RONALD C. MARTELLA and KATHLEEN M. WALDRON-SOLER, Eastern Washington University

# Language for Writing *Program Evaluation*

**Abstract: This evaluation investigated the effects of the *Language for Writing* program. Ten classrooms were involved including 201 students at the beginning of the project. Posttest scores were obtained for 126 students. This evaluation was conducted over 2 years. Evaluation I was conducted over a 5-month period from January 2003 to May 2003. Evaluation II was conducted over a 9-month period from September 2003 to May 2004. Results from Evaluation I found that students in general and special education made statistically and educationally significant improvements in their writing performance. The results from Evaluation II replicated those of the 1st year; that is, all students in general and special education made statistically and educationally significant improvements in their writing skills. There was also some evidence that English language learners experienced improved performance across both evaluations. Additionally, it was found that the length of lesson had decreased and teacher satisfaction with lesson length had improved as compared to the 1st year. In general, teachers were pleased with all aspects of the program.**

Writing is perhaps the most complex of all the language skills we need to teach to students (Bain, Bailet, & Moats, 1991; Hall, Salas, & Grimes, 1999; Harris, Schmidt, & Graham, 1997). Good writing means that writers must take on two roles simultaneously—namely the roles of author and secretary (Dixon, Isaacson, & Stein, 2002). Unfortunately, these two roles are not typically taught in an explicit manner. Without formalized instruction and practice throughout much of the school day, students are not likely to become better writers (Dixon et al.). According to Graves (1985), children need to practice writing several times a week to see any significant change in the quality of their writing. Thus, if we want to improve the writing skills of students, we need to teach these skills early and explicitly and have our students write more.

Writing is critical to successful school performance. Students are required to use writing as a means of communication with themselves (e.g., taking notes from class lectures and discussions) and others (e.g., writing answers to questions posed by the teacher). Students also use writing to facilitate the learning process. That is, students use their writing skills to practice skills introduced in the classroom. For example, students are often taught to write information on note cards and to quiz themselves prior to tests. Students are also asked to demonstrate skills they have learned through writing (e.g., papers, tests, or homework; Fredrick & Steventon, 2004).

Writing is also important because of its relationship to reading. Tierney and Shanahan (1991) found that reading and writing are related; however, the nature of this relationship is not completely clear. What is clear is that students who have difficulty in their written expression often have difficulties in their reading (Isaacson, 1994). Adams (1990) suggested that students' reading influences their writing. Snow, Burns, and Griffin (1998) also

*Journal of Direct Instruction, Vol. 5, No. 1, pp. 81–96.* Address correspondence to Ronald Martella at rmartella@ewu.edu

noted the need for increased time spent on reading and writing activities.

Demands on writing continue to increase in upper elementary, middle school, high school, and college (Harris & Graham, 1996). Recognizing the relationship between writing and successful school performance, high-stakes state tests often include a writing component. According to Fredrick and Steventon (2004), "Forty-nine of fifty states require a measure of writing competency for high school graduation or include writing assessments as part of statewide testing" (p. 142). Fredrick and Steventon also indicate that the SAT will include a writing component that is one third of the total SAT score in 2005. Beyond the public school system, writing continues to be an ever-increasing skill used in the workplace (Agnew, 1992). For example, many jobs require writing reports, taking notes related to job activities, and/or communicating through email with colleagues and/or other concerned parties.

Given these ever-increasing writing demands, instruction in writing skills is too important to leave to chance. According to Fredrick and Steventon (2004), "The importance of writing begins at an early age and continues for a lifetime. As a literate society we rely on writing as an effective means of communication in all walks of life. Because we rely on writing as a major means of communication in our society, writing instruction is critical" (p. 140).

Therefore, it seems prudent to teach young children basic writing skills early and well when they are in elementary school. One such program that was developed to teach early writing skills was the *Distar Language III* program (Engelmann & Osborn, 1987). This program was revised into the new *Language for Writing* program (Engelmann & Osborn, 2003). The *Language for Writing* program is designed for second- through fifth-grade students who have been through the *Language for Learning* (Engelmann & Osborn, 1999) and *Language for Thinking* programs (Engelmann & Osborn, 2002). However, the *Language for Writing* program can also be used with students who have not yet completed the first two language programs if their placement test scores indicate they are ready for this program. As stated by Waldron-Soler and Osborn (2004), "Students placed in the program should be reading and writing at an end of second-grade or beginning of third-grade level, and have adequate knowledge of basic spoken school English" (p. 73). Older students can also be placed in the program if they possess the aforementioned skills and pass the placement test.

The purpose of this paper is to describe a program evaluation of the *Language for Writing* program. The overall program evaluation was completed in two parts. Evaluation I was completed from January 2003 to May 2003. Evaluation II was completed from September 2003 to May 2004. The purpose of this program evaluation was two-fold. First, the improvement in student performance during the implementation of the *Language for Writing* program was assessed. Second, suggested areas for revision in the *Language for Writing* program were made based on feedback during Evaluation I. The evaluation of the *Language for Writing* program followed the program evaluation guidelines outlined by Martella, Nelson, and Marchand-Martella (1999).

# *Method*

## Classrooms

*Evaluation I.* Six classrooms served as Evaluation I sites. All students in Classrooms 1–5 were in the second grade in general education classrooms. Classroom 1 (24 students; 20% Hispanic, 5% Asian; 30% free or reduced-price lunch) was located in the Pacific Northwest. Classroom 2 (17 students; 70% Hispanic, 30% Caucasian), Classroom 3 (17 students; 95% African American; 100% free or reduced-price lunch), and Classroom 4 (18 students; 30% African American; 30–40%

Hispanic; 100% free or reduced-price lunch) were located in the Southwest. Classroom 5 (25 students; 100% African American; 99% free or reduced-price lunch) was located in the Midwest. Classroom 6 (16 students in Grades 3 through 5; 25% African American, 6% Hispanic; 44% free or reduced-price lunch) was located in the South. All students in Classroom 6 received special education services in a self-contained cross-categorical class. (Note: Statistics for Classroom 4 were discarded from analyses due to an absence of posttest scores.)

