comprising that construct were each separately standardized on the sample to mean 100 and standard deviation 15. The equally weighted average of these two was then restandardized to mean 100 and standard deviation 15, to become the composite measure. These constructs, and the tests comprising them, were as follows.

A. Total Phonemic Awareness (PHONEMIC AWARENESS). The first of the two variables comprising PHONEMIC AWARENESS was a Phonemic Analysis Cluster (PAC), derived from the pioneering work of Stanovich, Cunningham, and Cramer (1984). Their "Strip Initial Consonant" component re-

quired the subject to delete the initial phoneme of a word spoken by the examiner, pronouncing the word that remains, and their "Final Consonant Different" task requires the subject to listen to four words and choose the one with a different ending sound. As the two tests each have only 10 items, the scores from both tests were added to generate a single, 20-item PAC score. The second member of the PHONEMIC AWARENESS composite was the Lindamood Auditory Conceptualization Test (LAC) (Lindamood & Lindamood, 1979), which required the child to manipulate wooden blocks of different colors to indicate speech sound patterns in two categories: isolated sounds in sequence, and sounds within a syllable (Lindamood & Lindamood, 1979). Untimed accuracy was scored. It may be noted that the PAC is expressive in the sense of requiring a spoken response, whereas the LAC is receptive in the sense of requiring only a pointing or manual manipulation response to the examiner's instruction.

B. Total Picture Vocabulary (PICTURE NAMING VOCABULARY). Its first member variable is the Boston Naming Test (BNT) (Kaplan, Goodglass, & Weintraub, 1983). This test simply requires the subject to name the single picture depicted in a line drawing. Originally developed for use in testing for the naming difficulty that defines the anomic symptom complex within aphasia, the test offers items ranging from very common (suitable for kindergartners) to relatively infrequent (suitable for adults). Simple, uncued accuracy of word naming, untimed, was scored. The second variable in PICTURE NAMING VOCABULARY was the Peabody Picture Vocabulary Test-Revised (PPVT-R) (Dunn & Dunn, 1981). This test presents sets of four simple line drawings and asks the subject to point to the one that matches the word spoken by the examiner. Like the Boston Naming Test, it is untimed and its items span a difficulty range from late preschool years to adulthood. The total correct, untimed, was scored. As in the PHONEMIC AWARENESS composite above, the first of the PICTURE NAMING VOCABULARY tests (BNT) is expressive, requiring a spoken response to the picture, whereas the second test (PPVT-R) is receptive, requiring only a pointing response to a picture depicting the examiner's spoken word.

C. Rapid Automatized Naming (RAN, or RAPID NAMING). Developed by Denckla and Rudel (1976), the tests consist of four cards of 50 items each (digits, letters, color squares, and simple line drawings of objects, presented in five rows of 10 items each). The subject's task is to name the items on a card as rapidly as possible; the score for each card is the number of

seconds required to do so. The first of the two RAN scores was obtained by first adding the number and letter time scores, then taking the natural logarithm of that sum to generate the number-letter (NUMLET) time score. Similarly, the second RAN score (COLOB) was the natural logarithm of the sum of the color and object naming time scores. These logarithmic transformations successfully normalized the distributions; their sum was then standardized, in the manner described above, to yield the overall RAPID NAMING score. NUMLET and COLOB, while both expressive in the sense of requiring spoken responses, are nonetheless conceptually distinct. NUMLET is inherently a kind of "reading" task inasmuch at it involves alphanumeric stimuli, while COLOB is more narrowly a rapid "naming" task involving stimuli that are not alphanumeric.

D. Total Single Word Reading SINGLE WORD READING COMPOSITE. The first member of this composite is the Real Word subtest from Part II of the Decoding Skills Test (DST-WORD) (Richardson & DiBenedetto, 1985). DST-WORD has 60 items, half monosyllabic and half polysyllabic. The other member of the SINGLE WORD READING COMPOSITE construct is the Letter Word Identification subtest from the WJBR itself (WJ-WORD). These two real word reading tests are also subtly different: DST-WORD is a criterion referenced test where all items are expected to be correctly answered by typical fifth graders, and WJ-WORD is norm referenced and has items across a much greater range of difficulty, including adult levels.

