A Longitudinal Study of Enhancing Critical Thinking and Reading Comprehension in Title I Classrooms

Joyce VanTassel-Baska and Bruce Bracken
The College of William and Mary

Annie Feng
National Institute of Health

Elissa Brown
North Carolina Department of Public Instruction

A longitudinal study of student growth gains was conducted in Title I schools to assess growth in reading comprehension and critical thinking. Results suggested that all students benefited from the intervention of Project Athena units of study designed for high-ability learners. In addition, the study suggested that the comparison curriculum also benefited learners. Implications for practice include the use of high-level curriculum with all learners to elevate instruction and enhance critical thinking. Implications for scholarship include the need for studies that examine the specific nature of gains for different types of learners and schools using hierarchical linear modeling techniques.

Over the past decade, studies have continued to suggest the importance of critical thinking and reasoning to high-level production within and across domains (Csikszentmihalyi, 2000; Gardner, 1991). Although most K–12 programs for gifted students include some components of critical thinking as a fundamental part of the curriculum (Chandler, 2004), only recently has the efficacy of such curriculums been tested in respect to student growth in this integral area of learning at various

Joyce VanTassel-Baska is Professor Emerita at The College of William and Mary where she has served as Executive Director of the Center for Gifted Education. Bruce Bracken is Professor of Education at The College of William and Mary. Annie Feng is Research Behavioral Scientist with the National Institute of Health where she is engaged in various studies on cancer. Elissa Brown is Director of the Middle/High School Council at the North Carolina Department of Public Instruction.

Journal for the Education of the Gifted. Vol. 33, No. 1, 2009, pp. 7–37.

stages of development. At the secondary level, proxy outcome data like Advanced Placement (AP) scores, International Baccalaureate (IB) scores, and SAT scores are used to inform educators about how these students are performing at higher levels of thought (VanTassel-Baska, Feng, Brown, Baytops, Henshon, & Bai, 2002). However, we have not systematically assessed the performance of elementary school students on tasks that require higher level critical thinking that become the preparation for more advanced work in secondary programs like AP and IB. Such an effort involves the development and testing of advanced curriculum in relevant subject areas that stress higher level thinking in the domain and the training of teachers to deliver it.

Although studies have shown that students show significant and important gains in content-specific higher order skills such as literary analysis and persuasive writing on performance-based language arts measures (VanTassel-Baska, Zuo, Avery, & Little, 2002) or designing experiments in science on performance-based measures (VanTassel-Baska, Bass, Ries, Poland, & Avery, 1998), studies have not demonstrated that a content-based intervention has provided students with enhanced generic critical thinking and reasoning skills at these same grade levels.

Moreover, it requires the testing of curriculum, designed for high-ability learners, to be used with all learners, especially those from low-income backgrounds, to make the case that higher level thinking can be positively improved for all through targeted interventions.

Purpose of the Study

The purpose of this study was to learn if an integrated language arts curriculum unit of study designed for gifted learners could impact the learning of all students in Title I settings on the dimension of higher level thinking. After using the William and Mary language arts units for 3 years, the researchers assessed both reading comprehension skills and critical thinking abilities in elementary age (grades 3–5) learners in six different school districts representing urban, exurban, and rural demographics across two states. Assessment that stressed critical behaviors and higher level reasoning skills was completed before and after the William and Mary language arts intervention.

The study was longitudinal in that some participants experienced the intervention for 3 years while other participants experienced it for 1 or 2 years.

The William and Mary Curriculum Intervention

The curriculum units of study used in this study were designed for high-ability learners according to the Integrated Curriculum Model (ICM; VanTassel-Baska, 1988, 1998, 2003; VanTassel-Baska & Stambaugh, 2006), which posits that high-ability learners need a curriculum that provides advanced-level work and high-level thinking processes, and is organized around a relevant concept or theme that encourages reflective thinking about real-world issues and problems. In addition, the units were designed to be responsive to features of exemplary curriculum for all learners in the language arts, including multicultural literature, persuasive writing, oral communication, and language study.

The William and Mary curriculum outcomes emphasized literary analysis and interpretation skills, persuasive writing, oral communication, and vocabulary development. Use of short reading selections in the genres of poetry and short stories encouraged students to analyze their understanding of the reading selections in respect to vocabulary, reader response, meaning, images and symbols, and structure. Teachers also asked probing questions about each selection that encouraged interpretation and application to the concept of change. In writing, students were asked to use a persuasive writing model to compose essays based on relevant prompts limited to the reading selections. Vocabulary enrichment was also stressed through a focus on 20–25 words in the unit being examined for meaning and definition, antonyms and synonyms, and word analysis, including stems, etymology, and word families. A grammar packet for student self-study was included for upper level grades 4 and 5 to teach form, function, and the selective combination of words to make meaning. Each unit also included a research project linked to the topic of the unit.

Instructional strategies to implement the units were encouraged through professional development and included the asking of higher level thinking questions; the deliberate teaching of graphic organizers to help students structure their thinking about literature writing

and vocabulary; and models for teaching thinking and research that embedded metacognition. Teachers were also encouraged to use instructional grouping to accommodate different reading levels.

The Comparison Curriculum

The experimental curriculum differed from the comparison curriculum in its emphases on the use of a higher level concept and higher level thinking and in its integrated approach to teaching the language arts in a connected way, rather than only focusing on reading. Moreover, the curriculum design deliberately differentiated for strong students in respect to depth and complexity in the tasks, questions, and advanced readings. Comparison curriculum in each district followed the Reading First approved program materials that stressed specific reading skills and employed small-group instruction, discussion, and worksheets to implement the program.

Implementation

Each unit of study required at least 6 weeks to implement each fall from October through December in most settings. The units for each of the grade levels involved were organized according to the description above and implemented as described. Use of flexible grouping was employed by most teachers in the implementation of the various emphases in the curriculum. Teachers worked with students in discussion groups ranging in size from 6 to 10 students, facilitated small-group learning on various activities, and provided direct instruction on the graphic organizers and models employed throughout the unit.

Professional Development

Teacher training consisted of 4 days of 6 hours each annually, focused on the strategies and models described above. Three days were provided in the summer preceding each intervention year, and the fourth day was provided at the end of each intervention year and included teacher feedback and reflection on the implementation. A total of 12 days of professional development was provided across the 3 years.

Literature Review

The evidence for use of the William and Mary language arts units as the basis for this study emerges from two literature bases. One base comes from what works in teaching language arts at elementary levels. The second emerges from prior work on using the units in school-based settings with gifted learners over the prior 10 years.

