Learning Disabilities: A Contemporary Journal 5(2), 1–29, 2007

The Effects of Modified Classwide Peer Tutoring Procedures on the Generalization of Spelling Skills of Urban Third-Grade Elementary Students

Kazunari Hashimoto
Cheryl A. Utley[1]
Charles R. Greenwood
Carol L. Pitchlyn
The Juniper Gardens Children's Project
Schiefelbusch Institute for Life Span Studies
University of Kansas

A single-subject reversal design with counterbalanced phases across two classrooms was used to measure the effects of peer tutoring on the retention and generalization of spelling words in two third-grade general education classrooms. The results revealed that the mean pretest-posttest gain scores during all the peer tutoring phases of the two classes was 31.3, compared to 20.3 in the baseline phases. However, there was no apparent difference between peer tutoring and baseline phases in terms of the percentage correct on retention measures. Generalization test results showed that peer tutoring resulted in 13 to 18 percentage points higher than teacher-led instruction.

Key Words: Classwide Peer Tutoring, Fidelity of Implementation, Mastery Learning, Retention and Generalization of Spelling Words

One of the most well-studied strategies in spelling instruction is classwide peer tutoring (CWPT; Greenwood, Delquadri, & Carta, 1997; Utley, Mortweet, & Greenwood, 1997). Since its development, CWPT has been implemented with students with various needs and characteristics, such as mild mental retardation (Mortweet et al., 1999), learning disabilities (LD) and attention deficit hyperactivity disorder (ADHD) (Sideridis et al., 1997), English language learners (Greenwood, Arreaga-Mayer, Utley, Gavin, & Terry, 2001), and low-income ethnically diverse students (Maheady & Harper, 1987).

The research literature on the effectiveness of peer tutoring is extensive with more than 500 studies reviewed prior to 1982 (Kalkowski, 2001; Swanson & Hoskyn, 1998). As noted, studies have been conducted across different populations of children with disabilities, on diverse subject matter, and a variety of personality variables (e.g., self-concept). For example, earlier findings have shown that peer tutoring procedures conducted with students with mild mental retardation had an overall effect size (*ES*) of .36 and was more effective than traditional reading instruc-

1. Send correspondence to: Cheryl A. Utley, Ph.D., Juniper Gardens Children's Project, 650 Minnesota Ave., Second Fl., Kansas City, KS 66101; E-mail: cheryl @ku.edu

tion, regardless of classroom setting (Mathes, 1994). More recently, Kunsch, Jitendra, and Sood (2007) found a moderate *ES* (.59) for peer-mediated instruction implemented in classrooms for more than 16 weeks compared to an *ES* of .43 for 4-16 weeks of instruction in students' mathematics performance. Leung, Marsh, and Craven (2002) conducted an updated, comprehensive meta-analysis evaluating the effects of peer tutoring on academic achievement and self-concept. The findings demonstrated that peer-tutoring programs impacted positively on academic achievement (unweighted mean *ES* =0.81, *SD* =0.79; weighted *ES* =0.65, $p<.05$, 95% confidence interval = 0.59-0.71) and self-concept (unweighted mean *ES* =0.82, *SD* =0.80; weighted *ES* =0.88, $p< .05$, 95% confidence interval = 0.69-1.07). In examining the effectiveness of CWPT procedures with children from low-socioeconomic backgrounds, Greenwood (2006) found effect sizes between a low-SES CWPT treatment group versus a low-SES no-treatment group averaged .72, ranging from .37 (math) to .57 (reading), -.83 (a reduction in inappropriate behavior), and 1.41 (academic engagement) in a longitudinal study of elementary schools. Using Cohen's (1988) criteria, these effects are moderate to large in educational significance. At the middle school follow-up, the average *ES* was .44 (a moderate effect), ranging from .35 (language), .39 (reading), and .57 (math) on achievement test measures.

Despite its effectiveness in improving spelling accuracy, the generalization effects of CWPT have been seldom studied. Generalization is considered to occur if a relevant behavior occurs under different, non-trained conditions, when no training or incomplete training is provided (Cooper, Valentine, & Charlton, 1987; Stokes & Baer, 1977). With regard to a spelling instructional strategy, at least two aspects of generalization must be examined: generalization over time and generalization across other writing activities.