AUTOCORRELATION BETWEEN PREDICTORS AND OUTCOME CRITERIA

When the SINGLE WORD READING COMPOSITE construct is used in models that predict the Woodcock-Johnson Broad Reading score, then the WJ-WORD subtest is included both in the predictor and the outcome composites, necessarily creating an autocorrelation artifact that locks in some shared variance due to the same test. In a somewhat broader context, even the DST-WORD component of SINGLE WORD READING COMPOSITE could arguably be thought of as generating a type of autocorrelation artifact since it, too, is a single word reading test, albeit one where items are different. Despite this, various predictive models routinely use single word reading tests (e.g., Foorman et al., 1998), justifiably, it may be remarked, since they can improve the strength of prediction. As described below in the analyses of results, we consider the potential autocorrelation artifact in three different contexts. (1) The complete autocorrela-

tion artifact occurs when the full WJBR in first grade is used to predict the third and eighth grade WJBR outcomes, so in that case, all components of the predictor are present in the criterion, and correlations between present and future WJBR define the maximum possible autocorrelation. This can then be compared to the results from the four variable model, and if the four variable models predict future WJBR at least as strongly the first grade WJBR predicts future WJBR, then autocorrelation artifact cannot reasonably explain the full strength of the prediction by the four predictor variables. (2) In this context, it also becomes useful to consider separately the strength by which the three other variables, not including SINGLE WORD READING COMPOSITE, jointly predict the WJBR outcomes; such predictions are free of autocorrelation artifact and so provide an estimate of the predictive importance of the other three variables (PHONEMIC AWARENESS, PICTURE NAMING VOCABULARY, and RAPID NAMING). (3) The eighth grade Gates-MacGinitie contains no single word reading subtest, so if the four variables predict eighth grade GMs as well as they predict eighth grade WJBR, then autocorrelation cannot explain the strength of the predictive relationships. (When reading skill is at issue, the PICTURE NAMING VOCABULARY predictor would not be considered auto-correlated with the word reading subtest of the Gates-MacGinitie.)

RESULTS

Statistical analyses were directed first toward the accuracy or strength by which the four predictors predicted reading outcomes. Those results are shown in table I, and reveal an uncommonly strong prediction, with high percentages of variance accounted for in prediction of the respective first, third, and eighth grade outcomes (ranging from 89% to 69%). Table I also enables a comparison to the fully autocorrelated predictions by first grade WJBR, of future WJBR outcomes. That comparison shows in each case that the four predictor variables predict future WJBR slightly better than the first grade WJBR itself does, suggesting—as discussed above—that autocorrelation cannot explain the full strength of the four variable predictive models. Similarly, it is noted that the four variable prediction of the Gates-MacGinitie—which is free of autocorrelation artifact—is as strong as the prediction of the WJBR, both predictions accounting for 69% of the variance.

The WJBR outcome variable can be categorically dichotomized and accuracies of prediction then calculated. Two

Table I. Multiple regression analyses for predictions of first, third, and eighth Grade Woodcock-Johnson Broad Reading and eighth Grade Gates-MacGinitie, from first Grade four variable predictor model and from first Grade Woodcock-Johnson Broad Reading, for $N = 220$ longitudinal sample.

	Predictions from first Grade Four-Variable Model				
Outcome Predicted	$F(4, 215\ df)$	R	R^2	Standard error	Probability
Grade 1	442.15	0.94	0.89	4.5	<.0001
Grade 3 WJBR	174.56	0.87	0.76	7.04	<.0001
Grade 8 WJBR	121.58	0.83	0.69	7.83	<.0001
Grade 8 GM	119.3	0.83	0.69	8.43	<.0001
	Predictions from first Grade WJBR				
Outcome Predicted	$F(1, 219\ df)$	R	R^2	Standard error	Probability
Grade 3 WJBR	610.2	0.86	0.74	7.39	<.0001
Grade 8 WJBR	455.78	0.82	0.68	7.98	<.0001
Grade 8 GM	267.29	0.74	0.55	10.08	<.0001