*Evaluation II.* Four classrooms participated in Evaluation II. Classroom 7 (10 third-grade through fifth-grade resource-room students; 95% Caucasian; 19% free or reduced-price lunch) was located in the Pacific Northwest. Classroom 8 (41 third-grade students; 95% African American; 90% free or reduced-price lunch) was located in the Midwest. Classroom 9 (16 third-grade students; 73% Hispanic, 12.6% Caucasian; 100% free or reduced-price lunch) was located in the West. Classroom 10 (17 second-grade students; 70% Hispanic, 30% Caucasian; 100% free or reduced-price lunch) was located in the Southwest. (Note: Statistics for Classroom 10 were discarded from analyses due to an absence of posttest scores.)

## *Language for Writing*

*Language for Writing* is a 140-lesson Direct Instruction program designed for students in Grades 2 through 5. The goal of the program is to help students learn to communicate effectively through spoken and written language. Teachers in five of the Evaluation I sites (Classrooms 1 through 5) taught up to 70 lessons (50%) of the program from February 2003 to May 2003. Students in Classroom 6 (Evaluation I) received all 140 lessons from January 2003 to May 2004. Students in three of the Evaluation II sites (Classrooms 7, 8, and 10) received at least 70 lessons while one classroom (Classroom 9) received all 140 lessons from September 2003 to May 2004.

### Assessments

Assessments conducted throughout the evaluation included the Test of Written Language—3 (TOWL—3; Hammill & Larsen, 1996), documentation of specific student performance such as the number of errors per lesson and mastery-test performance, documentation of the length of each lesson across program evaluation sites, lesson ratings, a social-validity survey, and a curriculum-based measure of written language (Classroom 6 only).

*Test of Written Language—3 (TOWL—3).* The TOWL—3 was provided to measure student gains in writing performance. For Evaluations I and II, the TOWL—3 was provided as a pretest (Form A) and posttest (Form B). Form A was used again only for Classroom 6 in May of 2004 (Posttest 2). The alternate forms reliability for the TOWL—3 is above .80.

Twenty-seven percent ($N = 86$) of all of the assessments given ($N = 319$) were randomly selected for interscorer agreement. (Note: There were 215 assessments included in the program evaluation analysis due to dropping students who did not have pre- and posttest assessments completed.) Interscorer agreement ranged from 92.3% for Style to 98.1% for Spelling. The agreement percentages for composite scores were 99.5% for Contrived Writing, 94.6% for Spontaneous Writing, and 97.6% for Overall Writing.

*Student errors, lesson duration, lesson ratings, mastery-test performance, social-validity survey, and curriculum-based measure.* For each lesson, teachers documented student errors and the duration of the lesson. Every fifth lesson, lesson ratings and teacher responses were recorded on a teacher response form. Every 10 lessons, mastery tests were scored for each student. At the end of each school year, the classroom teachers completed a social-validity survey to

determine their overall satisfaction with the program. A curriculum-based measure was constructed based on guidelines set by Shapiro (1996) and administered to Classroom 6 immediately after the summer of 2003 and after the program was completed (May 2004). The measure included the following parts: written expression, mechanics, and quality evaluation.

# *Results*

Descriptive and inferential statistical analyses were conducted on the TOWL—3 data. Classrooms 4 and 10 were excluded from the analyses due to the absence of posttest assessment scores.

## TOWL—3

*Descriptive data.* As shown in Table 1, improvements were noted for Classrooms 1, 2, 3, and 5 from pre- to posttest assessment on all eight subtests as well as on the three composite scores. Students in Classroom 6 showed an average increase from the pretest to Posttest 1 across all eight subtests. However, there were decreases seen in six of the eight subtests from Posttest 1 to Posttest 2. (Note: The mean for each of the subtests for the normative sample is 10 with a standard deviation of 3.)

For the composite scores, improvements were seen in Classrooms 1, 2, 3, and 5. For Classroom 6, there were increases in all composites from the pretest to Posttest 1. However, there were increases in all three composites from Posttest 1 to Posttest 2. (Note: The average for the normative sample on composite scores is 100 with a standard deviation of 15.)

For Evaluation II, there were increases from the pretest to the posttest for Classroom 8 across all eight subtests (see Table 2). For Classrooms 7 and 9, there were increases across all of the subtests with the exception of Sentence Combining.

As with Evaluation I, improvements were seen in every classroom across the composite scores. For Evaluation II (Classrooms 8 and 9), improvement was seen from scoring below the normative average to scoring near or above the average. For Classroom 7, the students made large gains (at least 16.2 points).

*Inferential statistics and effect sizes.* Table 3 shows the pre- and posttest paired sample *t*-test results for the three composite scores for Classrooms 1, 2, 3, and 5. Only those students who had pretest and posttest scores were included in the analysis (Classrooms 1, 2, 3, and 5: $N = 72$ for Contrived Writing; $N = 75$ for Spontaneous Writing; and $N = 71$ for Overall Writing).

