

The NRP Comparison of Whole Language and Phonics: Ignoring the Crucial Variable in Reading

Stephen Krashen

Editors' Note: In 1997, Congress asked the Director of the National Institute of Child Health and Human Development (NICHD) to form a national panel to assess the effectiveness of different approaches used to teach children to read. For over two years, the National Reading Panel (NRP) reviewed research-based knowledge on reading instruction. On April 13, 2000, the NRP concluded its work and submitted the Report of the National Reading Panel: Teaching Children to Read. Part of the charge of this panel was to disseminate their findings to parents, teachers, administrators, and "anyone else interested in learning about reading research."

As Garan (2001) states, the NRP was fraught with controversy from its inception. First, panel members were primarily researchers whose work was experimental in nature, and the panel limited its review to "studies that were experimental or quasi-experimental in design." Therefore, qualitative research was disregarded. The NRP's database narrowed the original 100,000 studies to 1,373 phonics studies, which were further reduced to 38 (Garan, 2002, personal communication). Second, the professional backgrounds of the fourteen panel members raised many questions (for example, one was a physicist, several were educational psychologists, and one was a certified public accountant for a law firm).

The panel focused on several areas within reading: alphabets, including phonemic awareness and phonics instruction; fluency; comprehension, including vocabulary instruction, text comprehension instruction, and teacher preparation and comprehension strategies; teacher education and reading instruction; and computer technology and reading instruction. The phonics instruction section looked at 38 studies that were done since 1966, resulting in 66

comparisons of "skills-based" approaches and "whole language" approaches, as defined by the panel.

In this article, Krashen examines the results of the NRP's comparison of skills-based and whole language approaches through the lens of reading comprehension. His findings reveal that even when one accepts the restrictions on what is acceptable research imposed by the panel, when one considers the actual amount of reading done by children and examines the results for tests of reading comprehension, the research does not show that skills-based methods are superior.

In 2000, the National Reading Panel (NRP) completed its assessment of experimental and quasi-experimental research in reading and the effectiveness of various approaches to teaching reading, and published its findings in the *Report of the National Reading Panel: Teaching Children to Read*. The panel concluded that "skills"-based approaches are superior to whole language approaches in helping children learn to read. In this article, I argue that when one considers tests of reading comprehension and the amount of real reading done, this claim is not substantiated. To show this difference, I reanalyze studies used in the NRP's final report, as well as several that were inappropriately excluded, and interpret the results based upon calculations of effect sizes.

In its report, the NRP summarized the effect of each study using effect sizes. Effect sizes reflect the impact of a treatment and are calculated in most cases by subtracting the mean of the comparison group from the mean of the experimental group and dividing this result by the pooled standard deviation. All calculations in the NRP report, as well as additional calculations done in this paper, were done with DSTAT software (Johnston, 1993).

Table 1 shows the NRP's findings comparing phonics and whole language. In their analysis, effect sizes were not analyzed separately for each kind of measure, but represent the average effect size for all reading tests. A positive effect size indicates an advantage for phonics. Overall, the NRP found a positive effect size of .32 in favor of phonics approaches to teaching reading.

I re-analyze these studies here, presenting effect sizes for tests of reading comprehension. Reading comprehension is, after all, the goal of reading instruction. Of major importance in examining reading instruction is considering how much reading children actually do. The well-known hypothesis that we "learn to read by reading" (Smith, 1994; Goodman, 1982) predicts that this variable will be central. Real reading is, in addition, the central element of the whole language approach; the core of whole language is providing children with interesting texts and helping them understand these texts using all cueing systems simultaneously. Unfortunately, in most studies, it was impossible to determine which group did more real reading. In these cases, I accept the labels of whole language and phonics as provided by the NRP. At the end of this paper, I provide a separate analysis of those studies in which it was clear that children in one group were reading more than children in the other, as well as an analysis of all studies.

In Krashen (1999), I presented a narrative review of studies claiming to compare whole language and "skills" approaches. I concluded that when whole language was

defined as including a great deal of real reading, students in these classes performed as well as or better than children in skills classes on tests of reading comprehension, were equivalent on tests of skills (e.g., reading nonsense words), had more positive attitudes toward reading, and read more on their own. I focus here only on tests of reading comprehension, providing effect-size calculations where possible, and compare my results with the National Reading Panel's conclusions.

