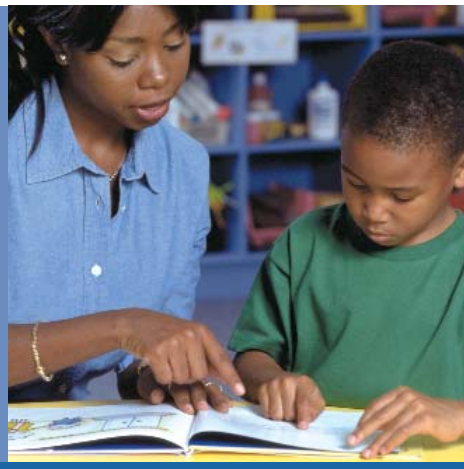
Longitudinal Study of the Effect of Universal Literacy

A Hierarchical Linear Modeling Analysis of Curriculum-Based Measurement Data



Greg Roberts, Ph.D.
Program Evaluation Consultant

E R S
Evaluation Research Services

Abstract

This technical report presents findings from multilevel growth modeling of longitudinal data for 16,443 students enrolled in Voyager's Universal Literacy System® at 291 schools across the U.S. The data were collected as part of the Vitals Indicators of Progress system (VIP), an alternate form of the curriculum-based Dynamic Indicators of Basic Early Literacy Skills (DIBELS). The Voyager program relies on VIP to monitor student progress against early literacy benchmarks, to identify struggling readers, and to evaluate program impact. Like DIBELS, the VIP system is fluency based and focuses on the "five big ideas" described in the National Reading Panel's report, *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction* (National Reading Panel, 2000).

Students in Voyager's first grade program made considerable progress over the course of the school year. Regardless of skill level at the school year's outset, children made notable gains in oral reading fluency, improving by an average of almost 30 words from November to May. Further, less skilled readers progressed at a somewhat faster rate, although initial performance was the primary indicator of end-of-year fluency rates. Children who attended Voyager kindergarten classrooms prior to entering first grade were more likely to be on track and less likely to be struggling early in first grade than children attending non-Voyager kindergarten classrooms.

Introduction

Twenty years of research in early reading instruction underscore the importance of assessing student progress at regular intervals throughout the school year (Deno, 1992; Shinn, 1998). Successful teachers systematically collect data on students' ability with important early literacy skills (Fuchs & Fuchs, 1999), and they use that data to plan and alter instruction, at the group level and for individual students (Deno, 1986; Fuchs, Fuchs, & Hamlett, 1989; Stoner, Scarpati, Phaneuf, & Hintze, 2002). For these teachers, decisions about what to teach, to whom, using what grouping strategy, etc. are based on evidence collected during these recurring assessment periods. Curriculum-based measurement offers a useful platform for this type of teacher activity.

Curriculum-based measurement is a time- and resource-efficient, classroom-oriented means of indexing academic standing at points in time and of quantifying progress over time (Deno, Fuchs, Marston, & Shin, 2001). Curriculum-based systems of measurement provide reliable and relevant data for estimating status and charting growth (Marston, 1989), for making instructional decisions and refining instructional plans (Marston, Mirkin, & Deno, 1984), and for evaluating program effectiveness (Deno, 1986; Stoner, Scarpati, Phaneuf, & Hintze, 2002). A considerable body of research supports the technical adequacy of curriculum-based measurement (e.g., Fuchs & Deno, 1991; Good & Jefferson, 1998; Markell & Deno, 1997).

This paper describes data collected using Voyager's Vital Indicators of Progress (VIP), a curriculum-based series of measures that address key elements of early reading proficiency, as outlined in the National Reading Panel's report, *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and its Implications for Reading Instruction* (National Reading Panel, 2000). VIP measures include: 1) initial sound fluency, 2) letter naming fluency, 3) phoneme segmentation fluency, 4) nonsense word fluency, and 5) oral reading fluency. VIP was developed by Dr. Roland Good and Dr. Ruth Kaminski as an alternate form of the Dynamic Indicators of Basic Early Literacy Skills (Kaminski & Good, 1996), which is widely used to monitor early reading progress within classroom settings. Dynamic Indicators of Basic Early Literacy Skills (DIBELS) and VIP are based on the ongoing program of research of Good, Kaminski, and their colleagues (Good, Simmons, & Kame´enui, 2001) and the benchmarks that serve as foundations of both systems are based on the most current work on early literacy (see website http://dibels.uoregon.edu).