There was an overall improvement from the pretest to the posttest of 6.65 points (a .45 change in effect size compared to the normative sample) for Contrived Writing. This result reached statistical significance beyond the .000 level. For Spontaneous Writing, the pretest to posttest gain for the students in Classrooms 1, 2, 3, and 5 was 19.32 points with an effect size gain of 1.29 compared to the normative sample mean. The pretest to posttest change was statistically significant beyond the .000 level. For Overall Writing, the pretest to posttest change for Classrooms 1, 2, 3, and 5 was 12.31 (effect size change of .82 compared to the normative sample). The change from pretest to posttest was statistically significant beyond the .000 level.

As stated previously, Classrooms 1, 2, 3, and 5 were combined for the analysis in Evaluation I. Classroom 6 was analyzed separately. This site was a special education classroom involving older students with disabilities (see Table 4). Only those students who had pretest and posttest scores were included in the analysis ($N = 15$ for pretest to Posttest 1 analyses; $N = 10$ for pretest to

Posttest 1 and Posttest 1 to Posttest 2 analyses). As with Classrooms 1, 2, 3, and 5, Classroom 6 demonstrated improvements. However, there was a nonstatistically significant change for Contrived Writing. Classroom 6 increased from the pretest to Posttest 1 by 4.2 points. This result represents an effect-size increase of .28 compared to the normative sample. Comparing the pretest to Posttest 2 scores revealed little change in student performance (i.e., .40 of a point or an effect size of .03 compared to the normative sample). This change was not statistically significant. Additionally, there was a negative change from Posttest 1 to Posttest 2 of 3.50 points or a loss of .23 of a standard

**Table 1**

*Pretest and Posttest Scores for TOWL—3 Subtests and Composite Scores Across Classrooms During Evaluation I*

| Lessons completed | Classroom 1 40 | | Classroom 2 70 | | Classroom 3 70 | | Classroom 5 50 | | Classroom 6 140 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pre-test | Post-test | Pre-test | Post-test | Pre-test | Post-test | Pre-test | Post-test | Pre-test | Post-test 1 | Post-test 2 |
| *Subtest* | | | | | | | | | | | |
| Vocabulary | 10.6 | 11.3 | 9.7 | 10.5 | 8.7 | 10.5 | 9.2 | 9.9 | 7.5 | 8.1 | 7.3 |
| Spelling | 10.0 | 10.7 | 9.5 | 9.9 | 8.9 | 9.9 | 8.8 | 9.0 | 6.6 | 6.8 | 6.8 |
| Style | 8.2 | 8.6 | 7.8 | 9.2 | 7.8 | 10.8 | 6.4 | 7.7 | 6.3 | 6.5 | 7.6 |
| Logical Sentences | 8.8 | 9.0 | 7.1 | 8.8 | 5.6 | 8.2 | 5.4 | 8.0 | 4.9 | 6.5 | 4.4 |
| Sentence Combining | 10.5 | 10.6 | 9.9 | 10.6 | 9.2 | 10.1 | 9.3 | 10.5 | 7.3 | 7.7 | 6.6 |
| Contextual Conventions | 8.9 | 11.8 | 8.1 | 10.7 | 8.5 | 11.2 | 8.0 | 9.6 | 6.1 | 10.0 | 7.5 |
| Contextual Language | 8.2 | 10.9 | 5.7 | 10.2 | 5.8 | 10.1 | 5.2 | 9.0 | 4.7 | 8.9 | 7.6 |
| Story Construction | 8.5 | 11.0 | 6.8 | 10.8 | 7.1 | 10.7 | 7.0 | 9.7 | 5.7 | 9.8 | 8.7 |
| *Composites* | | | | | | | | | | | |
| Contrived Writing | 97.2 | 99.9 | 91.6 | 97.7 | 86.4 | 99.4 | 84.7 | 92.0 | 76.1 | 80.3 | 76.2 |
| Spontaneous Writing | 90.6 | 107.7 | 79.8 | 102.0 | 81.5 | 104.2 | 78.8 | 96.3 | 70.9 | 95.6 | 86.9 |
| Overall Writing | 94.4 | 103.2 | 86.5 | 99.5 | 84.1 | 101.4 | 81.5 | 94.2 | 73.1 | 85.6 | 79.6 |

deviation compared to the normative sample. This difference was statistically significant at the .05 level.

There was a much larger change in Spontaneous Writing for Classroom 6. Classroom 6 students had a pretest to posttest change of 24.73 points. This represents a change in effect size of 1.65 compared to the normative sample. This result was statistically significant beyond the .000 level. When comparing the pretest performance to Posttest 2 scores, the changes were also statistically significant beyond the .000 level. There was a difference of 11.8 points, indicating an effect size improvement of .79 compared to the normative sample. Unfortunately, there was a statistically significant ($p < .02$) decrease in scores from Posttest 1 to Posttest 2 of 9.3 points. This decrease represented a change in the effect size of –.62 compared to the normative sample.

Finally, for Overall Writing there was a change from the pretest to Posttest 1 of 12.54. The effect-size change was .84 compared to the normative sample. This result was statistically significant beyond the .000 level. There was a statistically nonsignificant change from the pretest to Posttest 2 assessment. However, there was a 5.10 point increase representing an