Note: WJBR is Woodcock-Johnson Broad Reading, 1977 edition, a composite of Single Word Reading, Word Attack, and Passage Comprehension, individually examiner-administered. GM is Gates-MacGinitie Reading Test, a composite of Word Reading Vocabulary and Text Comprehension, both in a multiple choice paper and pencil format. The four variable predictor model comprises Phonemic Awareness, Picture Vocabulary, Rapid Naming, and Single Word Reading as described in the "Assessment Instruments" section of the Method in the text.

Note also: Standard errors are reported in terms of the actual score being predicted: for the Woodcock-Johnson Broad Reading, it is the grade referenced Standard Score delivered by the Woodcock-Johnson norms, having a nationally standardized mean of 100 and standard deviation of 15. The eighth Grade Gates-MacGinitie was restandardized to mean 100 and standard deviation 15 so as to make its standard error values comparable to those for the Woodcock-Johnson Broad Reading.

separate categorical cut score thresholds (for dividing the outcomes into "low" versus "typical") were considered: the 15th percentile of the outcomes (argued by Catts, et al., 2001, as the best compromise among available studies), and the 30th percentile of outcomes (as advocated by Torgesen, 1998, 2004). The former threshold—one standard deviation below the mean—has not only a certain face validity (any learner in the bottom 15% of the population might for that reason alone be considered as having some difficulty) but also closely resembles the 18th percentile threshold found in our genetic studies (Grigorenko et al., 1997, 2001; Grigorenko, Wood, Meyer, & Pauls, 2000) to be the cut score that best models the distinction between genetically affected versus unaffected cases. Torgesen's (1998, 2004)

higher 30th percentile threshold recommendation, on the other hand, derives from practical educational experience: educators are properly concerned as much for learners performing in a below average range (e.g., 15th to 30th percentile) as for the distinctly lower performing (below 15th percentile) learners. Accordingly, we report sensitivities and specificities for both thresholds. These are shown in table II, revealing uncommonly high accuracies for the concurrent prediction in first grade.

As described above, each of the four predictive constructs is itself a composite of two variables, each of them in turn measuring the same general construct by two somewhat different methods. That makes it possible to "split" the constructs into two alternate forms: Version A comprising PAC, BNT, NUMLET, and DST-WORD; and Version B comprising LAC, PPVT-R, COLOB, and WJ-WORD. Each version predicts the outcome criteria (of concurrent first, third, and eighth WJBR, and eighth GM) with accuracies approaching those derived from the composite constructs. In particular, between 0% and 6% less of the variance was explained by either version A or B than by the battery using the composite constructs; the median was 2% less variance explained. The alternate forms reliability compares the predictions by version A and version B; these are all at $r >= .90$, median $r = 92.5$, for each of the predictions (of concurrent first, third, and eighth WJBR, and eighth GM).

Table II. Accuracy of prediction of concurrent and future reading scores from a brief predictive battery in Grade 1, expressed as percentages. N = 220.

	Prediction of the 15% cut score for the WJBR outcome					Prediction of the 30% cut score for the WJBR outcome				
Grade	Sens	Spec	FP	FN	Acc	Sens	Spec	FP	FN	Acc
1	93.0	91.0	28.6	1.8	91.4	86.4	84.9	23.1	8.5	85.5
3	81.4	81.4	48.5	5.3	81.4	80.5	82.0	25.5	13.5	81.4
8	80.0	80.0	45.9	6.8	80.0	85.7	83.1	24.2	9.6	84.1

Note: The cut scores define, respectively, the bottom 15% or bottom 30% of the sample as measured on the outcome test (WJBR). The thresholds for prediction encompassed approximately 5% additional cases above the cut score, thereby minimizing false negatives.

Sens = sensitivity; Spec = specificity; FP = false positive; and FN = false negative. See the "Accuracy of Prediction" section of the Introduction for mathematical definitions. Acc = overall accuracy, the percentage of all the cases in the sample that were correctly predicted.