Interpreting VIP Data

Voyager classroom teachers administer the VIP on four occasions during the school year, at baseline (Benchmark 1), in the fall (Benchmark 2), the winter (Benchmark 3), and the spring (Benchmark 4). The data are used to monitor student progress, to identify struggling readers, and to plan and modify classroom instruction. The data are also used to evaluate the impact of the Universal Literacy System (ULS), Voyager's core reading program. Student performance benchmarks are central to each of these uses. A benchmark is a criterion score that determines "on level performance," where "on level" is defined as performance associated with an increased likelihood of achieving subsequent literacy goals. Children who score at or above a given benchmark are likely (i.e., 85% to 90% of the time) to be at benchmark at the next assessment period and to be proficient readers by the end of third grade (see Good, Simmons, & Kame´enui, 2001). Table 1 presents VIP benchmarks for children in kindergarten and first grade. All of the measures are fluency-based so scores are interpreted as "correct responses (words, phonemes, letters, etc.) per minute."

Table 1.

TreeHouse: Kindergarten Benchmark Schedule				
	BENCHMARK 1	BENCHMARK 2	BENCHMARK 3	BENCHMARK 4
Timeline	**Semester 1: Wks 3–4**	**Semester 1: Wks 12–13**	**Semester 2: Wks 4–5**	**Semester 2: Wks 15–16**
Measure	Initial Sound Fluency Goal: 11 Letter Naming Fluency	Initial Sound Fluency Goal: 11 Letter Naming Fluency	Initial Sound Fluency Goal: 26 Phoneme Segmentation Fluency Goal: 35 Letter Naming Fluency	Phoneme Segmentation Fluency Goal: 35 Nonsense Word Fluency Goal: 40 Letter Naming Fluency
SeaCastles: First Grade Benchmark Schedule				
	BENCHMARK 1	BENCHMARK 2	BENCHMARK 3	BENCHMARK 4
Timeline	**Semester 1: Wks 3–4**	**Semester 1: Wks 12–13**	**Semester 2: Wks 4–5**	**Semester 2: Wks 15–16**
Measure	Phoneme Segmentation Fluency Goal: 35 Nonsense Word Fluency Goal: 40 Letter Naming Fluency	Phoneme Segmentation Fluency Goal: 35 Nonsense Word Fluency Goal: 40 Reading Connected Text Goal: 40	Phoneme Segmentation Fluency Goal: 35 Nonsense Word Fluency Goal: 40 Reading Connected Text Goal: 40	Phoneme Segmentation Fluency Goal: 35 Nonsense Word Fluency Goal: 40 Reading Connected Text Goal: 40

For instance, the first grade winter benchmark for nonsense word fluency (NWF) is 40, meaning that students who are "on track" to becoming proficient readers are able to read at least 40 nonsense words in the allotted minute. Students who fail to meet the benchmark are considered to be either "emerging" or "struggling" readers depending on the extent to which they missed the mark. Emerging readers are those for whom a clear prediction of subsequent performance is not possible. Approximately one-half of

the group of emerging readers at any given benchmark will go on to achieve on track status on subsequent VIP assessments (Good, Simmons, & Kame´enui, 2001). By contrast, only 20% of the group of struggling readers are likely to achieve subsequent benchmark goals without additional intervention. Each category of reader suggests the intensity of instruction required to bring its members to benchmark. Struggling readers will require intensive intervention to succeed at subsequent assessment periods. Emerging readers will require strategic help to move to on track status. Benchmark readers will require regular quality instruction.

Evaluation Questions

The analyses described in this report use two years worth of VIP data to estimate the learning trajectories of first grade children who have participated in the Voyager Universal Literacy System. Treatment outcomes and learning trends are compared to criteria (the aforementioned "benchmarks") derived from data on a very large cross section of U.S. students (see Methods section for a description or participants).

Student performance is considered in terms of differing exposures to treatment, as well. Outcomes and rates of change for students participating in Voyager classrooms during both their kindergarten and first grade school years are compared to the learning trends for children attending Voyager during first grade only (i.e., attended non-Voyager kindergarten). This analysis, using levels of treatment dosage as the grouping variable, has been characterized as a *dose-response function* design by Lipsey (1990), and qualifies as quasi-experimental because of its two-group structure and the effect of multilevel modeling of pretreatment status.