What Works in Teaching Literature

Theoretical support for the William and Mary language arts units, used as the foundation of this study, emerges from the literature on effective instruction in language arts at elementary levels. Response-based approaches to teaching literature have been advocated strongly in the work of English educators during the past 2 decades (e.g., Langer, 1994; Rosenblatt, 1982), and action research in classrooms, even at the primary level, has substantiated student growth in more complex thinking when the instructional approach is balanced between teacher-initiated activities and student response (Baumann & Ivey, 1997; Jewell & Pratt, 1999). The importance of "discussion moves" such as recapping, focusing, and reframing students' responses in teaching reading comprehension and interpretation is highlighted in the work of Beck and McKeown (1999).

What Works in Teaching Writing

Augmentation of reading comprehension strategies with writing instruction is yet another approach that has been shown to improve reading comprehension significantly (Langer, 2000). In a study investigating the efficacy of writing instruction, Applebee and Langer (2006) found that 67% of eighth-grade students are expected to write an hour or less a week, thus raising the question of insufficient writing time as a factor in literacy underdevelopment. Recent research also has revealed the learning benefits of integrating reading and writing tasks (Bottomley, Truscott, Marinak, Henk, & Melnick, 1999; Henry & Roseberry, 1996; Newell, 1996). Reviews and meta-analyses on the teaching of writing during the past 15 years have demonstrated the significance of key instructional variables in the process (e.g., Hillocks,

1986; Levy & Ransdell, 1996; Sadoski, Willson, & Norton, 1997). It has been demonstrated that a combination of (a) inquiry activities, (b) analyzing or responding to literature, (c) prewriting preparation, and (d) the use of scales reflecting specific criteria all contributed to enhanced student gains in writing. Making the activity of writing about literature a central component of the instructional approach also is predictive of higher test scores on measures of writing quality (Applebee, Langer, Mullis, Latham, & Gentile, 1994).

Teaching Critical Reading Behavior

Some recent studies assert that the combination of "teaching for reading" skill development and "teaching for comprehension and text meaning" produce the largest gains in student achievement, suggesting the use of literature-based materials in combination with structured basal series (Dahl, Scharer, Lawson, & Grogan, 1999; Morrow, Pressley, Smith, & Smith, 1997). This combined approach is perceived as a balanced instructional model, which allows for both direct instruction by the teacher and constructivist inquiry by the learner (Fitzgerald & Noblit, 2000; Snow, Burns, & Griffin, 1998). The National Reading Panel (2000) suggested that direct instruction in comprehension is vital, advocating a model that explains, guides practice, and provides independent practice with feedback and discussion. Stevens (2003) found the most helpful instructional reading strategies for 4,000 urban middle school students were summarizing, finding the main idea, and clarifying. Writing and cooperative learning were also successfully employed as a complement to this approach. Alvermann (2002) emphasized these same strategies but also employed questioning as another beneficial instructional approach. The role of strategic instruction has been shown to be critically important at the middle school level (Schorzman & Cheek, 2004) for all ability groups (Applebee, Langer, Nystrand, & Gamoran, 2003), as well as for students with learning disabilities (Gersten, Fuchs, Williams, & Baker, 2001). Other researchers have found that open discussion among peers was an essential strategy for improving literacy (Alvermann, Umpleby, & Olson, 1996; Applebee et al., 2003). Scaffolded instruction and peer interaction combined resulted in greater gains in reading comprehension than either approach alone (Langer, 2001).

The instructional time expended ensuring that students become autonomous readers would suggest the need for methodologies that deliberately move students' skills from basic decoding and fluency to comprehension of text and beyond. Such an approach to reading instruction ensures that students also can traverse the path from basic comprehension skills to higher level critical reading skills, while employing the same reading methods and strategies. Instructional scaffolding that embeds strategic instruction in text reading, as described above, has been shown to effectively enhance reading comprehension (Fielding & Pearson, 1994; Villaume & Brabham, 2002). Moreover, teachers who emphasize higher order thinking among their students thoughtfully employ reflective questioning strategies and provide tasks that promote greater reading growth (Knapp et al., 1995; Taylor, Pearson, Peterson, & Rodriguez, 2003).

Prior Evidence of Effectiveness of the William and Mary Language Arts Curriculum

Evidence for the effectiveness of the language arts intervention was established in earlier but less well-controlled studies. These studies employed quasi-experimental conditions with teachers volunteering their classrooms for piloting purposes across multiple sites and states. This phase used only curriculum-based assessment techniques to assess pre/post student learning gains in literary analysis and persuasive writing. Experimental groups were predominantly high-ability learners. Comparison groups were gleaned from the same district and selected according to ability and socioeconomic status (SES) considerations (VanTassel-Baska, Johnson, Hughes, & Boyce, 1996; VanTassel-Baska Zuo, et al., 2002). The studies focused explicitly on student application of literary analysis and interpretation, persuasive writing, and linguistic competency (VanTassel-Baska et al., 1996; VanTassel-Baska, Zuo, et al., 2002).

Results of these studies showed that there were significant pre/post student gains and significant differences between the experimental and comparison groups, favoring experimental students who were exposed to the William and Mary language arts curriculum. Using the eta squared statistic to assess effect sizes, results were medium for literature (.070) and high for persuasive writing

(.242). Gender differences found were small and not educationally important.

Findings from a 6-year longitudinal study using performance-based assessment in literary analysis and interpretation and persuasive writing that examined the effects over time of using the William and Mary language arts program in a suburban school district suggested that gifted student learning at grades 3 to 5 was enhanced at significant and educationally important levels in critical reading and persuasive writing. Effect sizes, using Cohen's *d*, ranged from .52–.79 for literary analysis and interpretation and from .66–1.28 for persuasive writing. Repeated exposure over a 2- to 3-year period demonstrated increasing achievement patterns, and the majority of stakeholders reported the curriculum to be beneficial and effective (Feng, VanTassel-Baska, Quek, Bai, & O'Neill, 2005). Moreover, an interview study of selected school district leaders documents that the curriculum also impacted positive school change in respect to climate, collegiality, and district policy change (VanTassel-Baska, Avery, Little, & Hughes, 2000).

Although these studies had been well-designed and implemented, they suffered from the limitations of the type of instrumentation employed as outcome measures; performance-based assessments are not as technically sound as standardized assessment tools. Moreover, the volunteer nature of the teacher samples limit generalizability. These studies were conducted in cluster-grouped classrooms, pull-out settings, and self-contained settings for gifted learners. The findings cannot be generalized to using the curriculum in heterogeneous settings with learners not identified as gifted.

Method

To assess the efficacy of the targeted language arts curriculum, this study employed a quasi-experimental research design and included a randomized assignment of classroom teachers (grades 3–5) into experimental and comparison conditions. The students for the study were therefore intact in the experimental and comparison teacher classrooms. Classrooms were heterogeneous and randomly assigned

by the school principals each year of the study, according to their self-report of how classes were constructed.