**Table 2**

*Pretest and Posttest Scores for TOWL—3 Subtests and Composite Scores Across Classrooms During Evaluation II*

| Lessons Completed | Classroom 7 70 | | Classroom 8 90 | | Classroom 9 130 | |
|---|---|---|---|---|---|---|
| | Pretest | Posttest | Pretest | Posttest | Pretest | Posttest |
| *Subtest* | | | | | | |
| Vocabulary | 7.4 | 9.0 | 9.0 | 9.5 | 8.9 | 10.3 |
| Spelling | 7.1 | 7.8 | 9.1 | 10.0 | 9.6 | 9.7 |
| Style | 5.8 | 8.5 | 8.8 | 10.3 | 7.5 | 10.9 |
| Logical Sentences | 5.4 | 6.5 | 7.0 | 7.9 | 7.0 | 8.1 |
| Sentence Combining | 7.9 | 7.3 | 9.1 | 9.3 | 9.5 | 8.1 |
| Contextual Conventions | 6.9 | 10.3 | 8.1 | 13.6 | 9.3 | 13.4 |
| Contextual Language | 5.6 | 8.2 | 6.8 | 10.8 | 8.7 | 11.8 |
| Story Construction | 6.0 | 9.7 | 7.4 | 16.0 | 8.2 | 11.5 |
| *Composites* | | | | | | |
| Contrived Writing | 79.0 | 85.2 | 90.4 | 95.8 | 91.0 | 98.9 |
| Spontaneous Writing | 75.8 | 96.2 | 83.3 | 111.7 | 93.1 | 114.1 |
| Overall Writing | 72.7 | 89.0 | 87.3 | 102.1 | 91.4 | 105.6 |

effect-size improvement of .34 compared to the normative sample. Again, there was a decrease from Posttest 1 to Posttest 2 of 5.9 points, representing a change in effect size of –.39 compared to the normative sample. This negative change was statistically significant beyond the .01 level.

For Evaluation II (combined Classrooms 8 and 9; $N = 30$), there was a pretest to posttest change of 6.5 points (an effect-size change of .43 compared to the normative sample; see Table 5) for Contrived Writing. This result reached statistical significance beyond the .01 level. For Spontaneous Writing, the pretest to posttest change was 24.9 points. This change represented an effect size change of 1.66 compared to the normative sample. The pretest to posttest change was statistically significant beyond the .000 level. For Overall Writing, the pretest to posttest change was 14.5 points. This change reflected an effect-size increase of .97 compared to the normative sample. The change from pretest to posttest was statistically significant beyond the .000 level.

Classroom 7 ($N = 6$) was analyzed separately given that it was a special education classroom including students with disabilities. For Contrived Writing, Classroom 7 student scores increased from the pretest to the posttest by 6.2 points. This result represents an effect-size change of .41 compared to the normative sample. This change was not statistically significant beyond the .05 level. For Spontaneous Writing, Classroom 7 students had a pretest to posttest change of 20.4 points. This change represents an effect-size change of 1.36 compared to the normative sample. This result was statistically significant beyond the .02 level. For Overall Writing, there was a pretest to posttest change in Classroom 7 student scores of 16.3 points, or an effect-size change of 1.09 compared to the normative sample. This result was not statistically significant beyond the .05 level.

*English language learners.* As shown in Table 6, there were three identified English language learners (ELL) in Evaluation I and two ELL students in Evaluation II. ELL students in Evaluation I completed 70 lessons of the program. As shown in the table, there were

**Table 3**

*Pretest and Posttest Means, Standard Deviations, Effect Sizes, and* t*-test Results for Composite Scores for Classrooms 1, 2, 3, and 5 for Evaluation I*

| | Pretest | | | Posttest | | | Statistics | |
|---|---|---|---|---|---|---|---|---|
| Composites | *M* | *SD* | Effect Size | *M* | *SD* | Effect Size | Difference | *t* test |
| *Classrooms 1, 2, 3, & 5* | | | | | | | | |
| Contrived Writing ($df = 71$) | 90.31 | 11.98 | –.65 | 96.96 | 11.43 | –.20 | 6.65 | $p < .000$ |
| Spontaneous Writing ($df = 74$) | 82.79 | 9.64 | –1.15 | 102.11 | 12.96 | .14 | 19.32 | $p < .000$ |
| Overall Writing ($df = 70$) | 87.10 | 11.09 | –.86 | 99.41 | 11.91 | –.04 | 12.31 | $p < .000$ |

improved performances for all three students for Contrived Writing, Spontaneous Writing, and Overall Writing. The average increase from pretest to posttest assessments for Contrived Writing was 7.4 points, or .49 of a standard deviation compared to the normative sample. The average increase for Spontaneous Writing was 11.3 points, or .75 of a standard deviation compared to the normative sample. Finally, the average increase for Overall Writing was 9 points, or .60 of a standard deviation compared to the normative sample.

**Table 4**

*Pretest and Posttest Means, Standard Deviations, Effect Sizes, and* t*-test Results for Composite Scores for Classroom 6 for Evaluation II*

| | Pretest | | | Posttest 1 | | | Statistics (Pretest to Posttest 1) | |
|---|---|---|---|---|---|---|---|---|
| Composites | *M* | *SD* | Effect Size | *M* | *SD* | Effect Size | Difference | *t* test |
| *Classroom 6* | | | | | | | | |
| Contrived Writing (*df* = 14) | 76.06 | 10.76 | –1.60 | 80.26 | 11.73 | –1.32 | 4.20 | $p < .100$ |
| Spontaneous Writing (*df* = 14) | 70.87 | 14.18 | –1.94 | 95.60 | 11.23 | –.29 | 24.73 | $p < .000$ |
| Overall Writing (*df* = 14) | 73.06 | 11.38 | –1.80 | 85.60 | 10.94 | –.96 | 12.54 | $p < .000$ |