Reanalysis and Commentary on Studies Included by the National Reading Panel

Based upon my own reanalysis, I comment on the studies included in the NRP's report. I did not reanalyze Freppon's (1991) study because the study did not include a test of reading comprehension.

Foorman et al. Study

In Table 1, there are four comparisons from Foorman, Francis, Fletcher, Shatschneider, and Mehta (1998). Contrary to what was indicated in Table 1, Foorman et al. did not present a separate analysis for grades 1 and 2 or for at-risk/low-achieving children, at least not in their 1998 paper. The "direct code" group was classified as systematic phonics, the "embedded code" group as "blending large units" (i.e., focused on whole-word and syllable), and the "implicit code group" as whole language. All three methods

Study	Type of Phonics	Students	Effect Size
Foorman et al., 1998	Systematic	Grade 1, at risk	0.91
Foorman et al., 1998	Systematic	Grade 2, low achieving	0.12
Foorman et al., 1998	Blending LU	Grade 1, at risk	0.36
Foorman et al., 1998	Blending LU	Grade 2, low achieving	0.03
Evans and Carr, 1985	Misc phonics	Grade 1	0.6
Freppon, 1991	Misc phonics	Grade 1	0
Traweek & Berninger, 1997	Systematic	Grade 1, at risk	0.07
Wilson & Norman, 1998	Systematic	Grade 2	-0.47
Santa & Høien, 1999	Blending LU	Grade 1, at risk	0.76
Klesius et al., 1991	Misc phonics	Grade 1	0.2
Griffith et al., 1992	Misc phonics	Grade 1	-0.33
Stuart, 1999	Systematic	K, at risk	0.73

Overall mean effect size = .32 in favor of phonics

Blending LU = blending large units; a positive number indicates an advantage for phonics

Table 1. National Reading Panel Results: Phonics vs. Whole Language

"existed within a literature-rich environment in the classroom" (p. 39), but it was not clear which group did the most real reading.

Results were reported for both the regular implicit code group and a special "research implementation" of implicit code. In Table 2, I present my effect-size calculations for the passage comprehension test and the "formal reading inventory" (p. 41). Both regular and "research implementation implicit code" groups are presented; the regular group effect sizes are in parentheses.

There were problems with the measure used. First, the passage comprehension test consisted only of "a cloze test at the sentence level" (p. 41); second, the formal reading inventory was "too difficult for these children" (p. 41). In fact, many children did not take the reading inventory because it was too difficult.

Evans and Carr Study

In the Evans and Carr (1985) study, the "traditional" group actually did significantly more silent reading than the whole language group (14.6 minutes per day versus 9.88 minutes per day) and emphasized "contextual meaning" significantly more (p. 333). The effect-size sign should thus be changed from plus to minus, in favor of the group doing more reading.

Results for reading comprehension were reported separately depending on the difficulty of the passage used on the test. I calculated the following effect sizes:

Primer passage =	-0.76
Grade 1 passage =	-0.8
Grade 2 passage =	-0.24
Mean =	-0.6

Freppon Study

The study by Freppon (1991) did not include a test of reading comprehension and is therefore not included in my reanalysis.

Traweek and Berninger Study

Traweek and Berninger's (1997) study did not include a test of reading comprehension. The effect size reported by the panel appears to be for a test of word reading.

Wilson and Norman Study

The panel reported an overall effect size of -.47 in favor of whole language for Wilson and Norman's (1998) study. My calculations from three different measures of reading comprehension are:

Passage comprehension (based on reading text aloud):	-.32
Cloze (supply missing word):	-.78
Passage comprehension (questions based on cloze passage):	-.45
Mean =	-.52

Santa and Hoiem Study

Santa and Hoiem's (1999) study resulted in an effect size of 1.0 for reading comprehension in a test given immediately after the treatment, at the end of the school year, and 1.8 for a reading comprehension test given the next fall, in favor of phonics. This difference was due only to the performance of their "high risk" group, with a sample size of 13 (experimental) and 12 (control); only 12 experimentals and 9 controls took the delayed posttest. The entire sample was taken from the bottom 20% of performers in each class that was involved. Thus the effect was limited to the lowest 10%, the bottom half of the lowest 20% of a "lower middle class" group of students. The reading comprehension measure was a cloze test, with a context of only two to three sentences for each missing word. Also, it was not clear which group did more real reading; the whole language group appeared to devote more time to reading, but this included some strange practices, including students chorally reading the same page together, something called "mumble reading."