Participants

Over a 3-year implementation cycle, there were a total of 2,771 students who participated in the study. There was a balanced distribution of students in the experimental (52–54% of the student sample) and comparison (46–48% of the student sample) groups over the 3 years. The data showed that there was a balanced distribution of male (49%) and female (51%) students in the sample. This pattern was consistent across the 3 years of the study although student attrition occurred each year.

With respect to ethnicity, 43% of the students were Caucasian, 28% were African American, 18.7% were of Hispanic background, and 3.5% were Asian. Less than 2% of the student sample were Native Americans or Hawaiian or Pacific Islanders. The minority population, including Asian, African American, Hispanic, and others, comprised 50.2% of the student sample, attaining one project goal of scaling up the curriculum to a population of low-income and/or minority background students.

Students were divided into multiple classrooms at each of the school sites ($N = 11$), according to district teacher-pupil ratios. In all sites except one, experimental and comparison classrooms were in separate buildings. Classroom size ranged from 12 to 25 students. The total number of teachers in each year of the study was, on average, 74, including 38 experimental and 36 comparison teachers, representing intact classrooms that implemented the curriculum. The teacher portion of this study is reported elsewhere (see VanTassel-Baska, Bracken, Feng, & Brown, in press). Elementary schools across six districts in two states in the Mid-Atlantic area of the country participated in this study. All 11 schools in the study were designated as Title I, indicating that the majority of the student body (varying from 50% of the student population up to 84%) was on a free or reduced lunch program. Districts represented in the study included urban, exurban, and rural; therefore, accounting for some differences in the overall numbers of students in the study as well as the percentage of diverse populations attending each school. Number of participating students by district ranged from 78 to 213.

Instruments

Four pretest instruments given to the entire sample were used to assess students' incoming levels of cognitive and academic functioning and to match or covary the participating students' entering abilities across treatment groups. These instruments included one group-administered test of cognitive functioning (CogAT; Lohman & Hagen, 2001), one individually administered comprehensive nonverbal intelligence test (UNIT; Bracken & McCallum, 1998), a test of critical thinking (TCT; Bracken et al., 2003), and the reading comprehension subtest portion of a group-administered achievement test (ITBS; Hoover, Dunbar, & Frisbie, 2001). Posttest instruments that were used to measure students' learning outcomes included the ITBS Reading Comprehension subtest and the TCT. Two performance-based measures, literary analysis and persuasive writing, were also administered to experimental students as pre/post tests during each implementation year. Additionally, a structured observation scale, the Classroom Observation Scale–Revised (COS–R; VanTassel-Baska et al., 2003) was used to monitor treatment fidelity as well as teachers' instructional practices in both experimental and comparison classes.

The CogAT. The Cognitive Abilities Test (CogAT; Lohman & Hagen, 2001) is a well-known and established group-administered measure of cognitive functioning. The verbal and nonverbal components of the CogAT were group administered to each participant. Internal consistency reliabilities for this instrument ranged from .93 to .95. The CogAT technical manual indicates strong evidence for the instrument's technical adequacy (Lohman & Hagen, 2001).

The UNIT. The Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998) is a nonverbal intelligence test with a special appeal for application with low-income and minority students and those who speak English as a second language (Bracken & McCallum, 1998). The abbreviated battery, Symbolic Memory and Analogic Reasoning and Cube Design, was administered to participating students. Average internal consistency coefficients for the Abbreviated Battery are reported as .91 for the entire sample, .96 for combined clinical/exceptional samples, and .96 for a gifted sample. The corresponding reliability coefficient

was .96 for African Americans and .94 for Hispanics. It has a stability coefficient of .83 for the total sample over a 4- to 6-week retest interval.

The ITBS. The Iowa Tests of Basic Skills (ITBS; Hoover et al., 2001) is a commonly used group-administered achievement test, with strong evidence of technical adequacy. Using KR-20 coefficients, ITBS internal consistency coefficients for the Reading Comprehension scale for the third, fourth, and fifth grades used in this study were .88, .87, and .86, respectively. This study used only the subtest of reading comprehension, which requires students to select the best answer based on reading a passage.

The TCT. The Test of Critical Thinking (TCT; Bracken et al., 2003) is a 45-item instrument designed to assess the critical thinking skills of students in grades 3, 4, and 5. The TCT was developed using Paul's (1992) model of critical thinking, including his eight elements of thought (i.e., issue, purpose, concept, point of view, assumptions, evidence/information, inferences, and implications/consequences). The TCT is a group-administered test that consists of 10 short stories or scenarios, each of which is followed by several multiple-choice questions. The TCT presents a balanced framework of critical thinking elements within interesting stories that reflect seven important life domains for children and adolescents (Bracken, 1993, 1996; Wasserman & Bracken, 2003), making it both useful and relevant to the lives of young students. The internal consistency of the instrument was .81. It has a 6-month stability coefficient of .66 and strong convergent correlations with the ITBS Reading Comprehension scale and the CogAT Verbal scale (i.e., .63 and .63, respectively) and expected lower discriminate coefficients with the CogAT Nonverbal Scale (r = .45) and UNIT full scale IQ (r = .25).

Literary Analysis and Persuasive Writing. Literary analysis is a performance-based measure of literary analysis and interpretation. This test, modeled on the NAEP assessment in reading (National Assessment Governing Board, 1992) addresses four task demands: (a) main idea, (b) analysis of a quote, (c) relationship of the concept of change to selection, and (d) creating a title with a rationale to support it. The second performance-based assessment, persuasive writing, asks students

to develop an argument to support or reject a statement. Both assessments were reviewed for content validity by experts in English and gifted education and were given favorable reviews. Interrater reliability estimates for scoring each instrument exceeded .90 for each scorer team comprised of three teams of graduate students and project staff (VanTassel-Baska, Zuo, et al., 2002).

The COS–R. The Classroom Observation Scale–Revised (COS–R; VanTassel-Baska et al., 2003) is a scale developed for assessing teachers' instructional practice against expectations derived from best practices in mainstream and gifted education classrooms. The instrument was developed with theoretical bases from the reform literature, general teaching practices, as well as literature in differentiation strategies, and has gone through several revisions and reiterations. The COS–R total scale has evolved into a scale comprising 25 expected teaching behaviors subsumed under six subscales. The presence of a certain teacher behavior is measured using a Likert scale of 1 to 3, with 1 being *not effective*, 2 being *somewhat effective*, and 3 being *effective*. The internal consistency reliability for the COS–R was .91–.93. The content validity established by expert review agreement using intraclass coefficient was .98 (VanTassel-Baska, Quek, & Feng, 2007). Observation data were used for measuring treatment fidelity throughout the project. A companion student observation scale (SOS) is embedded in the COS–R that documents students' responding behaviors to teachers' instruction. The internal consistency reliability for the SOS ranged from .85–.92. A research-staff developed supplemental scale was also used to assess teacher implementation of the curriculum models during each observation period to augment COS–R findings.