**Table 4, continued**

*Pretest and Posttest Means, Standard Deviations, Effect Sizes, and* t*-test Results for Composite Scores for Classroom 6 for Evaluation II*

| | Pretest | | | Posttest 1 | | |
|---|---|---|---|---|---|---|
| Composites | *M* | *SD* | Effect Size | *M* | *SD* | Effect Size |
| *Classroom 6* | | | | | | |
| Contrived Writing (*df* = 9) | 75.80 | 13.24 | –1.61 | 79.70 | 12.35 | –1.35 |
| Spontaneous Writing (*df* = 9) | 75.10 | 13.67 | –1.66 | 96.20 | 13.32 | –.25 |
| Overall Writing (*df* = 9) | 74.50 | 13.10 | –.170 | 85.50 | 12.17 | –.97 |

For Evaluation II, there were five identified ELL students. Unfortunately, pretest and posttest data were available for only two of these students. These ELL students completed at least 123 lessons of the program. As shown in Table 6, there were large improvements on all three composites. The largest improvement was seen in Spontaneous Writing. The two students combined showed the following increases from pretest to posttest assessments: 9 points (.60 of a standard deviation compared to the normative sample) for Contrived Writing, 26.5 points (1.77 of a standard deviation compared to the normative sample) for Spontaneous Writing, and 17 points (1.13 of a standard deviation compared to the normative sample) for Overall Writing.

## Curriculum-Based Measure

On the assessment completed after the summer, 7 of the 10 students who took the posttest in Classroom 6 demonstrated improved performance for Written Expression (see Table 7). For Quality Evaluation, six students demonstrated improved performance. Seven of the 10 students demonstrated improved performance on Mechanics. Overall (with the exception of Student 17), there was improved performance of six more words for Written Expression, a 3.5% increase in Quality Evaluation, and a 3.3% improvement in Mechanics.

## Lesson Ratings

*Likert-like ratings.* Ratings ranged from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*) for each of five lessons across all classrooms for Evaluations I and II. For Evaluation I, there was an overall decrease in ratings for the statement, "The lessons were easy to teach," from an average high of 4.7 for Lessons 1 to 5, to an average low of 3.3 for Lessons 31 to 35. In other words, the teachers felt the lessons became more difficult to teach as the program progressed. A decrease in teacher ratings was observed for the statement, "The exercises were at the appropriate level of difficulty," from an average high of 4.2 for Lessons 6 to 10, to an average low of 3.0 for Lessons 66 to 70. Also, the statement, "Students responded enthusiastically to the lessons," resulted in decreased ratings for many of the later lessons.

| Posttest 2 | | | Statistics (Pretest to Posttest 2) | | Statistics (Posttest 1 to Posttest 2) | |
|---|---|---|---|---|---|---|
| *M* | *SD* | Effect Size | Difference | *t* test | Difference | *t* test |
| 76.20 | 12.01 | –1.59 | .40 | $p > .200$ | –3.50 | $p < .050$ |
| 86.90 | 9.84 | –.87 | 11.80 | $p < .050$ | –9.30 | $p < .020$ |
| 79.60 | 10.36 | –1.36 | 5.10 | $p < .200$ | –5.90 | $p < .010$ |

Statement 4, "The lessons were the right length for the class period," was perhaps the best indicator of the concern for lesson length. Ratings were low for the later lessons (i.e., decreased from an average high of 4.8 for Lessons 1 to 5, to an average low of 2.0 for Lessons 66 to 70), indicating that the teachers did not believe the lessons were of the right length (especially with regard to the writing tasks). The teacher for Classroom 6 reported that the lesson length did improve over time.

For Evaluation II, responses to the statement, "The lessons were easy to teach," had overall average ratings of 4.2. For Classroom 9, the ratings averaged 4.9 (range 4.0, Lessons 136 to 140, to 5.0, Lessons 101 to 135). For the statement, "The exercises were at the appropriate level of difficulty for my students," the average rating across the first 100 lessons was 3.5. For Classroom 9, the average rating over the final 40 lessons was 3.0. However, there were low ratings of 2.0 for Lessons 111 to 125. For the statement, "Students responded enthusiastically to the lessons," the overall average rating was 3.0 across the first 100 lessons. For Classroom 9, the average rating was 3.4 (range 2.0, Lessons 116 to 120, to 5.0, Lessons 136 to 140). For statement 4, "The lessons were the right length for the class period," the overall average ratings across the first 100 lessons were generally favorable (overall average 3.0). However, overall average ratings were lower for Lessons 36 to 45 (overall average of 2.3) to Lessons 76 to 80 (overall average of 1.0). Therefore, there was still a concern over lesson length primarily for Lessons 36 to 80. Classroom 9 provided ratings for Lessons 100 to 140. The overall average rating was 3.7.

**Table 5**

*Pretest and Posttest Means, Standard Deviations, Effect Sizes, and* t*-test Results for Composite Scores for Classrooms 7, 8, and 9 for Evaluation II*

| | Pretest | | | Posttest | | | Statistics | |
|---|---|---|---|---|---|---|---|---|
| Composites | *M* | *SD* | Effect Size | *M* | *SD* | Effect Size | Difference | *t* test |
| *Classrooms 8 & 9* | | | | | | | | |
| Contrived Writing ($df = 29$) | 90.7 | 8.7 | –.62 | 97.2 | 10.0 | –.19 | 6.5 | $p < .010$ |
| Spontaneous Writing ($df = 29$) | 87.9 | 10.8 | –.81 | 112.8 | 10.0 | .85 | 24.9 | $p <. 000$ |
| Overall Writing ($df = 29$) | 89.2 | 9.0 | –.72 | 103.7 | 9.1 | .25 | 14.5 | $p < .000$ |
| *Classroom 7* | | | | | | | | |
| Contrived Writing ($df = 5$) | 79.0 | 8.1 | –1.40 | 85.2 | 11.9 | –.99 | 6.2 | $p > .200$ |
| Spontaneous Writing ($df = 5$) | 75.8 | 6.0 | –1.61 | 96.2 | 13.3 | –.25 | 20.4 | $p < .020$ |
| Overall Writing ($df = 5$) | 72.7 | 15.5 | –1.82 | 89.0 | 12.6 | –.73 | 16.3 | $p < .100$ |