Methods

Participants

The data used in this study were collected by teachers in first-grade Voyager classrooms during the 2001-2002 and 2002-2003 school years. There are 16,443 cases with at least a single VIP data point in the original data set with 291 schools represented. Across all schools, 82% of children qualified for free/reduced lunch and 30% qualified as limited English proficiency. Ninety-six percent of schools had at least 50% of students receiving free or reduced lunch, and 23% of schools reported at least 50% of students with limited English. Student-level demographic data were not available.

Measures

VIP Oral Reading Fluency (ORF) is a standardized, individually administered test of accuracy and fluency with connected text. Oral reading fluency is a widely used curriculum-based measure of reading competence and reading comprehension (Good, Simmons & Kame´enui, 2001). Generally operationalized as the number of words read correctly in one minute (Fuchs & Fuchs, 1999), oral reading fluency is a reliable and remarkably efficient predictor of elementary-school students' scores on more traditional (and more-lengthy) measures of reading ability (Fuchs & Deno, 1991; Fuchs, Fuchs, Hosp, & Jenkins, 2001) and reading comprehension (Fuchs, Fuchs, & Maxwell, 1988; Jenkins, Fuchs, Espin, van den Broek, & Deno, 2000), with bivariate correlations generally in the .6 to .9 range (Good & Jefferson, 1998).

Information processing models provide a framework for understanding oral reading fluency and its relationship to reading comprehension. The prevailing theory is that both decoding and comprehension require cognitive resources, and the more processing capacity is devoted to decoding, the less capacity is available for understanding what has been decoded (LaBerge & Samuels, 1974; Stanovich, 2000). Recent extensions (Perfetti, 1995) of this view suggest that comprehension-related tasks may be subject to varying levels of automaticity, much like word recognition and decoding tasks. Inferential processes, for instance, may become increasingly automatic, freeing up additional cognitive resources for even deeper levels of text processing (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Logan, 1997), suggesting that oral reading fluency may be more than a proxy for reading competence and comprehension; instead, fluency may relate to reading ability in more substantive ways (Nathan & Stanovich, 1991).

Performance scores on measures of oral reading fluency are widely used for intra- and inter-individual comparisons (Fuchs, Fuchs, Hosp, & Jenkins, 2001), and they may represent a more sensitive indicator of individual and group change than more-broadly conceptualized (i.e., the commercially-available measures of reading) measures of reading comprehension (Marston, Fuchs, & Deno, 1985), underscoring their utility for detecting even small changes in student standing and for documenting

relatively small-sized program effects. There is also evidence that oral reading fluency scores may reliably forecast performance on high-stakes, state-mandated tests of reading ability (Crawford, Tindal, & Stieber, 2001; Buck & Torgesen, 2003), although this area of research continues to develop.

The VIP ORF reading passages are calibrated by grade level, with first grade passages set at a Spache readability level of 2.0 and second grade passages at 2.6. Words omitted, substituted, and hesitations of more than three seconds are scored as errors. Words self-corrected within three seconds are scored as accurate. The number of correct words read per minute is the oral reading fluency rate. ORF is administered for the first time at the second benchmark in the first grade. For purposes of this analysis, all three first grade ORF assessments were used (i.e., middle 1, middle 2, end).

Plan for Analysis

Growth curve modeling is well suited to analyzing time series data. Conceptualizing measurement times as "nested within" individuals resolves many of the limitations of traditional regression models and the HLM for Windows software (Bryk & Raudenbush, 1992) is a powerful tool for analyzing multilevel data of this sort. HLM provides more reliable parameter estimates for multilevel, or hierarchical, data than do other regression models. It yields growth curve coefficients at the individual level and for subgroups or persons within a data set. Further, HLM can "explain" variation in individual curves in terms of person-level variables and, in three-level models, in terms of group characteristics.

A recommended first step when using HLM is to estimate a model that does not include "explanatory" variables (Bryk & Raudenbush, 1992). This *unconditional model* represents growth averaged over all children and schools and represents a baseline against which to evaluate subsequent, more sophisticated models (i.e., *conditional models*). The unconditional model also provides data that are useful for partitioning variability of the person-level parameters into level-2 and level-3 components, estimating correlation among growth parameters, and estimating reliability of effects at each level. In this context, Level 1 parameters correspond to within person effects, Level 2 parameters correspond to between person effects.