Selection of Outcome Measures

The content analysis of assessment measures is needed for intervention studies because differential measures will produce differential results (Schoenfeld, 2006). The same logic may be applied to different curricula.

The choice of the TCT and ITBS Reading Comprehension subtest as outcome measures was a deliberate one. In order to appropriately assess critical thinking, it was thought that a technically adequate

measure that used the same approach to critical thinking as the curriculum was needed. The TCT then was developed and piloted as described.

In the case of selecting the ITBS, we wanted a standardized measure of reading comprehension that could apply to all states in the study rather using the individual state tests that would be difficult to calibrate for comparison. Moreover, the ITBS was the test used to document gains in the Reading First program at primary levels in Michigan (U.S. Department of Education, 2004). Reading First was the comparison curriculum used in the majority of school sites. Therefore, using this measure allowed easier comparison of results to a program already deemed successful for low-income students.

Because the performance-based measures had been used in earlier studies, we decided to retain them for purposes of this study, as well as to provide deeper insight into student learning in the core areas of the curriculum: literary analysis and interpretation and persuasive writing. Moreover, the character of curriculum implementation is what matters; thus, there was a need to include the COS–R, which assessed fidelity of implementation.

Procedure

At each participating site, researchers randomly assigned grades 3–5 classes into an experimental or comparison condition. A pre/post design using ANCOVA was employed to covary any initial differences in reading and/or critical thinking skills. Participating experimental teachers were trained on the William and Mary language arts curriculum and were provided the necessary materials for implementation.

Pre/Post Testing. The ITBS Reading Comprehension and the TCT were administered before and after each implementation period of the intervention. Each new participating student was also administered the CogAT and UNIT at the beginning of his or her participation to determine incoming cognitive functioning levels. ANCOVA's were run to control for pretest differences between experimental and comparison groups.

Treatment Fidelity. After pretesting was complete, teachers implemented the language arts curriculum in their respective classrooms.

Each experimental teacher implemented a unit of study, comprising 24 lessons over the course of 6 to 8 weeks in the fall of each year. Treatment fidelity was addressed in several ways. Teachers maintained and completed implementation logs, noting lessons taught and judgments about their efficacy. Teachers and students were observed twice during each year of the intervention period by project staff using the COS–R and the SOS to assess both differentiation strategy use and project-specific lesson implementation. Teachers also were observed for their effective use of specific unit teaching models. Taken together, these procedures constituted a multidimensional approach to treatment fidelity.

Comparison Classroom Intervention

While the William and Mary language arts curriculum units were used by experimental classes, comparison classes continued to use district-selected curriculum over the same period of time. The major program used in participating school comparison classrooms was the Reading First Program (U.S. Department of Education, 2004). Reading First emphasizes reading fluency and comprehension and was presented daily in small reading and discussion groups. Students took turns doing oral reading and then responded in a group to teacher-generated questions that focused on text comprehension. Specific comprehension skills such as inference and prediction were emphasized. The Reading First Program has showed positive results enhancing reading comprehension in elementary school students (Armbruster & Osborn, 2003) although specific materials used varied.

Data Analyses

The primary research question for this study was the longitudinal impact of the language arts curriculum intervention; therefore, factorial mixed design repeated measures analyses were the major analytical tool used. The two outcome variables were the TCT and ITBS Reading Comprehension pre/post tests over 3 years of curricular intervention. Although we had a large sample of participants over 3 years (N = 2,771), there was also a substantial attrition of student participants for various reasons; one of the largest participating districts had

approximately a 50% attrition rate by the second year of the project due to military moves of families of participants. The number of students who had been with the project for 3 years (i.e., those who were third graders in Year 1) became low by the end of Year 3 implementation (approximately 130 in both experimental and comparison classes). Consequently, we had a substantially lower number of participants for data analyses than expected. Therefore, a liberal p value of .05 was chosen as the criterion for statistical significance testing, despite the fact that multiple tests had been conducted, in order to retain statistical power under a small sample while giving attention to Type I error.

Due to different forms of the ITBS that were administered at grades 3 to 5, all ITBS standardized scores were converted into an IQ metric with a mean of 100 and a standard deviation of 15 to report and analyze ITBS data in the same metric. Specifically, we subtracted from the ITBS standardized reading comprehension score the national norm mean (Fall 2003 Survey Battery) and divided that number by the standard deviation of the national norm, creating a z score. We then multiplied that z score by the IQ standard deviation of 15 and added to it the population mean IQ of 100. By converting ITBS raw scores into a standard metric, we were able to conduct analyses with aggregated data across grade levels, alleviating the likelihood of Type I error (Kaplan, 2004).

Results

Students' Learning Gains in 3 Years by Condition, Gender, and Ethnicity

To examine students' overall learning gains across 3 years as measured by the TCT and ITBS Reading Comprehension subsection as well as between two treatment condition groups, mixed designed (Time x Condition x Gender x Ethnicity) repeated measures analyses were conducted on the TCT and the ITBS Reading Comprehension subtest. The six pre/post tests of the TCT and the ITBS Reading Comprehension over the 3 implementation years comprised the six levels of the within-subject factor, and the treatment condition in

Year 3 implementation, gender, and ethnicity served as between-subject factors. Table 1 and Table 2 present means and standard deviations of the pre- and postassessments on the TCT and ITBS Reading Comprehension of both experimental and comparison students across 3 years. Experimental students in this sample obtained higher mean scores than control students at each assessment data point. The repeated measures test results showed that there was a significant time main effect in the multivariate testing using Wilks' Lambda ($\Lambda = 6.6$, $p = .000$). There was a significant time effect on the TCT, $F(5, 86) = 3.5$, $p = .004$, suggesting students' increased critical thinking skills over time. However, no significant growth effect was found on the ITBS Reading Comprehension subtest, $F(5, 86) = 1.0$, $p = .40$, across 3 years of implementation.

The results also showed that there was not a significant treatment effect on the TCT across 3 years, $F(1, 86) = 3.3$, $p = .07$, $\eta^2 = .034$; the 95% confidence interval for the experimental group was 20.4–23.1 and 17.9–21.4 for the comparison group. There was no significant treatment effect on the ITBS Reading Comprehension ($p = .47$), either. Cohen's d was calculated for each assessment point to show the magnitude of differences between experimental and comparison students over the 3 years; the data showed that there was a pattern of increasing effect sizes between experimental and comparison students' TCT performance from Year 1 to Year 3 implementation. By Year 3 postassessment, the magnitude of differences between experimental and comparison students reached $d = .40$ or above on the TCT assessment, suggesting that by the end of the third year of project implementation, experimental students achieved a .4 standard deviation above that of the comparison students on the TCT. Despite a lack of statistically significant differences between experimental and comparison group students on ITBS Reading Comprehension across the 3 years, a similar pattern of increasing performance differences was reflected in the effect size d (see Table 2) suggesting an increasing difference in rate of growth between the two groups, favoring the experimental student sample across the 3 years.