## Social Validity Survey

The social validity survey data for Evaluations I and II showed that teachers were generally pleased with the program. They saw improvements in student performance and liked the sequencing of the skills taught. For Evaluation I, four teachers indicated they would continue to use the program in the future. Perhaps the most enthusiastic teacher was from Classroom 6 (special education teacher). Finally, the most significant feedback from the social validity survey had to do with the lesson length. Overall, if the lesson length could be shortened, the teachers saw this program as an excellent writing program. For Evaluation II, the program was not seen as difficult to implement. There was also an overall low concern for lesson length. Three teachers indicated they would use the program in the future. However, there was concern that the program was more appropriate for second-grade students or for students with disabilities in the higher grades.

# *Discussion*

Results from Evaluations I and II showed that students in general and special education made statistically and educationally significant improvements in their writing performance. There was also some evidence that ELL students benefited from the program.

Writing skills were directly assessed by the TOWL—3. For Evaluation I, general education students in Classrooms 1, 2, 3, and 5 showed improvements in Contrived and Spontaneous Writing skills. Students in a special education classroom (Classroom 6) also experienced improved skills across Contrived and Spontaneous Writing composites. Overall writing scores across all five classrooms that provided pre- and posttest results were also impressive. If an effect size of .25 is considered educationally significant (see Adams & Engelmann, 1996), there were educationally significant improvements across all five classrooms that provided pre- and posttest data.

**Table 6**

*ELL Student Performance on Evaluations I and II*

| | Pretest | | | Posttest | | |
|---|---|---|---|---|---|---|
| Evaluation I Students | Contrived Writing | Spontaneous Writing | Overall Writing | Contrived Writing | Spontaneous Writing | Overall Writing |
| 1 | 68 | 72 | 69 | 78 | 85 | 80 |
| 2 | 97 | 79 | 90 | 105 | 89 | 99 |
| 3 | 73 | 70 | 70 | 77 | 81 | 77 |
| *M* | 79.3 | 73.7 | 76.3 | 86.7 | 85.0 | 85.3 |

| | Pretest | | | Posttest | | |
|---|---|---|---|---|---|---|
| Evaluation II Students | Contrived Writing | Spontaneous Writing | Overall Writing | Contrived Writing | Spontaneous Writing | Overall Writing |
| 1 | 78 | 83 | 79 | 78 | 91 | 83 |
| 2 | 68 | 74 | 69 | 86 | 119 | 99 |
| *M* | 73.0 | 78.5 | 74.0 | 82.0 | 105.0 | 91.0 |

Improvements for Classrooms 1, 2, 3, and 5 ranged from an effect-size change of .45 (Contrived Writing) to 1.29 (Spontaneous Writing) compared to the TOWL—3 normative sample. The Overall Writing composite showed an effect-size change of .82 compared to the normative sample.

The results of the 2-year evaluation replicated and extended those of the 1st year. The improvement in performance of Classrooms 8 and 9 compared to the normative sample was .43 for Contrived Writing, 1.67 for Spontaneous Writing, and .97 for Overall Writing.

There were also improvements in the special education classrooms. For Classroom 6, the Contrived Writing composite effect size improved from pretest to Posttest 1 by .28, while the Spontaneous Writing composite effect size improved by 1.65 compared to the normative sample. The Overall Writing composite effect size improved by .84. Therefore, the changes that were seen in this evaluation

**Table 7**

*Curriculum-Based Measure for Classroom 6*

| | After Summer Assessment (after 70 lessons) | | | End of Program Assessment (after 140 lessons) | | |
|---|---|---|---|---|---|---|
| Student | Written Expression | Quality Evaluation | Mechanics | Written Expression | Quality Evaluation | Mechanics |
| 1 | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported |
| 2 | –30 | 93.9% | 87.5% | –19 | 100% | 97.0% |
| 3 | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported |
| 4 | –43 | 100% | 87.5% | –6 | 100% | 100% |
| 5 | –11 | 97.0% | 81.3% | –27 | 100% | 97.0% |
| 6 | –29 | 81.8% | 100% | –25 | 100% | 90.9% |
| 7 | –9 | 97.2% | 81.3% | –9 | 100% | 93.9% |
| 8 | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported |
| 9 | –41 | 93.9% | 87.5% | –26 | 100% | 93.9% |
| 10 | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported |
| 11 | –22 | 93.9% | 100% | –14 | 93.8% | 90.9% |
| 12 | –10 | 87.9% | 100% | –9 | 100% | 75.8% |
| 13 | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported |
| 14 | –9 | 100% | 93.8% | –11 | 93.8% | 97.0% |
| 15 | Not reported | Not reported | Not reported | Not reported | Not reported | Not reported |
| 16 | –29 | 87.9% | 81.3% | –27 | 81.3% | 97.0% |
| 17 | –32 | 100% | 81.3% | Not reported | Not reported | Not reported |
| Average* | –23.3 | 93.4% | 90.0% | –17.3 | 96.9% | 93.3% |

*Excluding Student 17

can be considered educationally significant. Unfortunately, the scores for Classroom 6 regressed from Posttest 1 (covering Lessons 1 to 70) to Posttest 2 (covering Lessons 71 to 140). A likely reason for this decrease is the loss of 5 of the 15 students who provided Posttest 1 scores. An analysis of these five students shows that their improvements from pre- to Posttest 1 assessments were near or higher than the improvement for the entire group. However, the top three students from this group of five had an average increase of 10.7 points for Contrived Writing versus 4.2 for the 15-member group, 35.7 for Spontaneous Writing compared to 24.7 for the 15-member group, and 21.0 for Overall Writing versus 12.5 for the 15-member group. Thus, the decrease seen from Posttest 1 to Posttest 2 may have been due to the loss of these high achievers.