The conditional model represents an extension of the unconditional model by including conceptually meaningful variables to explain level-1 and/or level-2 variability. Within these two general classes of models, three subtypes are generally fit, including an unrestricted model, a model with Level-1 homogenous variance, and a model with Level-1 heterogenous variance. The model subtype with the best ratio of total deviance to total estimated parameters is generally accepted as the best fitting model. For this analysis, a multi-level model was assembled according to the questions of interest, with the emphasis on simplicity and clarity of interpretation.

A 3-level unconditional growth model was fit to the full first grade data set (i.e., across all cases). The model was estimated with performance for a given child at a given VIP administration modeled in terms of time at level 1, within school variance modeled at level 2, and between school variance at level 3. The conditional model included two level-2 factors, one representing the type of kindergarten program attended (Voyager or another program) and the second indicating students' benchmark status at Benchmark 1 in September of first grade (based on the first benchmark assessment which does not include a measure of oral reading fluency), either struggling, emerging, or on track (*kindergarten attended and performance at first benchmark*). The final model reflected the relative effects of Voyager instruction for children beginning first grade at different skill levels. It also made possible examination of the Voyager kindergarten program's potential effect on students' initial status.

Results

Initial inspection of the ORF data indicated that scores were positively skewed at all three time points. While skewed data do not bias parameter estimates, they can affect estimation of standard errors, confounding interpretation of parametric tests (Bryk & Raudenbuch, 1992). For this reason, data were transformed into natural logarithms, and tests were conducted on the transformed data. For purposes of reporting parameter estimates, natural logs were back transformed to raw score units (word read correctly per minute). General findings about the unconditional model are presented in raw score units. Findings related to the conditional model are presented in terms of raw-score units, as well.

Unconditional Model

The unconditional restricted model converged in 8 iterations, an informal indication of good model fit (Bryk & Raudenbush, 1992). The model with homogenous level-1 variance provided a significantly *less* adequate fit than the unrestricted model $\{\Delta\chi^2\ (1) = 1498.38, p < .001\}$, while the fit (and the parameter estimates) provided by the heterogenous model did not differ significantly $\{\Delta\chi^2\ (1) = .00008, p > .5\}$ from the unrestricted model. All fixed effects estimates for the heterogenous model differed significantly from 0 (t = 84.6 for intercept and t = 35.2 for slope) with p values of less than .001.

For all children across all schools, the average words read per minute at Benchmark 2 in November of first grade was 17.4, and the average learning rate was 12.42 additional words per benchmark. The average oral fluency rate at the end of the first grade school year was 42.24. Also of note is the ratio of between school variance to total variance; 26% of all variance in initial status was between-schools. For learning rate, 36% of the total variance was due to differences between schools, suggesting that in some cases two schools may have had the similar average ORF score at time 1, but experienced different learning rates during the first grade year, due possibly to differences in fidelity of program implementation. Finally, the within-school correlation of initial status and learning rate was -.77, meaning that students with lower ORF scores at Benchmark 2 in November of first grade tended to gain at a considerably faster rate than children with higher Benchmark 2 ORF scores.