There was no significant gender effect on the TCT, $F(1, 86) = 2.5$, $p = .11$, nor on the ITBS Reading Comprehension, $F(1, 86) = 44$, $p = .51$. The results also showed that there was a significant ethnicity effect on TCT, $F(4, 86) = 4.3$, $p = .003$, but no significant ethnicity effect on

Table 1
Student Longitudinal Gains on TCT by Condition (*N* = 97)

	Experimental (n = 60)		Comparison (n = 37)		Effect size $(E\text{-}C)/SD_c$
Test	Mean	*SD*	Mean	*SD*	*d*
Year 1 TCT Pretest	15.6	6.3	13.8	5.6	.32
Year 1 TCT Posttest	18.7	6.8	18.6	4.5	.02
Year 2 TCT Pretest	21.0	6.9	19.2	5.5	.33
Year 2 TCT Posttest	23.5	6.7	20.5	6.7	.45
Year 3 TCT Pretest	25.0	7.4	22.2	5.8	.48
Year 3 TCT Posttest	26.1	6.3	23.7	6.0	.40

Note. Within-subject effect on TCT: F (5, 86) = 88.9, p = .000, η^2 = .48.

Table 2
Student Longitudinal Gains on ITBS by Condition (*N* = 97)

	Experimental (n = 60)		Comparison (n = 37)		Effect size $(E\text{-}C)/SD_c$
Test	Mean	*SD*	Mean	*SD*	*d*
Year 1 ITBS Pretest	107.1	12.6	106.8	14.2	.02
Year 1 ITBS Posttest	116.9	15.4	114.0	14.9	.19
Year 2 ITBS Pretest	109.8	15.4	107.3	14.8	.17
Year 2 ITBS Posttest	116.8	14.8	113.2	18.4	.19
Year 3 ITBS Pretest	110.1	14.0	105.2	15.8	.31
Year 3 ITBS Posttest	115.0	13.6	111.3	14.0	.26

Note. Within-subject effect on ITBS: F (5, 86) = 23.3, p = .000, η^2 = .20.

the ITBS Reading Comprehension, $F(4, 86) = .8.9$, $p = .000$, across 3 years. Post hoc analyses showed that White American students did significantly better than African American students, who did better than Hispanic American students ($p < .05$).

Student Longitudinal Gains on ITBS Reading Comprehension and TCT by Ability Level

In order to examine the similar or different achievement gains of participating students who are at different ability levels, further exploratory

analyses were performed. The baseline data as measured by the CogAT verbal and nonverbal components and the UNIT test scores were used to define ability levels. Specifically, a student who scored at or above 130 on any component of the ability tests (i.e., CogAT verbal or nonverbal, or the UNIT) was categorized as gifted. A student who scored at or above 115 but below 130 on any component of the ability measures was categorized as a promising learner; students who scored at or above 100 but below 115 were categorized as typical learners; students who scored at or above 85 but below 100 were categorized as low-end learners; and finally students who scored below 85 on any of these ability measures were classified as atypical learners. Therefore, there were five ability levels based on above-mentioned definitions: gifted, promising, typical, low-end, and atypical learners.

A two-way (Gifted x Condition) repeated measures analysis was explored to examine the longitudinal but differential gains due to ability levels. The six assessment points of the TCT and the ITBS Reading Comprehension were the six levels of the within-subject factors. Year 3 treatment condition and ability level were the between subject factors (i.e., time or growth factor). The condition and time, and ability and time, as well as the ability and condition interaction effects were tested using the multivariate criterion of Wilks' Lambda (Λ) and was found not significant ($p > .05$). The time effect (within subject factor) was significant, $\Lambda = 13.46$, $p = .000$ as well as the ability main effect, $\Lambda = 10.34$, $p = .000$. The condition main effect was not significant, $\Lambda = .66$, $p = .518$. Significant ability effects were registered on both the TCT, $F(4, 89) = 17.23$, $p = .000$, $\eta^2 = .44$, and the ITBS Reading Comprehension, $F(4, 89) = 21.8$, $p = .000$, $\eta^2 = .49$, longitudinally. The partial eta squared index suggested that the differences among different ability levels in terms of their performance on the TCT and ITBS Reading Comprehension were large. Post hoc analyses showed that promising learners did significantly better than typical learners ($p < .05$), who subsequently did significantly better than low end learners ($p < .05$). Gifted learners and atypical learners were excluded from the post hoc analyses due to the lower number of cases available for analysis ($n = 11$ and $n = 1$, respectively).

Table 3 and Table 4 present the means and standard deviations by ability level and treatment condition across the 3 years. Again, gifted and atypical learners were not included due to the low number of cases.

The descriptive statistics showed that over 3 years of project intervention, promising learners in the experimental group had an increase of 10.8 points on the TCT raw score scale, meaning they were able to answer 10 to 11 more questions correctly compared to when they were third graders; typical and low-end learners in the experimental group also had a similar increase on the TCT raw score scale. Promising learners and typical learners in the comparison group had an increase of 10 to 11 points; low-end learners in the comparison group had an increase of 5.4 points on the TCT after 3 years' implementation, less than their experimental counterparts (see Table 3). With regard to longitudinal gains on the ITBS Reading Comprehension at different ability levels, promising learners scored 11.4 more on the ITBS reading comprehension battery after 3 years of participation, typical learners had an increase of 6 points, and low-end learners had an increase of 7.6 points; the same corresponding ability level students in the comparison group had an increase of 2.7, 8.4, and 1.3 points respectively; the gains of the promising and low-end learners in the comparison groups were less than the increases made by the corresponding learners in the experimental groups.

Experimental Students' Learning Growth on Performance-Based Assessment

A subanalysis was also conducted of different ability levels with respect to student outcomes longitudinally on the two performance-based measures. No statistically significant ability main effect was registered in the multivariate testing, $\Lambda = 1.1$, $p > 05$. Neither was there a significant ability and time interaction; however, there was a significant time main effect, $\Lambda = 7.5$, $p = .000$), suggesting significant learning gains across 3 years. The longitudinal learning gains were also significant on both literary analysis, $F(5, 27) = 7$, $p = .000$, $\eta^2 = .21$, and persuasive writing, $F(5, 27) = 3.2$, $p = .027$, $\eta^2 = .11$, suggesting that experimental students made statistically significant and educationally important gains on performance-based measures after 3 years of curricular intervention. However, the longitudinal gains within each ability group cannot be generalized due to the low number of students who went through all 3 years of intervention and assessment.