	Passage Comprehension	Formal Reading Inventory	Mean ^a
Systematic versus whole language	.31 (.54)	.03 (-.15)	0.17
Blending LU versus whole language	-.04 (.19)	-.08 (-.29)	-0.06

^a Research group only

Table 2. Foorman et al. Study

Klesius, Griffith, and Zielonka Study

Klesius, Griffith, and Zielonka (1991) reported little difference between skills-based and whole language groups for the reading comprehension section of the Comprehensive Test of Basic Skills (CTBS) ($d = .05$). This difference, however, could be $-.05$ because Klesius et al. did not indicate which group did better. McQuillan (1998) points out another problem: The whole language program was new, and only one day of inservice was provided. In addition, "two of the three whole language teachers needed additional assistance" (p. 51). No measure was made of how much reading was done in either program.

Griffith, Klesius, and Kromrey Study

Griffith, Klesius, and Kromrey (1992) presented results for children with high and low phonemic awareness (PA) separately. On the CTBS comprehension test, I calculated an effect size of $.71$ for low PA students (favoring traditional students), and -1.63 for high PA students (favoring whole language). The mean was thus $-.46$, but there were only six students in each group.

Stuart Study

Stuart (1999) made a comparison between Jolly Phonics and Big Books for kindergarten children, most of whom spoke English as a second language. Big Books involved the use of large size books, but it is not clear that much more real reading was included in this program, or that there was in fact a great deal of focus on meaning. Here is the description of the advice given to Big Books teachers:

Teachers were asked to spend time on word level work, that is, to emphasize words and letters, by drawing children's attention to written words in the text, and talking about the letters in words. Work with letters should involve introduction to their names and sounds, and children should be encouraged to notice and learn words and letters in the classroom environment (590).

I calculated effect sizes favoring Jolly Phonics of $.37$ and $.26$ (using F ratio) for the test of reading comprehension, given one year after the intervention ended. This study is included in my final analysis (Table 7), even though children in Big Books apparently had quite a bit of skills instruction, and it is not clear how much real reading was done.

Omissions

The panel omitted the following studies from their final analysis. A closer look at these studies is nevertheless merited.

Eldridge Study

The panel did not classify the Eldridge (1991) study as one involving whole language. The "modified" whole language group in this study had 15 minutes per day of phonics instruction, but "most of the classroom time was spent in recreational and functional reading and writing activities" (p. 32), and the whole language group clearly did more real reading. The effect size for reading comprehension was $-.81$ in favor of a whole language approach.

Hagerty, Hiebert, and Owens Study

Hagerty, Hiebert, and Owens (1989) compared the reading achievement of students in a "literature-based" program with the reading achievement of students in a "skills-based" program. The study compared second-, fourth-, and sixth-grade classrooms. The literature-based program included reading trade books and writing on topics chosen by students, while the skills-based program consisted of teacher-directed instruction and "filling out teacher-assigned worksheets that provided practice on particular skills or reading assigned textbook passages" (p. 455). Some free reading was included. From the description, it was clear that the literature-based classes did a lot more real reading.

Table 3 shows results of the Gates-MacGinitie Reading Comprehension Test, administered at the beginning and end of the academic year. The scores presented are residual scores; they have a mean of 0 and a standard deviation of 1.0. Two classes were utilized at each grade level. As Hagerty et al. note, one literature-based class showed a slight decline (negative residual score), but overall the literature-based classes clearly did better than the skills-based classes.

Grade	2	4	6
Literature-based	-.20, .35	.12, .04	.37, .22
Skills	.11, -.35	-.45, -.36	.09, -.03

Table 3. Hagerty, Hiebert, and Owens Study

Morrow Study

Morrow's (1992) study examined the effects of two reading approaches on second-grade students. Second graders in the literature-based group spent about 3.5 hours per week with basals and about four hours with literature. They were read to daily, engaged in at least three "literacy activities" per week (e.g., retelling and rewriting stories, book sharing), and had at least three 30-minute sessions per week in a "literacy center," during which time they read, wrote, and performed stories. Students in the basal program were read to no more than twice a week, and instruction focused nearly entirely on the basal and workbook. Free reading was allowed only when children had finished their seatwork.