Conditional Model

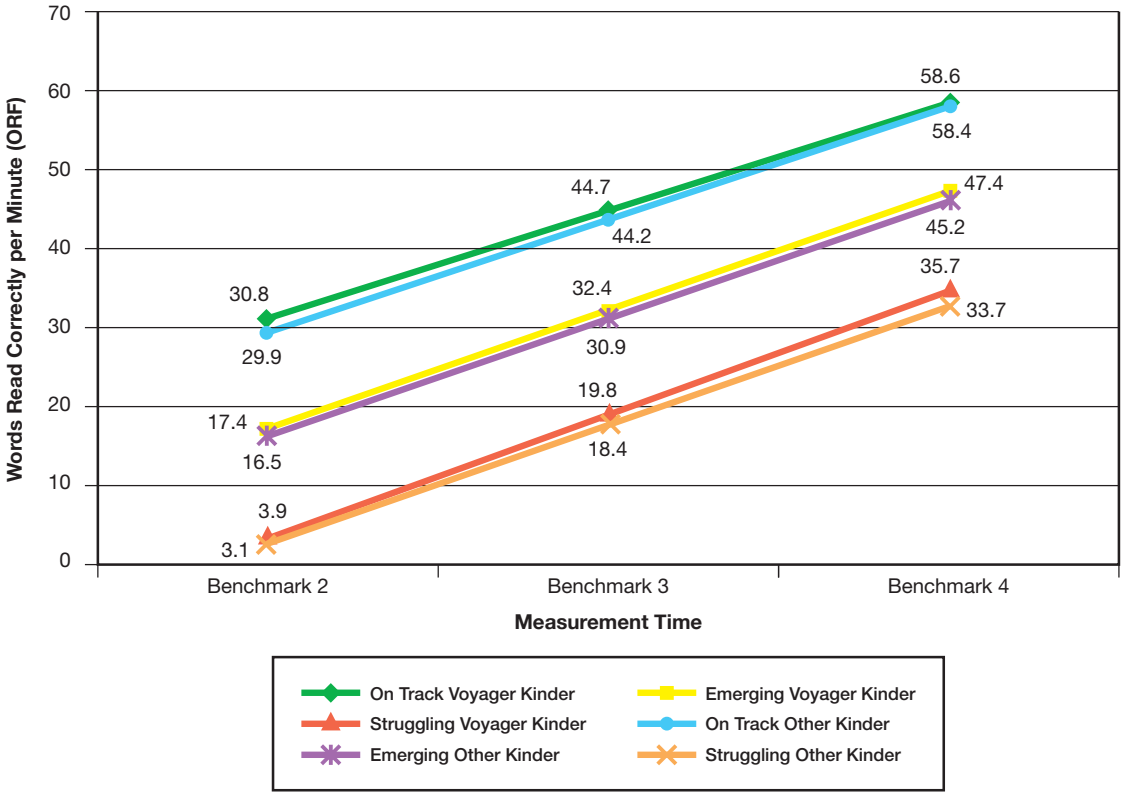
The *conditional* model with two level-two predictors was fit by specifying level-1 variance as homogenous; models with heterogeneous level-1 variance failed to converge. The level-2 predictors together (*kindergarten attended and performance at first benchmark*) accounted for 52% of the variance in initial status (i.e., intercept) and 8.3% of the variation in rate of learning (slope). Variance explained statistics in hierarchical analysis provide the strongest evidence for evaluating the importance of level-2 predictors (Bryk & Raudenbush, 1992). There is no consensus regarding the amount of variance-

explained required to deem a predictor "important," but using 5% of total variance as a guide, the level-2 predictors together appear to add meaningfully to explaining variation in individual growth curves. Considered separately, however, *kindergarten attended* did not contribute significantly to the prediction of students' learning rates {t (16,443) = 1.61, p =.107}, though it did contribute to the prediction of initial status {t (16,443) = 6.80, p < .001). *Performance at first benchmark* also contributed to the explanation of differences in initial status {t (290) = 72.50, p < .001), as well as to the prediction of learning rate {t (290) = - 38.18, p <.001}. Overall mean values were significant for both initial status {t (290) = 130.90, p < .001} and learning rate {t (290) = 70.60, p < .001}.

The trend lines in Figure 1 represent ORF results for the conditional model. Six groups are included, reflecting the possible combinations of level-2 predictors. As suggested, students' initial status as indicated by *overall performance at first benchmark* is predictive of their later performance. Students on track at the outset of first grade read an average of from 30 to 31 words per minute at Benchmark 2 in November and progressed at rate of about 14 additional words per benchmark, ending the school year with a fluency rate of about 58.5 words per minute. Emerging readers at Benchmark 2 improved by about 14.5 to 15 words per benchmark, and ended the year reading between 45.2 and 47.4. Struggling readers read fewer than 4 words correctly in a minute in November. This group gained about 15.5 words per benchmark to finish the year reading between 33.7 and 35.7 words per minute, on average.

It is also helpful to consider the results in terms of expected benchmark performance. The end-of-first-grade benchmark for oral reading fluency is 40 words per minute; this is the level of performance that is predictive of subsequent achievement. In this sample, the end-of-year average for the group of emerging readers exceeds this threshold, while the average performance level for strugglers is within .33 standard deviations of the benchmark (the standard deviation in time 3 scores for the struggling group is approximately 18 words).