Table 3
Student Longitudinal Gains on TCT by Ability and Condition (*N* = 85)

		Year 1 TCT pretest		Year 1 TCT posttest		Year 2 TCT pretest		Year 2 TCT posttest		Year 3 TCT pretest		Year 3 TCT posttest	
Ability	Condition	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
130 > IQ ≥ 115	Exp (n = 16)	17.8	6.3	21.1	5.5	22.6	5.4	24.9	5.2	26.9	5.5	28.6	6.0
	Control (n = 15)	14.8	6.1	20.2	4.2	21.4	3.7	23.6	3.9	24.2	4.9	25.3	5.4
115 > IQ ≥ 100	Exp (n = 23)	14.8	5.4	16.4	4.3	19.9	5.1	22.9	5.0	23.5	5.7	25.4	4.9
	Control (n = 16)	13.8	6.1	18.8	4.2	19.4	5.9	20.6	5.9	22.7	5.8	24.8	5.7
100 > IQ ≥ 85	Exp (n = 9)	8.9	4.1	12.2	3.5	14.0	5.3	16.7	4.7	17.8	6.9	20.3	5.3
	Control (n = 6)	11.8	1.5	14.2	3.6	13.3	3.7	12.5	8.2	15.7	3.1	17.2	4.5

Note. TCT = Test of Critical Thinking; Exp = Experimental group; Control = Control group.

Table 4
Student Longitudinal Gains on ITBS by Ability and Condition (*N* = 85)

		Year 1 ITBS pretest		Year 1 ITBS posttest		Year 2 ITBS pretest		Year 2 ITBS posttest		Year 3 ITBS pretest		Year 3 ITBS posttest	
Ability	Condition	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Promising learners 130 > IQ ≥ 115	Exp (n = 16)	111.6	8.2	121.7	12.2	115.2	14.0	122.2	9.4	112.6	13.0	122.0	13.1
	Control (n = 15)	115.4	10.7	122.5	13.0	116.9	11.8	122.6	12.3	112.8	13.5	118.1	11.3
Typical learners 115> IQ ≥ 100	Exp (n = 23)	106.4	10.3	114.3	10.5	109.0	8.7	115.3	11.9	109.0	12.8	112.8	10.7
	Control (n = 16)	103.5	12.3	112.0	11.9	104.6	11.7	112.6	15.4	105.1	14.2	111.8	11.6
Low end Learners 100 > IQ ≥ 85	Exp (n = 9)	91.7	8.2	98.0	12.6	87.5	12.3	95.0	11.4	95.9	7.2	99.3	11.2
	Control (n = 6)	91.3	11.8	91.5	9.4	90.7	12.5	91.0	21.5	86.2	9.3	92.6	9.3

Note. ITBS = Iowa Test of Basic Skills; Exp = Experimental group; Control = Control group.

Discussion and Implications

The results of this study suggest that across 3 years of curricular intervention, both experimental and comparison students made statistically significant and educationally important pre/post learning gains on the TCT but not on ITBS Reading Comprehension. No statistically significant treatment main effects were found on the TCT or ITBS Reading Comprehension subtest, suggesting that experimental and comparison students performed equally well on the outcome measures across 3 years. Experimental students also showed significant and important gains in the curriculum-based assessment areas of literary analysis and persuasive writing. However, despite a lack of statistically significant differences between the two groups on critical thinking and reading comprehension measures, an increasing gap in score results, favoring the experimental group students in the sample, was evident across the 3 years.

The lack of significance in longitudinal results on TCT and ITBS Reading Comprehension may be explained in a number of ways. One explanation might be attributed to the student attrition rate. In fact, the extent of student attrition for each year was somewhat alarming, ranging from 50% attrition between Year 1 and Year 2 in one district where the school is home primarily to military personnel to 50% attrition in another district between Year 2 and Year 3 due to principal reassignment of half of the experimental classes to the regular curriculum because of lower than anticipated results on the state assessment test the preceding spring. In addition to these unforeseen losses of study students, the staggered design across 3 years (i.e., third graders would experience the curriculum for 3 years, fourth graders for 2 years, and fifth graders for only 1 year) meant we would lose two thirds of the participants anyway for purposes of the longitudinal study. Given a significantly reduced sample size for the longitudinal data due to above mentioned reasons, significance might well have been impacted.

Moreover, the use of Reading First in these same schools as the comparison curriculum made it more challenging to show results as administrators believed that Reading First provided the targeted focus for reading that their students needed; the program was district-wide and state-approved and received ongoing funding support. The issue of multiple innovations being used at the same time, especially in the

same subject area, has been cited earlier as a problematic context for obtaining favorable results (Coburn, 2003; Desimone, 2002).

The lack of a significant growth effect on the ITBS Reading Comprehension might also be related to the lack of equivalence among the ITBS forms; students took grade-level correspondent forms of the test as they graduated from a lower to a higher grade. Moreover, it should be acknowledged that the ITBS is an assessment measure less sensitive to change after transforming it to a standard score. However, the ITBS was converted into a standard score metric mainly because different levels of the ITBS had been administered as students matriculated from a lower grade to a higher grade level.

Although the classroom observation data did not support the contention of teacher contamination, it is possible that some contamination effect was at work in some of the school settings where the experimental and control classrooms were in the same school. In an age of reform where schools are encouraged to engage staff in grade-level teaming, it is highly possible that teachers shared relevant information about the curriculum and especially what worked for them with colleagues in the comparison group, even though the researchers explicitly asked them to agree not to do so. Given the level of teacher attrition, the need to share strategies that work with new staff may have superseded concerns about research protocol. As the study progressed across years, the concerns about contamination may not have been stated as strongly to new teachers joining the project, and principals may have become less vigilant in monitoring its presence. Moreover, data on teacher use of differentiation in Project Athena favored veteran teachers who stayed with the project across all 3 years ($N = 16$). Thus teacher attrition in the project may also have adversely affected longitudinal results.

Although we know that the Reading First curriculum was the comparison curriculum in most schools, we do not know as much about its implementation as we would like. Given that the emphases appeared to vary by school and district site and across two states, it is difficult to ascertain all the ways its implementation was both similar and different from the experimental curriculum. On the narrow gauge of ITBS Reading Comprehension, we know that it appeared to be as or more successful than the William and Mary curriculum. Yet, its impact on critical thinking growth was slightly less powerful. In the

absence of an implementation plan that could be analyzed, we were left with describing this alternative intervention in more general terms, based on written accounts in the districts and on direct observation in comparison teacher classrooms at six different points across the 3 years. A more thorough analysis of fidelity of implementation of this program would have allowed better inferences about the relative merits of the program in relation to the experimental curriculum.