In the context of the above discussion of autocorrelation, separate analyses were done in which the SINGLE WORD READING COMPOSITE construct was dropped from the predictor variables. The three variable multiple regression models accounted then for 71%, 65%, 61%, and 65% of the variance, respectively, in first, third, and eighth Grade WJBR and eighth Grade GM outcomes. Consistent with the double-deficit model and our own work (Meyer, Wood, Hart, & Felton, 1998; Wolf & Bowers, 1999), the interaction term (PHONEMIC AWARENESS x RAPID NAMING) replaced the two separate predictors in the optimized prediction model, accounting for, but only in prediction of, the eighth grade Gates-MacGinitie. The individual constructs alone, in a bivariate correlation, predicted the following percentages of variance (r^2): for PHONEMIC AWARENESS, the r^2 values (x100) were 66, 55, 49, and 49 for the respective first, third, and eighth Grade WJBR and eighth Grade GM outcome predictions; for PICTURE NAMING VOCABULARY, they were 37, 36, 37, and 52, respectively; and for RAPID NAMING, they were 23, 20, 20, and 17, respectively. Notably, for the predictors PICTURE NAMING VOCABULARY and RAPID NAMING, there is no substantial decrease over time in their correlations with outcome criteria, but PHONEMIC AWARENESS, on the other hand, does lose predictive power over this same interval.

A general discussion of the findings from Study 1 is deferred until after the presentation of the results of Study 2.

STUDY 2: CROSS-VALIDATION OF THE CONCURRENT PREDICTION IN A NATIONALLY REPRESENTATIVE KINDERGARTEN THROUGH THIRD GRADE SAMPLE, N = 500.

New items were developed for each of the subtest domains in Study I. Only one test was developed for each domain: in each case, the briefest of the two types of tests used in Study 1 for that domain. Thus, the new test battery consisted of (a) 30 single letter or word identification items; (b) 35 line drawings of objects for naming vocabulary; (c) 20 phonemic awareness items requiring same-different judgments of beginning or ending consonants and phoneme deletion from beginning, end, or middle of the spoken word; and (d) rapid naming of 50 letters and 50 digits presented separately as five rows of 10 digits and five rows of 10 letters. Except for the phonemic awareness sub-

test, the other three were "wide-range" in the sense that the same items, across a range of difficulty, were used for all grades (second semester kindergarten through end of third grade). The phonemic awareness test for kindergartners included 10 same-different beginning consonant judgments and 10 initial phoneme deletion items; for first graders, it included the initial phoneme deletion task accompanied by 10 same-different final consonant judgments; and for second and third graders alike, it included the final consonant task accompanied by 10 new phoneme deletion items from beginning, end, or middle of the word. Preliminary versions of this test battery, containing more items, were administered along with the Woodcock-Johnson III (Woodcock, McGrew, & Mather, 2001) at the invitation of other schools in North Carolina for their educational use. These preliminary administrations permitted the selection of a final set of items that in the aggregate, were equally good predictors of reading across ethnic groups. This predictive assessment of reading (PAR) became the new field test.

The Woodcock Johnson third revision (WJ–III) Broad Reading was used as the concurrent criterion. It differs from the WJBR by replacing the word attack (nonword reading) subtest with a sentence reading fluency subtest. New schools in North Carolina, New York, Minnesota, Colorado, Arizona, and California invited the final version of PAR as well as the criterion (WJ–III) testing in their second semester kindergarten through third grade classes for their educational use. The general procedure, with few exceptions, was for teachers or substitute teachers to administer the PAR and for WFUHS psychologists, blind to the predictor testing, to administer the WJ-III, never less than a day apart and never more than four days apart. In no case did the same person administer both assessments, and examiners were always blind to the prior assessment results. These schools included a wide demographic range from which it was possible to sample randomly within ethnic strata, and within a normal distribution of performance on the WJ-III Broad Reading. This achieved a sample where percentages of African-American, Hispanic, Asian, and other, and majority-race (Caucasian) students were highly similar to those found in the national early school grades population. At each grade level, the African American and Hispanic-Latino students comprised 20% each; majority race 57%; and 3% Asian and other. The mean WJ-III Broad Reading score was 100.1 with standard deviation of 15.3. There was no significant departure from normal curve parameters on any test of normality. There

were at least 100 cases in each grade from second semester kindergarten through third grade.