Generalization over time (i.e., retention) refers to the extent to which the spelling words taught during instruction are retained in use at a later time. Retention is important because the acquisition of words has very little value if a student cannot produce correct spellings at a later time (Mallette, Harper, Maheady, & Dempsey, 1991). Generalization across other writing activities, in turn, refers to the extent to which spelling words taught during instruction are spelled correctly when used in various other writing forms and activities (McNaughton, Hughes, & Clark, 1994). For the acquisition of spelling words to be truly meaningful, the words should be used in other writing activities, such as compositions and essays (Harper, Mallette, Maheady, Parkes, & Moore, 1993). To date, only three CWPT studies have examined these two aspects of generalization.

The measurement of generalization in CWPT studies has been questioned. Research by Maheady and Harper (1987) and Mallette et al. (1991) demonstrated the effectiveness of CWPT on retention tests for randomly selected words, with the span between instruction and the retention tests ranging from two weeks to three months. In both studies, it was unclear if the students mastered the words beforehand because the words were not screened prior to instruction. Thus, the acquisition of the spelling words might not have been taken into account, which potentially confounded the results. It was also unclear if the tests included words that were not in the repertoire of the students at the end of instruction. If the retention tests included non-acquired word, the incorrect spellings on the retention tests would have

reflected not only failure to retain the words that were acquired during CWPT, but also failure to acquire the words during CWPT, potentially underestimating the effects.

Another CWPT study by Harper et al. (1993) addressed both retention and generalization across other writing activities. The retention tests for randomly selected words were administered twice, one week and 18 days after the last instruction. The sentence dictation tests for all words were administered weekly (one week after instruction). Unlike the other two studies, the spelling words had to be mastered on both the weekly posttests and the generalization tests (i.e., retention test and sentence dictation test). While this measure ensured that the words were in the students' repertoire at the end of instruction, the words considered correct on the generalization tests might have included words that the students already knew prior to CWPT. Thus, the results of the generalization tests might have reflected not only generalization of the words acquired through CWPT, but also generalization of the words acquired prior to CWPT, thereby potentially overestimating the effects. Therefore, in all three studies, generalization measures could have been inaccurate because acquisition was not well considered.

Finally, among these three studies, two were conducted in special education classroom settings (Harper et al., 1993; Mallette et al., 1991). The one other study (Maheady & Harper, 1987) was conducted in a general education classroom setting, but retention was not measured in a systematic way (only 54% of students took the follow-up test). Therefore, the effectiveness of CWPT on generalization in a general education classroom setting is not well established. Also, two of the three studies (i.e., Harper et al., 1993; Mallette et al., 1991) did not employ a control condition (i.e., conventional teacher-led instruction), and the third study did not report retention of spelling words taught during teacher-led instruction. Thus, generalization effects between CWPT and teacher-led instruction were not compared.

The empirical validation of interventions as evidence-based is critical to its implementation, given the federal legislation No Child Left Behind (NCLB; 2001) and the Education Sciences Reform Act (ESRA; 2002). The effectiveness of interventions (e.g., CWPT) using single-case studies has been examined by a number of researchers (e.g., Campbell, 2006; Cohen, 1988; Hershberger, Wallace, Green, & Marquis, 1999; Parker, Hagan-Burke, & Vannest, 2007). According to Busk and Serlin (1992), measurement of *ES* addresses the change in the level of behavior between baseline and the intervention phase, as indicated by a standardized mean difference (*SMD*). *SMD*s are calculated as the ratio of the difference between the mean of the baseline data points and the mean of the treatment data points to the standard deviation of the data in the baseline phase (Cooper et al., 2000).

Given the lack of studies, the current study examined the effectiveness of CWPT as a spelling intervention and generalization and retention of spelling words in two third-grade general education classrooms. A necessary precursor to these analyses was an examination of the (a) overall fidelity of CWPT implementation and (b) accuracy and initial mastery of spelling words. Thus, the following research questions were addressed:

1. What was the magnitude and variation in fidelity of CWPT implementation?
2. What was the magnitude and variation in students' increased spelling accuracy?

3. What was the magnitude and variation in students' generalization of mastered words on a sentence dictation task one week later?
4. What was the magnitude and variation in students' retention of mastered words two weeks later?
5. What were the overall mean *ES*s of the CWPT intervention across baseline and intervention phases?

5a. What were the *ES*s of the CWPT intervention on students' generalization of mastered vs. loss words on a sentence dictation task one week later?

5b. What was the *ES* of the CWPT intervention on students' retention of mastered words two weeks later?

Method

Participants

Two teachers and 40 students participated in the study. Both teachers taught general education classes at an elementary school with an ethnically diverse student body (78.0% African American, 18.7% Hispanic/Latino, 1.9% Caucasian, 1.0% Asian American, and 0.5% Native American) located in urban area of a major midwestern city. The students participated in the free or reduced-price lunch program at the school. Parent permission letters were distributed in each classroom at the beginning of the study.