The results that experimental students in the sample obtained higher mean scores than comparison students on the TCT ($p = .07$) across 3 years of implementation and registered a 95% confidence interval with its lower bound approximating the upper bound of the comparison group's confidence intervals suggested that the William and Mary language arts curriculum reinforced critical thinking more than the alternative curriculum employed in comparison classrooms. Yet, the fact that the curriculum did not enhance learning to a greater extent in reading comprehension or thinking may relate to the curriculum design that assumed a degree of automaticity or fluency in the reader since the units of study were designed for high-ability readers. It may be that more emphasis on reading fluency and lower level skill development in comprehension may have been necessary for some project students. A supplementary reading comprehension program was developed in Years 2 and 3 to provide additional scaffolding for comprehension development although it was used on a voluntary basis.

Although the numbers of gifted learners in this study were too small for meaningful analysis, it was gratifying to see other groups of learners benefit from a curriculum designed for gifted learners, especially promising learners on the cusp of becoming stronger. The extent of gain, however, as seen in earlier studies with gifted students, was related to their functional level in ability on the pretest measures. As in those earlier studies as well, grouping, teacher competency, and treatment fidelity all were issues that could have impacted results. While race and gender effects were less evident in earlier studies, the results in this study are more consistent with other studies that have examined learning gains of elementary students on both achievement and aptitude measures (Jensen, 1998).

Although both groups showed significant gains on relevant outcome measures, these longitudinal results across 3 years suggest that the intervention was promising for the experimental participants,

suggesting that the alternative curriculum was also successful. This finding bodes well for the use of alternative reading programs in Title I schools. While the comparison curriculum had prior evidence of effectiveness with all learners, the William and Mary curriculum had not been used with all students in earlier studies, but only with those students identified as gifted.

This scaling-up study of the William and Mary language arts curriculum provided evidence that high-powered curriculum designed for high-ability learners can be successfully used in regular classroom settings to the benefit of all learners. Performance-based measures also yielded significant and educationally important results for the experimental students in all ability groups (VanTassel-Baska, Bracken, Brown, & Feng, 2005), again suggesting that the curriculum is effective with a broad range of learners.

Limitations of the Study

This study has important limitations, however. One limitation was in the choice of instrumentation, especially the ITBS Reading Comprehension subscale, which does not allow easy calibration of longitudinal growth across years. Thus, the growth curves associated with a 3- year pattern of performance are compromised on the results reported for this instrument. The other limitation was the size of the attrition in the project across 3 years, both for students and teachers. Based on earlier data on the Title I schools involved, it was predictable that attrition would be a problem; it still was hard to fathom the extent of attrition of teachers across the 3-year period. Although we replaced experimental and comparison teachers each year, it may well have impacted results in ways that cannot be assessed. At the very least, it may have caused less expert implementation of the curriculum, given less time in training and exposure.

Implications for future research would include (1) a more careful study of the longitudinal effects of alternative curricula on discrete categories of students with larger sample sizes and guaranteed ongoing student cohorts; (2) the use of better outcome measures that are sensitive to student gains across years as the ITBS Reading Comprehension was not satisfying in this regard; (3) studies of both younger and older populations in Title I schools using a curriculum designed for

high-ability learners; and (4) a follow-up study that would examine growth curves, using hierarchal linear modeling (HLM), to partial out the attribution of variables at different levels (i.e., student, class, school) in order to better assess the contributions of key variables to student learning.

References

Alvermann, D. (2002). Effective literacy instruction for adolescents. *Journal of Literacy Research, 34,* 189–209.

Alvermann, D. E., Umpleby, R., & Olson, J. R. (1996). Getting involved and having fun: Dilemmas in building a literate community in one lower-track English class. *International Journal of Qualitative Studies in Education, 9,* 461–475.

Applebee, A. N., & Langer, J. A. (2006). *The state of writing instruction in America's schools: What existing data tell us.* Albany, NY: National Research Center on English Learning and Achievement.

Applebee, A. N., Langer, J. A., Mullis, I. V. S., Latham, A. S., & Gentile, C. A. (1994*). NAEP 1992 writing report card.* Washington, DC: Office of Educational Research and Improvement.

Applebee, A. N., Langer, J., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal, 40,* 685–730.

Armbruster, B. B., & Osborn, J. (2003). *Put reading first: The research building blocks of reading instructions.* Retrieved July 11, 2009, from http://www.nationalreadingpanel.org/Publications/researchread.htm

Baumann, J. F., & Ivey, G. (1997). Delicate balances: Striving for curricular and instructional equilibrium in a second-grade, literature/strategy-based classroom. *Reading Research Quarterly, 32*, 244–275.

Beck, I. L., & McKeown, M. G. (1999). Comprehension: The sine qua non of reading. *Teaching and Change, 6,* 197–211.

Bottomley, D. M., Truscott, D. M., Marinak, B. A., Henk, W. A., & Melnick, S. A. (1999). An affective comparison of whole language,

literature-based and basal reader literacy instruction. *Reading Research and Instruction, 29*, 115–129.

Bracken, B. A. (1993). *Multidimensional Self-Concept Scale*. Austin, TX: Pro-Ed.

Bracken, B. A. (1996). Clinical applications of a multidimensional, context-dependent model of self-concept. In B. A. Bracken (Ed.), *Handbook of self concept: Developmental, social, and clinical considerations* (pp. 463–505). New York: John Wiley and Sons.

Bracken, B. A., Bai, W., Fithian, E., Lamprecht, M. S., Little, C., & Quek, C. (2003*). The Test of Critical Thinking*. Williamsburg, VA: Center for Gifted Education, The College of William and Mary.

Bracken, B. A., & McCallum. R. S. (1998). *Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside.

Chandler, K. (2004). *A national study of curriculum policy and practice for gifted students in the fifty states*. Unpublished doctoral dissertation, The College of William and Mary, Williamsburg, VA.

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher, 32*(6), 3–12.

Csikszentmihalyi, M. (2000). Positive psychology: The emerging paradigm. *NAMTA Journal, 25,* 5–25.

Dahl, K. L., Scharer, P.L., Lawson, L. L., & Grogan, P. R. (1999). Phonics instruction and student achievement in whole language first-grade classrooms. *Reading Research Quarterly, 34,* 312–341.

Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research, 72,* 433–479.

Feng, A. X., VanTassel-Baska, J., Quek, C., Bai, W., & O'Neill, B. (2005). A longitudinal assessment of gifted students' learning using the Integrated Curriculum Model (ICM): Impacts and perceptions of the William and Mary language arts and science curriculum. *Roeper Review, 27,* 78–83.

Fielding, L. G., & Pearson, P. D. (1994). Reading comprehension: What works. *Educational Leadership, 51*(5), 62–67.

Fitzgerald, J., & Noblit, G. (2000). Balance in the making: Learning to read in an ethnically diverse first-grade classroom. *Journal of Educational Psychology, 92*, 3–22.