The four subtests were then examined for their ability jointly to predict the concurrent criterion, WJ-III Broad Reading (WJ-III BR) standard score. First, the scores were standardized to their respective grade levels with a correction for season of the year. Then, two analyses were done: (1) the four subtests were submitted to a multiple regression prediction of WJ–III BR to find the optimal regression weights for the prediction; and (2) the original regression weights from Study 1 were applied to the subtests, and the predicted reading score derived from those weights was then also correlated with the WJ-III BR criterion. The former procedure finds the maximum prediction formula, but it also capitalizes on chance variations in the new sample; the latter procedure simply attempts an exact cross-validation of the predictive power of the weights derived in the earlier Study 1, avoiding any capitalization on chance in Study 2. The similarity in strength of the regression prediction, between the two procedures, then reflects the robustness of the prediction.

RESULTS

A multiple regression solution for the optimized prediction of the WJ-III standard score, by the four subtests of the new PAR, yielded a multiple R of .929, R^2 of .863, and standard error of 5.69. However, regression weights from the concurrent (first to first grade) prediction in Study 1 were also applied to these new data from Study 2; these yielded R of .926, R^2 of .857, and standard error 5.80. The cross-validation was, therefore, highly successful, suggesting that the cross validated weights are robustly transferable to a new set of predictors and a new criterion. These were, therefore, considered the preferred weights for any future use. Grade level was tested for any contributory effect on the regression. It had nothing close to a significant effect (i.e., once the grade-standardized values were entered into the prediction equation), grade level made no significant additional contribution to the prediction, and there were no significant changes across grade level in the strength or accuracy with which this equation predicted the criterion. No significant slope or intercept bias was observed across the ethnic groups, and the results were equally accurate within these subgroups and within individual grade levels. The scatter plot of predictions using the original Study 1 weights is presented in figure 1.

Further evidence of the stability of the predictors is available from the Cronbach's Alpha coefficients of internal consis-

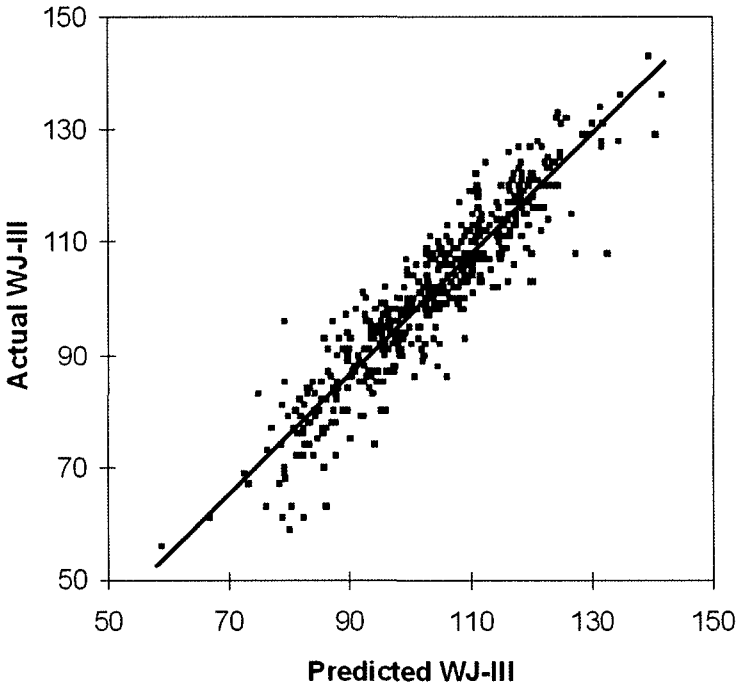

Figure 1. Scatter plot and regression line for prediction of actual Woodcock-Johnson III Broad Reading test performance from concurrently administered predictive assessment of reading, in a nationally representative sample of N = *500 kindergarteners, and first, second, and third graders.*

tency reliability calculated from the three item-based tests. Controlling for grade level, the alpha's were .90, .92, and .93, respectively, for picture naming vocabulary, phonemic awareness, and letter-word calling. Sensitivity and specificity values for study 2 were at least as high as those for study 1 (see table III).