I calculated effect sizes for performance on the reading subset of the California Achievement Test (CAT), and a probed comprehension test, based on comprehension questions asked the child by a researcher after the child read a passage. Groups 1 and 2 were identical, except that Group 1 also had a parental involvement component. The results show that children in the literature-based program performed better on measures of reading comprehension.

Knapp and Associates Study

Knapp and Associates' (1995) two-year study examined 66 classrooms "serving large numbers of children from low-income families" (p. xi). At the start of the study, the children were in grades one through five. Researchers analyzed each class according to its orientation to meaning; that is, differences were determined based upon the degree to which reading and writing were integrated, the extent of discussion in light of what was read, and (most relevant to this analysis) how much time was devoted to reading. In "high-meaning" emphasis classes ($n = 16$), children read an average of 48 minutes per day; in moderate-emphasis classes ($n = 29$), they read 18 minutes per day, and in low-meaning emphasis classes ($n = 22$), only 5 minutes per day. No mention was made as to how much of the reading was self-

	Group 1 vs. Control	Group 2 vs. Control
Probed Comprehension	-2.25	-1.84
CAT Reading	-0.66	-0.62
Mean		-1.23

Table 4. Morrow Study: Effect Sizes on a Probed Comprehension Test and the Reading Subset of the CAT

	Year 1	Year 2
High emphasis on meaning vs. low emphasis	0.26	0.06
Moderate emphasis on meaning vs. low emphasis	0.19	0.19

Table 5. Knapp and Associates Study: Differences in CTBS Reading Comprehension after One Academic Year

selected. The CTBS Reading Comprehension Test was used to measure achievement.

As shown in Table 5, high- and moderate-meaning groups did better than the low-meaning group in terms of reading achievement. In year 2, however, the effect size for the moderate-meaning emphasis group was larger than the effect size for the high-meaning emphasis group. Effect sizes were computed from NCE (normal curve equivalent) scores, controlling for initial level of reading achievement, poverty, teachers' background in language arts, and teachers' satisfaction with teaching.

Conclusions and Summary

Table 6 presents effect sizes for those studies ($n = 5$) in which the amount of reading done by children in one group was clearly more than the amount read by the other. In all cases those who read more outperformed those who read less (Hagerty et. al. was not included in Table 6, as it was not clear how to calculate effect sizes). The mean effect size was -.70, but this set is clearly not homogeneous. Note that in the study showing the smallest effect size, Knapp and Associates, it was not clear if all the reading was self-selected.

The conclusions parallel those in Krashen (1999) and are consistent with Smith's (1994) and Goodman's

	Effect Size
Evans and Carr, 1985	-0.6
Eldridge, 1991	-0.81
Morrow, 1992*	-1.23
Knapp and Associates, 1996	-0.17

*Group 2 results only
(Negative sign = group that read more was superior)

Table 6. Studies in Which One Group Read More than the Other: Tests of Reading Comprehension

(1982) original claims that we "learn to read by reading." They are also consistent with the results of sustained silent reading studies done with older readers (Krashen, 2001).

Table 7 presents all the studies reviewed here for which calculation of effect sizes was possible. The mean (unweighted) effect size for all studies is $-.17$, in favor of whole language, defined as the group that did more real reading or, when this classification was not possible, as the group labeled whole language. This result contrasts with the NRP's result of a $.32$ effect size in favor of phonics. Thus, even if we allow all studies into the analysis, including those in which it was not clear that one group did more reading, the overall results do not favor skills.