Gardner, H. (1991). *The unschooled mind: How children think and how schools should teach.* New York: Basic Books.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of the research. *Review of Educational Research, 71,* 279–320.

Henry, A., & Roseberry, R. L. (1996). A corpus-based investigation of the language and linguistic patterns of one genre and the implications for language teaching. *Research in the Teaching of English, 30,* 427–489.

Hillocks, G., Jr. (1986). *Research on written composition: New directions for teaching.* Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills and National Conference on Research in English.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). *Iowa Tests of Basic Skills.* Itasca, IL: Riverside.

Jensen, A. R. (1998). *The* g *factor: The science of mental ability.* Westport, CT: Praeger.

Jewell, T. A., & Pratt, D. (1999). Literature discussions in the primary grades: Children's thoughtful discourse about books and what teachers can do to make it happen. *The Reading Teacher, 52,* 842–850.

Kaplan, D. (Ed). (2004). *The Sage handbook of quantitative methodology for the social sciences.* Thousand Oaks, CA: Sage.

Knapp, M. S., Adelman, N. E., Marder, C., McCollum, H., Needels, M. C., Padilla, C., et al. (1995). *Teaching for meaning in high-poverty classrooms.* New York: Teachers College Press.

Langer, J. A. (1994). *A response-based approach to reading literature: Report Series No. 6.7.* Albany, NY: National Research Center on Literature Teaching and Language.

Langer, J. A. (2000). *Guidelines for teaching middle and high school students to read and write well: Six features of effective instruction.* Albany, NY: National Research Center on English Learning and Achievement.

Langer, J. (2001). Beating the odds: Teaching middle and high school students to read and write well. *American Educational Research Journal, 38,* 837–880.

Levy, C. M., & Ransdell, S. (Eds.). (1996). *The science of writing: Theories, methods, individual differences, and applications*. Mahwah, NJ: Erlbaum.

Lohman, D., & Hagen, E. P. (2001). *Cognitive Abilities Test norms booklet*. Itasca, IL: Riverside.

Morrow, L., Pressley, M., Smith, J., & Smith M. (1997).The effect of a literature-based program integrated into literacy and science instruction with children from diverse backgrounds. *Reading Research Quarterly, 32,* 54–78.

National Assessment Governing Board. (2009). *Reading framework for the 2009 National Assessment of Educational Progress*. Washington, DC: U.S. Government Printing Office.

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.

Newell, G. E. (1996). Reader-based and teacher-centered instructional tasks: Writing and learning about a short story in middle-track classrooms. *Journal of Literacy Research, 28,* 147–172.

Paul, R. (1992). *Critical thinking: What every person needs to survive in a rapidly changing world*. Sonoma, CA: The Foundation for Critical Thinking.

Rosenblatt, L. M. (1982). The literary transaction: Evocation and response. *Theory Into Practice, 21*, 268–277.

Sadoski, M., Willson, V. L., & Norton, D. E. (1997). The relative contributions of research-based composition activities to writing improvement in the lower and middle grades. *Research in the Teaching of English, 31,* 120–147.

Schoenfeld, A. (2006). What doesn't work: The challenge of the WWC to conduct meaningful reviews of studies of mathematics curricula. *Education Researcher, 35*(2), 13–21.

Schorzman, E., & Cheek, E. (2004). Structured strategy instruction: Investigating an intervention for improving sixth-graders' reading comprehension. *Reading Psychology,* 25(1), 37–60.

Stevens, R. (2003). Student team reading and writing: A cooperative learning approach to middle school literacy instruction. *Educational Research and Evaluation, 9,* 137–160.

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

Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal, 104*(1), 3–30.

U.S. Department of Education. (2004). *Reading first*. Retrieved July 24, 2006, from http://www.ed.gov/programs/readingfirst/index.html

VanTassel-Baska, J. (1988). Developing scope and sequence in curriculum: A comprehensive approach. *Gifted Child Today, 11*(3), 29–34 .

VanTassel-Baska, J. (1998). *Excellence in educating gifted and talented learners* (3rd ed.). Denver, CO: Love.

VanTassel-Baska, J. (2003). *Curriculum planning and instructional design for gifted learners*. Denver, CO: Love.

VanTassel-Baska, J., Avery, L. D., Little, C. A., & Hughes, C. E. (2000). An evaluation of the implementation: The impact of the William and Mary units on schools. *Journal for the Education of the Gifted, 23,* 244–272.

VanTassel-Baska, J., Avery, L., Struck, J., Feng, A. X., Bracken, B., Drummond, D., et al. (2003). *Classroom Observation Scale–Revised*. Williamsburg, VA: Center for Gifted Education, The College of William and Mary.

VanTassel-Baska, J., Bass, G., Ries, R., Poland, D., & Avery, L. (1998). A national pilot study of science curriculum effectiveness for high ability students. *Gifted Child Quarterly, 42,* 200–211.

VanTassel-Baska, J., Bracken, B., Brown, E., & Feng, A. (2005). *Project Athena year three student data report.* Williamsburg, VA: Center for Gifted Education, The College of William and Mary.

VanTassel-Baska, J., Bracken, B., Feng, A., & Brown, E. (in press). A longitudinal study of reading comprehension and reasoning ability of students in elementary Title I schools. *Journal for the Education of the Gifted*.

VanTassel-Baska, J., Feng, A., Brown, E., Baytops, J., Henshon, S., & Bai, W. (2002). *An evaluation study of the Greenville, South*

Carolina challenge program. Williamsburg, VA: Center for Gifted Education, The College of William and Mary.

VanTassel-Baska, J., Johnson, D. T., Hughes, C. E., & Boyce, L. N. (1996). A study of the language arts curriculum effectiveness with gifted learners. *Journal for the Education of the Gifted, 19*, 461–480.

VanTassel-Baska, J., Quek, C., & Feng, A. X. (2007). Developing structured observation scales for instructional improvement in classrooms accommodating gifted learners. *Roeper Review, 29,* 84–92.

VanTassel-Baska, J., & Stambaugh, T. (2006). Project Athena: A pathway to advanced literacy development for children of poverty. *Gifted Child Today, 29*(2), 58–63.

VanTassel-Baska, J., Zuo, L., Avery, L. D., & Little, C. A. (2002). A curriculum study of gifted student learning in the language arts. *Gifted Child Quarterly, 46,* 30–44.

Villaume, S. K., & Brabham, E. G. (2002). Comprehension instruction: Beyond strategies. *The Reading Teacher, 55,* 672–676.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric characteristics of assessment procedures. In J. R. Graham, J. A. Naglieri, & I. B. Weiner (Eds.), *Handbook of psychology: Volume 10: Assessment psychology* (pp. 43–66). Hoboken, NJ: John Wiley & Sons.