DISCUSSION

In Study 1's predictions from first to third or first to eighth grade, over two-thirds of the variance in outcomes is predictable, and the accuracies meet or exceed both the AAP (Committee on Children with Disabilities, 2001) standard, and also the accuracies in the relevant extant literature (see again the section "Prediction Studies" in the Introduction). Thus, Flynn (2000), predicting from one to four years forward, and Foorman et al. (1998) predicting from fall to spring within a school year,

Table III. Concurrent accuracy for predicting WJ–III BR from PAR, in Study 2, compared to Study 1.

	Prediction of 15th percentile Outcome Criterion					Prediction of 30th percentile Outcome Criterion				
	Sens	Spec	FN	FP	Acc	Sens	Spec	FN	FP	Acc
Study 1, $N = 220$ first graders	93.0	91.0	28.6	1.8	91.4	86.4	84.9	23.1	8.5	85.5
Study 2, $N = 500$ K, 1st, 2nd, & 3rd graders	91.3	87.99	36.8	2.18	88.6	94.12	89.14	16.2	3.79	91.0

Note: the results for Study 2 are combined across grade levels. As described in the text, the predictor variables (like the WJ–III itself) are standardized to mean 100 and standard deviation 15 within grade levels.

both had specificities (57% and 63.5%) below the AAP standard. Study 1's long-term prediction accuracies even exceed the within-month prediction by Shaw and Shaw (2002), which while technically meeting the AAP standard with sensitivity and specificity of 90.7 and 73.3, show a false negative rate of 27%, which is almost certainly too high for practical use (since, if 27% of learners classified as typical actually turn out low, a large number of needy individuals is being missed).

While the percentage of variance accounted for by the predictions of eighth grade outcomes is—as expected—less than that for the third grade outcome, it is interesting that the classification accuracies for eighth grade outcomes are actually slightly higher than (but not statistically different from) those of the third grade predictions. Were that a true reflection of longitudinal dynamics in the population, it would suggest some progressing separation of the low and typically functioning students, so that even though standard errors of prediction get larger as the learners progress through the grades, the difference between low and typically functioning students might also widen, thus preserving the overall classification accuracies in the low to mid 80% range, with comfortably small false negative percentages, in all cases less than 10%. In other words, it is possible that a bimodal distribution, separating low from typical readers, emerges over time: larger longitudinal samples would, however, be required to confirm that possibility.

Study 1's future predictions confirm the particular salience of phonemic awareness. On its face, it is astonishing that so nar-

row a cognitive skill, measured in first grade, involving no viewing of print, by itself predicts virtually half the variance in a standard individually administered reading achievement test given in eighth grade. At the same time, it is interesting that although the predictive power of phonemic awareness remains strong even for the eighth grade outcome, it nonetheless steadily diminishes from 65% of variance to 49%, confirming Badian's (1995) report suggesting that phonemic awareness has particular relevance for predicting early, more than later, reading achievement, even in this case when both predictions are from the same first grade time point.

Phonemic awareness yields priority to vocabulary when the prediction is to eighth GM instead of to eighth WJBR. It has to be assumed that the explicit reading vocabulary and text comprehension features of the GM are responsible for the somewhat greater predictive salience of picture naming vocabulary. More generally, it confirms an occasional caution in the literature (e.g., Scarborough, 1998) not to overlook the importance of vocabulary, particularly when dealing with the reading—for meaning—of extended text. Similarly, the double-deficit model (Wolf, 1991; Meyer et al., 1998) also gains some support from these results in the sense that rapid naming adds additional predictive variance, which, in predicting eighth grade GM, interacts with and, therefore, amplifies the effect of phonemic awareness.