Figure 1: Conditional Model



The difference in rates of change is statistically significant and potentially meaningful (based on the 8.3% variance-explained estimate) and translates into about a 3 word per year advantage in oral fluency for strugglers over on track readers. The most salient predictor of progress, however, is students' status at the year's outset (i.e., *performance at first benchmark*). While struggling and emerging readers made progress, and to an extent more rapid progress, being on track initially represents a considerable advantage. Given that, a reasonable next question is, "What factors contribute to being on track at the outset of first grade?"

Student-level demographic data would be necessary to fully consider this question, but a preliminary investigation is possible. Specifically, it is worth asking if participation in the Voyager kindergarten program was associated with initial status. Table 2 displays students' *performance at first benchmark in terms of kindergarten attended*. The number of children in each category is shown, along with column percentage (within *performance at first benchmark*). The trend suggests that kindergarten participation in the Voyager program is associated with more favorable benchmark status at the outset of first grade {Pearson χ^2 (2) = 1,013.5, $p < .001$}. Approximately 69% percent (69.3%) of Voyager kindergarten children were on track compared to 31.7% of children attending kindergarten programs other than Voyager. Also of note is the relative percentage of struggling readers at first benchmark. In the Voyager kindergarten group, only 3.8% were struggling to read at the outset of first grade. By comparison, 27% of children attending kindergarten programs other than Voyager were struggling.

Table 2: Performance at First Benchmark by Kindergarten Attended

	Struggling	Emerging	On Track
Other Kindergarten	N = 1,279 27%	N = 1,955 41.3%	N = 1,499 31.7%
Voyager Kindergarten	N = 86 3.8%	N = 619 27%	N = 1,588 69.3%
Total	1,365	2,574	3,087

It is important to consider this finding in terms of the sample on which it is based. The data used for this analysis are generally similar in terms of social indicators (percent receiving free/reduced lunch, etc.), meaning that the group of high scorers may be more similar to the low scorers on traditional achievement predictors. Future analyses will include person- and school-level demographic data, making possible more precise estimates of effects at level-2 and level-3 of the model and allowing for more comprehensive examination of the reasons for differences in students' learning trends over time.

Discussion

Internal validity is reflected in the question "how would the treatment group have fared if it hadn't participated in the intervention?" The extent to which this question can be reliably answered represents the degree to which one can infer that a treatment caused an observed outcome. In these data, the kindergarten comparison can be characterized as having reasonably high internal validity. There is no pretest for comparing status of comparison and intervention groups prior to treatment and there is no measure of implementation fidelity. But given the very large numbers in both conditions and the fact that cases are distributed across a number of schools and communities, it is more likely than not that the samples represent the respective populations of children. The groups are probably comparable at the outset and probably have comparable experiences during their kindergarten school years, *with the exception of the program in which they participated*, supporting the argument that Voyager program contributes significantly to the sizable group differences in the first grade benchmark 1 data.

In the first grade data, the question "how would the treatment group have fared if it hadn't participated in the intervention?" is less salient due to the absence of comparison data. One can, however, address the question "how did children do in relation to relevant criteria?" by considering the end of first grade results in terms of the VIP benchmarks. This is an important question, and the answer is that most Voyager participants achieve benchmark status. "Would first graders have performed similarly in programs other than Voyager?" is a next question; future analyses of the VIP dataset will include comparative data across first grade. Similarly, questions related to the sustainability of effects will be addressed in subsequent studies. This question will require comparison data across two years for a group of children participating in Voyager during kindergarten and first grade and for a similar group of children participating in kindergarten and first grade programs other than Voyager. In the absence of such, arguments related to sustainability are speculative.

Another potential threat concerns the collection of data. VIP is administered by classroom teachers, and while there is evidence of acceptable inter-rater reliability (see Roberts, 2002) in several of the schools represented in this dataset, a sample-wide estimate of reliability is presently unavailable. Finally, a number of quality curricula are undoubtedly included, a reality that should be considered when interpreting these results. The program effects are not in relation to "no treatment," but to alternative, and in some cases strong, treatment.

VIP data collection is ongoing. A next wave of analysis is scheduled for summer of 2004. At that time, second grade ORF data will be available. A far greater number of level-2 and level-3 "units" in each of the two conditions will be available, and student- and school-level data will be more accessible.

References

Bryk, A. & Raudenbusch, S. (1992). *Hierarchical Linear Models*. Newbury Park: Sage Publications.