For Classroom 7, the pretest to posttest improvement on the Contrived Writing composite was 6.2 points (.41 of a standard deviation). The improvements for the Spontaneous Writing composite score were 20.4 points (1.36 of a standard deviation), and the Overall Writing composite score improved by 16.3 points (1.09 of a standard deviation). As with Classroom 6, these increases are large and educationally significant.

On the curriculum-based measure completed after the summer, 7 of the 10 students in Classroom 6 who took the pretest demonstrated improved performance for Written Expression. Six of the 10 students demonstrated improved performance for Quality Evaluation, and 7 of the 10 students demonstrated improved performance on Mechanics. Overall (excluding Student 17), there was improved performance of six more words for Written Expression, a 3.5% increase in Quality Evaluation, and a 3.3% improvement in Mechanics.

The performance for Classroom 7 showed a .41 improvement in Contrived Writing compared to the normative sample, 1.36 for Spontaneous Writing, and 1.09 for Overall Writing. Interestingly, the largest improvement was seen in Spontaneous Writing. Students' writing scores on this composite were greater than their performance in Contrived Writing.

The ELL students also showed great improvement across all three composite test areas as a result of having been exposed to the program. It is also important to point out that the identified ELL students demonstrated large improvements in Spontaneous Writing.

The lesson ratings and social validity survey data for Evaluations I and II showed that teachers were generally pleased with the program. They saw improvements in student performance and liked the sequencing of the skills taught. For Evaluations I and II, most teachers indicated they would continue to use the program in the future. Finally, the most significant feedback from the social validity survey had to do with the lesson length. Overall, the lesson length improved from Evaluation I to Evaluation II. For example, the overall lesson duration across 70 lessons was 50 min for Evaluation I and 42 min across 90 lessons for Evaluation II. For Evaluation I, the overall average lesson duration increased across lessons up to Lesson 60. The average lesson duration had a high of 72 min (Lessons 51 to 60). In Evaluation II, the average lesson duration had a high of 53 min (Lessons 61 to 70). Therefore, the highest average duration decreased by 19 min per lesson compared to the highest average duration in Evaluation I.

The importance of these findings cannot be understated. Writing is a critical skill to obtain in a literate society (Fredrick & Steventon, 2004). With the *Language for Writing* program, significant improvements were seen in every area of writing. Essentially, students took on the role of author and secretary (Dixon et al., 2002), showing improvements in writing across Contrived and Spontaneous Writing composites. The results of this evaluation also

support the contention by Graves (1985) that children need to write multiple times per week to see any appreciable change in the quality of their writing.

These results are consistent with the findings of two studies that have included *Distar Language III* in their investigations (the original version of *Language for Writing*). Booth, Hewitt, Jenkins, and Maggs (1979) investigated the long-term effects of a 5-year program in *Distar Language I, II,* and *III* and *Distar Reading I*, *II*, and *III*. Twelve children with mental retardation (8 to 14 years of age at the beginning of the study) gained an average of 34 language-age months in 32 months of instruction. Many of the children were performing at approximately normal third- to fourth-grade levels in language and reading at the conclusion of the study. Gersten and Maggs (1982) also investigated the long-term effects of an intensive 5-year program in *Distar Language I*, *II*, and *III* and *Distar Reading I*, *II*, and *III*. Twelve children with mental retardation ranging in age from 6 years 10 months to 12 years 6 months at the beginning of the study made significant gains on cognitive, language, and reading measures.

Although several interesting and important findings were noted in this evaluation of the *Language for Writing* program, several caveats should be addressed. First, there was no control group in this evaluation. Therefore, definitive cause-and-effect statements cannot be made. In other words, it is not possible to state that the *Language for Writing* program caused the improvements in the students' writing skills. However, to overcome this weakness, the normative sample from the TOWL—3 was used essentially as a contrast group. Based on the comparison with the normative sample, significant changes were noted.

Second, there is a lack of reliability and validity information with regard to teacher reports/responses. However, there were several assessments that were consistent. For example, teacher responses that lessons were taking too long corresponded with the duration data as well as the increased student error rates. Therefore, a form of "triangulation" (Martella et al., 1999) took place that increases the believability of the data.

Third, there was no verification of the implementation of the program. Although there were trainers at each site, there is no information indicating that the program was implemented as intended. In each classroom, all students were placed at the same level at the beginning of the program and taught in a large-group format. Placement test results indicated that there were large differences among students in each classroom. Therefore, it is unknown what the effects of the program would be if implemented in a small-group format using homogeneous groupings. The heterogeneous implementation used in this evaluation is likely responsible for higher error rates on each lesson than would be allowed with a pure implementation. For example, during Evaluation I, the average number of errors increased as the program progressed, indicating that many students were not meeting mastery on the lessons. The overall average of errors across the six classrooms through Lesson 70 was 6.3. Average errors increased from a low of 1.3 errors across Lessons 1 to 10 to a high of 12.7 errors across Lessons 51–60. However, there was a reduction in the number of errors during Evaluation II. The overall average number of errors across the four classrooms up to Lesson 90 was 2.6. The average number of errors increased from a low of .8 errors across Lessons 1 to 10 to a high of 6.0 errors across Lessons 61 to 70.