Study 2 cross-validates the concurrent predictive strengths and predictive accuracies found in Study 1. The application of the regression weights from Study 1 to the variables in Study 2 results in little loss of concurrent predictive strength, with $R = .92$ and $R^2 = .85$. That the high predictive variance survives new versions of both the predictors and the outcome criterion gives confidence that major, stable domains of reading-related variance are being successfully measured. That confidence is further enhanced by the demonstrably strong internal consistencies of the three item-based tests in the Study 2 protocol.

In the WFUHS studies, genetic linkage and association analyses have played an increasingly important role in clarifying the behavioral data (see Wood & Grigorenko, 2001, for a methodological review). These studies suggest a degree of biological validity to the predictive models reported above. Specifically, each of these four behavioral constructs is individually linked through informative markers on chromosome 6 to a gene or genes in the vicinity of 6p21.3 (Grigorenko et al., 1997; Grigorenko et al., 2000).

McCardle, Scarborough, and Catts (2001) consider a variety of candidate models for explaining the predictive relationships: core phonological, auditory temporal processing, double deficit, and language-based. The present results suggest that such a typology should not be construed as series of alternatives for "the" single mechanism of reading deficit. Instead, the predictive relationships reported here suggest a continuous model in which each tested mechanism contributes variance to the overall outcome.

While it may be reasonable to suggest that the uncommonly strong predictive accuracy of these models may derive in part from their genetic validity, it must be emphasized that the biological validity of these theoretical constructs in no way implies a fixed, irremediable deficit. It is appropriate, therefore, to consider the implications for educational practice of these unexpectedly strong predictive relationships. The data on classification accuracy for these replicated concurrent predictions provide assurance that they could confidently be used to identify children needing additional remedial help.

The predictive assessment in Study 2 takes no more than 15 minutes to administer, and teachers were proven in Study 2 to generate scores that are as reliable and valid as those produced by psychologists. With such short administration times, teachers can feasibly give the tests themselves, and gain the opportunity to observe the theoretical mechanisms first hand, particularly the phonemic awareness and rapid naming constructs, which remain somewhat abstract or seemingly removed from ordinary classroom activities. To see a child struggle unexpectedly with the simplest phoneme deletion or rapid naming tasks is to appreciate their relevance in ways that can never be communicated with correlation coefficients, however high.

Notably, the subtests are individually reliable and, therefore, yield not only a composite screening result but also a plausible strength and weakness profile. For example, a learner whose low predicted reading score derives mostly from low vocabulary (as sometimes seen in English language learners) would not necessarily benefit from the intensive phonological remedial instruction, or would at least need remedial attention to vocabulary; that learner's needs would certainly differ markedly from those of the "double-deficit" learner with disproportionately low phonological and fluency skills. As noted in the results, the test given in English to Hispanic-Latino learners predicts their English reading as accurately as for native English speaking learners. An obvious next research question would address the practical utility of the test profile for guiding in-

struction differentially, according to the particular needs of individual learners.

ACKNOWLEDGMENTS

Supported by P01 HD 21887, and by a grant from The Dyslexia Foundation.

Address correspondence to: Frank Wood, Ph.D., Wake Forest University Health Sciences, Winston-Salem, NC 27157-1043. E-mail:fwood@wfubmc.edu

References

Badian, N. A. (1994). Preschool prediction: Orthographic and phonological skills, and reading. *Annals of Dyslexia, 44*, 3–25.

Badian, N. A. (1995). Predicting reading ability over the long term: The changing roles of letter naming, phonological awareness, and orthographic processing. *Annals of Dyslexia, 45*, 79–96.

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.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Committee on Children with Disabilities, American Academy of Pediatrics. (2001). Developmental surveillance and screening of infants and young children. *Pediatrics, 108*, 192–195.

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

Dunn, L. M., & Dunn, L. M. (1981). *Peabody picture vocabulary test–revised.* Circle Pines, MN: American Guidance Service, Inc.