Buck, J., & Torgesen, J. (2003). *The relationship between performance and on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test: FCCR Technical Report #1*. Florida Center for Reading Research, Tallahassee, FL.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment*, 7(4), 303-323.

Deno, S., Mirkin, P., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children*, 49, 35-45.

Deno, S. (1986). Formative evaluation of individual student programs: A new role for school psychologists. *School Psychology Review*, 15, 358-374.

Deno, S., (1992). The nature and development of curriculum-based measurement. *Preventing School Failure*, 36(2), 5-10.

Deno, S.,Fuchs, L., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review*, 30(4), 507-524.

Fewster, S., & MacMillan, P. (2002). School-based evidence for the validity of curriculum-based measurement of reading and writing. *Remedial and Special Education*, 23, 149-156.

Fuchs, L.,& Deno, S, (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, 57, 488-501.

Fuchs, L. & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review*, 28, 659-671.

Fuchs, L., Fuchs,D., & Maxwell,L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education*, 9(2), 20-28

Fuchs, L., Fuchs, D., Hosp, M., & Jenkins, J.(2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5, 239-256.

Fuchs, L. Fuchs, D. & Hamlett, C. (1989). Effects of instrumental use of curriculum-based measurement to enhance instructional programs. *Remedial and Special Education*, 10, 43-52.

Good, R. & Jefferson, G. (1998). Contemporary perspectives on curriculum-based measurement validity. In M.R. Shinn (Ed.), Advanced applications of curriculum-based measurements (pp. 61-88). New York: Guilford Press.

Good, R., Simmons, D., & Kame´enui, E. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, 5(3), 257-288.

Good, R. H., Simmons, D. C., & Kame´enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, 5, 257-288.

Hamilton, C. (2001). The "word caller" phenomenon: Teachers' judgments on reading comprehension and fluency. *Dissertation Abstracts International*, 62/06, 2027. (UMI No. 3018366) This is now published in SPR.

Jenkins, J., Fuchs, L., Espin, C., van den Broek, P. & Deno, S. (2000). *Effects of task format and performance dimension on word reading measures: Criterion validity, sensitivity to impairment, and context facilitation.* Paper presented at Pacific Coast Research Conference, San Diego, CA.

Kaminski, R. & Good, R.(1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review*, 25, 215-227.

LaBerge, D. & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293-323.

Lipsey, Mark W. (1990). *Design Sensitivity: Statistical Power for Experimental Research.* Thousand Oaks, CA: Sage Publications.

Logan, G. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading and Writing Quarterly*, 13, 123-146.

Markell, M. & Deno, S. (1997). Effects of increasing oral reading: Generalization across reading tasks. *The Journal of Special Education*, 31(2), 233-250.

Marston, D., Fuchs, L. & Deno, S., (1985). Measuring pupil progress: A comparison of standardized achievement tests and curriculum-related measures. *Diagnostique*, 11(2), 77-90.

Marston D., & Mirkin, P., & Deno, D. (1984). Curriculum-based measurement: An alternative to traditional screening, referral, and identification. *Journal of Special Education*, 18, 109-117.

Marston, D., (1989). Curriculum-based measurement: What is it and why do it? In M.R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford.

Nathan, R., & Stanovich, K., (1991). The causes and consequences of differences in reading fluency. *Theory Into Practice*, 30, 176-184.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Perfetti, C., (1995). Cognitive research can inform reading education. *Journal of Research in Reading*, 18(2), 106-115.

Roberts, Greg. (2002). *Technical Evaluation Report on the Impact of the Voyager Universal Literacy System 2001-2002.*

Shinn, M., (1998). Advanced applications of curriculum-based measurement. New York: Guilford Press.

Shinn, M., Good, R., Knutson, N., Tilly, W., & Collins, V. (1992). Curriculum-based measurement reading fluency: A confirmatory factor analysis of its relation to reading. *School Psychology Review*, 21, 459-479.

Shraw, G., & Bruning, R. (1999). How implicit models of reading affect motivation to read and reading engagement. *Scientific Studies of Reading*, 3, 281-302.

Stanovich, K. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York:Guilford.

Stoner, G., Scarpati, S., Phaneuf, R. & Hintze, J. (2002). Using curriculum-based measurement to evaluate intervention efficacy. *Child & Family Behavior Therapy*, 24(1/2), 101-112.