The studies from the NRP report that I have discussed in this article were not undertaken with what I consider to be the crucial variable in mind: the amount of genuinely interesting, real reading that children did. Thus, my conclusions are post-hoc and are only suggestive. What is clear, however, is that the National Reading Panel's interpretation of the results is not the only possible one and should not be the basis for national policy. ¹●

Note

1. One could argue that the analysis in Table 7 is biased against skills, because both reading comprehension measures and both conditions from Foorman et al. were included. If we only include the systematic phonics group and the passage comprehension test ($d = .31$), the overall advantage for whole language drops, but only from $d = -.17$

	Effect Size
Evans and Carr, 1985	-0.6
Foorman et al., 1998	0.17
Foorman et al, 1998	-0.06
Wilson & Norman, 1998	-0.52
Santa & Hoiem, 1999	1.4
Klesius et al., 1991	0.05
Griffith et al, 1992	-0.46
Stuart, 1999	0.37
Eldridge, 1991	-0.81
Morrow, 1992*	-1.23
Knapp and Associates, 1996	-0.17

(Negative effect size = whole language superiority)

*Group 2 results only

Table 7. Phonics vs. Whole Language: All Studies

to $d = -.168$. The mean in Table 7 is heavily influenced by the Santa and Hoiem study, which had a small sample size and utilized only high-risk students. If this study is removed, the mean for this group increases to $-.33$ in favor of whole language.

References

- Eldredge, L. (1991). An experiment with a modified whole language approach in first-grade classrooms. *Reading Research and Instruction, 30*, 21-38.
- Evans, M., and Carr, T. (1985). Cognitive abilities, conditions of learning, and the early development of reading skill. *Reading Research Quarterly, 20* (3), 327-350.
- Foorman, B., Francis, D., Fletcher, J., Shatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90*, 37-55.
- Freppon, P. (1991). Children's concepts of the nature and purpose of reading in different instructional settings. *Journal of Reading Behavior, 23*, 139-163.
- Garan, E.M. (2001). What does the report of the national reading panel really tell us about teaching phonics? *Language Arts, 79* (1), 61-71.
- Goodman, K. (1982). *Language, literacy, and learning*. London: Routledge Kagan Paul.
- Griffith, P., Klesius, J., and Kromrey, J. (1992). The effect of phonemic awareness on the literacy development of first grade children in a traditional or a whole language classroom. *Journal of Research in Childhood Education, 6*, 85-92.
- Hagerty, P., Hiebert, E., and Owens, M. (1989). Students' comprehension, writing, and perceptions in two approaches to literacy instruction. In S. McCormick and J. Zutell (Eds.), *Thirty-eighth yearbook of the National Reading Conference* (pp. 453-459). Chicago: National Reading Conference.
- Johnston, B. (1993). *DSTAT: Software for the meta-analytic review of research literatures*. Mahwah, NJ: Erlbaum.
- Klesius, J., Griffith, P., & Zielonka, P. (1991). A whole language and traditional instruction comparison: Overall effectiveness and development of the alphabetic principle. *Reading Research and Instruction, 30*, 47-61.

- Knapp, M. and Associates. (1995). *Teaching for meaning in high-poverty classrooms*. New York: Teachers College Press.
- Krashen, S. (1999). *Three arguments against whole language and why they are wrong*. Portsmouth, NH: Heinemann.
- Krashen, S. (2001). More smoke and mirrors: A critique of the National Reading Panel report on fluency. *Phi Delta Kappan* 83, 119–123.
- McQuillan, J. (1998). *The literacy crisis: False claims and real solutions*. Portsmouth, NH: Heinemann.
- Morrow, L. (1992). The impact of a literature-based program on literacy achievement, use of literature, and attitudes of children from minority backgrounds. *Reading Research Quarterly*, 27(3), 250–275.
- National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development. Available online at <http://www.nationalreadingpanel.org/Publications/summary.htm>.
- Santa, C., and Høien, T. (1999). An assessment of early steps: A program for early intervention of reading problems. *Reading Research Quarterly*, 34, 54–79.
- Smith, F. (1994). *Understanding reading* (5th ed.) Hillsdale, NJ: Erlbaum.
- Stuart, M. (1999). Getting ready for reading: Early phoneme awareness and phonics teaching improves reading and spelling in inner-city second language learners. *British Journal of Educational Psychology*, 69, 587–605.
- Trawick, D., and Berninger, V. (1997). Comparisons of beginning literacy programs: Alternative paths to the same learning outcome. *Learning Disabilities Quarterly*, 20, 160–168.
- Wilson, K., and Norman, C. (1998). Differences in word recognition based on approach to reading instruction. *The Alberta Journal of Educational Research*, 44, 221–230.

Stephen Krashen is Professor Emeritus at the University of Southern California.