Elbro, C., Borstrøm, I., & Petersen, D. K. (1998). Predicting dyslexia from kindergarten: The importance of distinctiveness of phonological representations of lexical items. *Reading Research Quarterly, 33*, 36–60.

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

Flynn, J. M. (2000). From identification to intervention: Improving kindergarten screening for risk of reading failure. In N. A. Badian (Ed.), *Prediction and prevention of reading failure* (pp. 133–152). Baltimore: York Press.

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

Grigorenko, E. L., Wood, F. B., Meyer, M. S., Hart, L. A., Speed, W. C., Shuster, A., et al. (1997). Susceptibility loci for distinct components of developmental dyslexia on chromosomes 6 and 15. *American Journal of Medical Genetics (Neuropsychiatric Genetics), 60*, 27–39.

Grigorenko, E. L., Wood, F. B., Meyer, M. S., & Pauls, D. L. (2000). The chromosome 6p influences on different dyslexia-related cognitive processes: Further confirmation. *American Journal of Human Genetics, 66*, 715–723.

Grigorenko, E. L., Wood, F. B., Meyer, M. S., Pauls, J. E. D., Hart, L. A., & Pauls, D. L. (2001). Linkage studies suggest a possible locus for developmental dyslexia near

the Rh region on chromosome 1. *American Journal of Medical Genetics (Neuropsychiatric Genetics), 105*(1), 120–129.

Kaplan, B. J., Goodglass, H., & Weintraub, S. (1983). *Boston naming test.* Baltimore: Lippincott Williams & Wilkins.

Lindamood, C. H., & Lindamood, P. C. (1979). *Lindamood auditory conceptualization test.* Boston: Teaching Resources Corporation.

MacGinitie, R. K. (1978). *Gates-MacGinitie reading tests.* Chicago: Riverside Publishing Company.

McCardle, P., Scarborough, H. S., & Catts, H. W. (2001). Predicting, explaining, and preventing children's reading difficulties. *Learning Disabilities Research and Practice, 16*(4), 230–239.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 3*, 195–216.

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.

No Child Left Behind Act of 2001. (P.L. 107–110 [20 U.S.C. 7801].)

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

Scarborough, H. S. (1989). Prediction of reading disability from familial and individual differences. *Journal of Educational Psychology, 81*, 101–108.

Scarborough, H. (1998). Predicting the future achievement of second grades with reading disabilities: Contributions of phonemic awareness, verbal memory, rapid naming, and IQ. *Annals of Dyslexia, 68*, 115–136.

Sénéchal, M., LeFevre, J.-A., Thomas, E. M., & Daley, K. E. (1998). Differential effects of home literacy experiences on the development of oral and written language. *Reading Research Quarterly, 33*, 96–116.

Shaw, R., & Shaw, D. (2002). DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado state assessment program (CSAP) (Technical Report). Eugene, OR: University of Oregon.

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

Stanovich, K. E., Cunningham, A. E., & Cramer, B. B. (1984). Assessing phonological awareness in kindergarten children: Issues of task comparability. *Journal of Experimental Child Psychology, 38*, 175–190.

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.

Torgesen, J. K. (2004). Preventing early reading failure—and its devastating downward spiral. *American Educator, 28*(3), 6–19, 12–13, 17–19, & 45–47.

Wolf, M. (1991). Letter naming, reading and the contribution of the cognitive neurosciences. *Reading Research Quarterly*, 123–141.

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

Wood, F., & Grigorenko, E. (2001). Emerging issues in the genetics of dyslexia: A methodological preview. *Journal of Learning Disabilities, 34*, 503–511.

Woodcock, R. W., & Johnson M. G. (1977). *Woodcock Johnson psycho-educational battery.* Allen, TX: DLM Teaching Resources.

Woodcock, R. W., & Johnson, M. G. (1989). *Woodcock Johnson psycho-educational battery–revised.* Allen, TX: DLM Teaching Resources.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson tests of achievement–III.* Itasca, IL: Riverside Publishing Co.

Manuscript received June 21, 2005.
Final version accepted October 3, 2005.