In the first classroom, Ms. A originally had 10 students in her fourth-grade classroom, but that number increased to 14 by the end of the study. At the beginning of the school year, students' ages ranged from 8.6 to 10.4 years old, with a mean age of 9.7. In the second classroom, Mr. B had 18 students in his third-grade general education classroom at the beginning of the study, but the number decreased to 17 students. At the beginning of the school year, students' ages ranged from 8.5 to 9.4 years old, with a mean age of 9.0. Data on all students in both classrooms, including those who moved in and out during the study, except for two students with learning disabilities whose grades were not determined by Mr. B, were used in analyses. According to anecdotal reports by the teachers and informal observations by the researcher, the students in both classes had difficulty staying on task during instruction. Both teachers had prior experience with CWPT as a part of an earlier schoolwide implementation effort (Cheryl Utley, personal communication, 2001).

Classroom Setting

The study was carried out during the regularly scheduled spelling periods of both classes. The classrooms were equipped with a blackboard, an overhead projector, a desk and a chair for each student and the teacher, a large desk for group work, and six computers. Additionally, Mr. B occasionally used the multipurpose room during peer tutoring phases for "a change of setting." The multipurpose room contained large movable tables with enough chairs for approximately 200 students. In both classes, the spelling instruction schedule varied weekly, depending on other subjects being taught per the school's master calendar or special events. Spelling periods were held almost every day and lasted from 20 to 40 minutes in both classes.

Design

A single-subject reversal design with counterbalanced phases across two classes was used. In Ms. A's class, the conditions were manipulated in the BAB

sequence. In Mr. B's class, they were manipulated in the ABAB sequence, where: A = teacher-led spelling instruction and B = spelling peer tutoring. This design was used because Ms. A was using peer tutoring prior to beginning the study. It also offered the additional benefit of controlling for the order of treatment effects (Rusch, Rose, & Greenwood, 1988).

The study began in Mr. B's class two weeks after the beginning of the study in Ms. A's class. The length of each phase was based on visual analyses of the trends and means in each phase and, secondarily, on the number of weeks left in the school year.

Teacher-led instruction (A). The teachers were asked to use their own instructional strategies, but not to use activities involving pairs of students working together as in CWPT. Teacher instructional strategies included solving problems in the textbooks (e.g., filling in the blank, sorting the words in alphabetical order, filling in crossword puzzles, writing definitions), creating sentences using the spelling words, and writing words three to five times each. Throughout the study, the words were taught in the order of the lessons in the textbooks used. One lesson was taught in a week, and each lesson contained 25 words for Ms. A's class and 18 words for Mr. B's class.

Peer tutoring (B). The key procedures of peer tutoring were reciprocal peer tutoring and group reinforcement (Greenwood et al., 1997). Students in both classrooms were paired either at random or by skill pairing. Pairs were assigned to one of two teams, which competed with each other by earning points for correct responding. Partners and teams changed weekly to allow students to learn how to work cooperatively with many other students.

Before tutoring began, weekly pretests were administered to check the mean difficulty of spelling words and to provide a basis for measuring posttest gains (Greenwood et al., 1987). Likewise, posttests were administered after tutoring each week to measure students' skill improvement, give feedback on their performance, and provide a basis for social reward.

During a tutoring session, one student in the pair performed the tutor role and the other performed the tutee role, while a teacher supervised the classroom (Greenwood et al., 1997). As the tutor gave a question from a tutoring list consisting of weekly words and an answer key prepared by the teacher or taken from a textbook, the tutee wrote a response on a tutoring worksheet. If the spelling was correct, 2 points were recorded by the tutor on the point sheet. If a word was misspelled, the correct spelling was provided by the tutor, and the tutee verbally spelled and practiced writing it three times. If all three practices were performed correctly, 1 point was awarded. In this way, all points were recorded on the point sheet by the tutor. To encourage correct tutoring behaviors, bonus points were awarded by a teacher to students who followed the tutoring procedures correctly. After 10 minutes had elapsed on the timer, the students switched roles (Greenwood et al., 1997), whereby the tutee now performed the tutor role and the tutor performed the tutee role. Upon completion of the reciprocal tutoring sessions, each student reported his or her points earned to the teacher.

The tutoring session typically lasted about 30 to 35 minutes, and peer tutoring was implemented a minimum of three days a week to maximize students'

learning. The fifth day was used to conduct the posttest covering the material for the week and the pretest covering the material for the upcoming week.