The heterogeneous groups did not seem to affect the groups as a whole. For Evaluation I, all classrooms were near or above 80% mastery on all mastery tests. Generally, all classrooms in Evaluation II met mastery of 80% or higher on every test with the exception of Classrooms 9 and 10 on Mastery Test 7 (66%). Thus, it is not known if a homogeneous implementation

were provided, whether or not the improvements in students' skills would be even more impressive. However, the implementation seen in this evaluation is likely more representative of the way the program would be implemented in many classrooms (i.e., in a large-group format).

Overall, the results of this program evaluation showed that the implementation of the *Language for Writing* program was correlated with an improvement in the writing skills of students across five classrooms that provided pre- and posttest assessment scores. Additionally, information was obtained regarding student performance per lesson, mastery test performance, lesson duration, teacher views of every five lessons, and overall teacher opinions of the program in general.

As with a program evaluation of this type, there were constraints that prevented the use of an adequate experimental design. Although there were flaws in the experimental design of this evaluation, significant information was gained. The primary purpose of this evaluation was to provide feedback to the program authors and publishers with regard to areas in need of improvement. The extent of this evaluation before the program was completed is important and should be considered for other program evaluations. Authors of programs should continue to refine the program up to the publication date based on program evaluation data gathered during the developmental stage of the program. Doing so will result in an improved version of the program with supporting field test data once the program is published. Thus, this evaluation should be seen as a first step in validating the program. Clearly, there is a need for further research on the *Language for Writing* program using adequate experimental designs to establish a cause-and-effect relationship between the *Language for Writing* program and student gains.

## References


Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Adams, G. L., & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR.* Seattle, WA: Educational Achievement Systems.

Agnew, E. (1992). Basic writers in the workplace: Writing adequately for careers after college. *Journal of Basic Writing, 11*(2), 28–46.

Bain, A. M., Bailet, L. L., & Moats, L. C. (1991). *Written language disorders: Theory into practice.* Austin, TX: Pro-Ed.

Booth, A., Hewitt, D., Jenkins, W., & Maggs, A. (1979). Making retarded children literate: A five-year study. *The Australian Journal of Mental Retardation, 5*(7), 257–260.

Dixon, R. C., Isaacson, S., & Stein, M. (2002). Effective strategies for teaching writing. In E. J. Kame'enui, D. W. Carnine, R. C. Dixon, D. C. Simmons, & M. D. Coyne (Eds.), *Effective teaching strategies that accommodate diverse learners* (2nd ed., pp. 93–119). Upper Saddle River, NJ: Pearson Education.

Engelmann, S., & Osborn, J. (1987). *Distar language III.* Columbus, OH: SRA/McGraw-Hill.

Engelmann, S., & Osborn, J. (1999). *Language for learning.* Columbus, OH: SRA/McGraw-Hill.

Engelmann, S., & Osborn, J. (2002). *Language for thinking.* Columbus, OH: SRA/McGraw-Hill.

Engelmann, S., & Osborn, J. (2003). *Language for writing.* Columbus, OH: SRA/McGraw-Hill.

Fredrick, L. D., & Steventon, C. (2004). Writing. In N. E. Marchand-Martella, T. A. Slocum, & R. C. Martella (Eds.), *Introduction to Direct Instruction* (pp. 140–177). Boston: Allyn and Bacon.

Gersten, R. M., & Maggs, A. (1982). Teaching the general case to moderately retarded children: Evaluation of a five-year project. *Analysis and Intervention in Developmental Disabilities, 2,* 329–343.

Graves, D. H. (1985). All children can write. *Learning Disabilities Focus, 1*(1), 36–43.

Hall, J. K., Salas, B., & Grimes, A. E. (1999). *Evaluating and improving written expression: A practical guide for teachers* (3rd ed.). Austin, TX: Pro-Ed.

Harris, K. R., & Graham, S. (1996). *Making the writing process work: Strategies for composition and self-regulation.* Cambridge, MA: Brookline Books.

Harris, K. R., Schmidt, T., & Graham, S. (1997). Every child can write: Strategies for composition and self-regulation in the writing process. In K. R. Harris, S. Graham, & D. Deshler (Eds.), *Teaching every child every day* (pp. 131–167). Cambridge, MA: Brookline Books.

Hammill, D. D., & Larsen, S. C. (1996). *Test of Written Language-3.* Austin, TX: PRO-ED, Inc.

Isaacson, S. L. (1994). Integrating process, product, and purpose: The role of instruction. *Reading and Writing Quarterly, 10,* 39–62.

Martella, R. C., Nelson, J. R., & Marchand-Martella, N. E. (1999). *Research methods: Learning to become a critical research consumer.* Boston: Allyn and Bacon.

Shapiro, E. S. (1996). *Academic skills problems: Direct assessment and intervention* (2nd ed.). New York: Guilford Press.

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

Tierney, R. J., & Shanahan, T. (1991). Research on the reading-writing relationship: Interactions, transactions, and outcomes. In R. Barr, M. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 246–280). New York: Longman.

Waldron-Soler, K. M., & Osborn, J. (2004). Language. In N. E. Marchand-Martella, T. A. Slocum, & R. C. Martella (Eds.), *Introduction to Direct Instruction* (pp. 66–99). Boston: Allyn and Bacon.

## Acknowledgements

We would like to thank the teachers who participated in this evaluation for their hard work and dedication. Their commitment to their students is inspiring. We would also like to thank the trainers who aided in the implementation of the program at the various sites. Finally, we would like to thank those who provided us feedback and encouragement throughout this evaluation including: Jane Schott, Karen Sorrentino, Linda Carnine, Jean Osborn, J. Ron Nelson, and Nancy Marchand-Martella.