Teacher training. Because of their previous professional development experiences with CWPT, no formal training was provided to Ms. A and Mr. B. Instead, the researcher conducted a brief review of the standard CWPT procedures individually with each teacher. The review covered having students work in pairs, assigning pairs to two competing teams, and administering pre- and posttests.

The CWPT procedures also were taught to the students during a single training session, which lasted for 40 to 50 minutes. The training consisted of verbal description of the procedures of CWPT, demonstrations of examples and nonexamples by the teacher and the researcher, and practice by the students while the teachers and researcher provided corrective feedback. The training took place during the first spelling period of the first CWPT phase in each class. Additionally, a brief review of procedures was given immediately before the second and third CWPT sessions either by the teachers or the researcher.

Modified CWPT procedures. Although the standard CWPT procedures (Greenwood et al., 1997) were initially planned to be used, considerable modifications were made prior to and during the current study. These included (a) no public posting of students' performance, (b) reduced number of peer tutoring sessions (e.g., a minimum of 3 days per week), (c) changes in pairing procedures, (d) no use of team competition, and (e) no public point recording on charts and verbal reporting by the students. These modifications were made due to school policies. Because of these considerable deviations from the standard CWPT procedure, the term *peer tutoring* is used below to differentiate this modified form of peer tutoring from the standard CWPT procedures.

Measures

A range of measures were used to address the research questions. Specifically, a checklist based on standard CWPT implementation was used to monitor implementation fidelity, weekly spelling and generalization tests were used to measure student spelling outcomes, and student and teacher satisfaction surveys were administered at the end of the study.

Fidelity. Fidelity observations were conducted by the researcher during peer tutoring phases. The daily fidelity score for each tutoring session was calculated by dividing the number of observed items by the total number of items on the fidelity checklist multiplied by 100. The weekly fidelity score was calculated by averaging scores of all available tutoring session scores in a week.

Spelling tests and administration. To measure spelling effectiveness and generalization, four tests were administered (i.e., pretests, posttests, sentence dictation test, and retention test). Every week, typically on Monday, all words from a new lesson in the textbook were pretested to evaluate pre-existing levels of spelling accuracy. The same words were posttested at the end of the week, typically on Friday, to evaluate the effects of instruction during the week. When these tests were not administered on these days for any reason, a given test was administered the next school day. When administering pre- and posttests, a teacher pronounced a word once, used the word in a sentence, and pronounced the word a few more times, as needed.

Further, a sentence dictation test was administered on Friday (a week after posttest) to evaluate generalization of spelling skills to this task. From a word list taught in a given week, five words were selected according to the procedures described below and tested in five sentences; each sentence included one word. When administering sentence dictation tests, the students were asked to write verbatim sentences that contained the spelling words on their test paper. The teachers read sentences and repeated them as necessary. A retention test was also administered on Friday (two weeks after posttest). From a word list taught in a given week, five or six words were selected according to the procedures described below and tested. Retention tests were administered in the same manner as the pre- and posttests. For both generalization tests, only the mastered spelling words were scored.

After each test, the researcher checked the words and recorded whether words were correct or incorrect. A spelling word was considered correct when the letters were readable and in the correct order in accordance with the textbook and a dictionary. After the scoring, the test paper was returned to the teachers to be graded and given back to the students.

Word selection procedures for sentence dictation and retention tests. From a weekly spelling word list, 10 or 11 target words were selected to be tested for either the retention (i.e., 5 or 6 words) or sentence dictation tests (i.e., 5 words) according to the order of the greatest number of students who mastered each word (i.e., incorrectly spelled the word on the pretest and correctly spelled the same word on the posttest). For example, if the word *apple* was mastered by 13 students and the word *orange* was mastered by 10 students, the word *apple* would be selected before the word *orange.* To distribute the words equally in terms of the number of students who mastered them, the words were matched for the number of students who mastered. When 11 words were selected, the extra word was always tested for retention. Some changes were made to the testing procedures during the course of the study because of students' attrition and the articulation of the researcher (not a native speaker of English). Table 1 shows a summary of these measures and relationship with the tests.

Pretest-posttest gain scores. To answer the second research question, "What is the overall effect of peer tutoring on weekly spelling gains?," a pretest-posttest gain score was calculated. This measure is commonly used in the CWPT literature to evaluate whether students improved their spelling accuracy during a given week (e.g., Greenwood, Terry, Arreaga-Mayer, & Finney, 1992). Pretest-posttest gain scores were calculated by subtracting the percentage correct on the pretest from the percentage correct on the corresponding posttest in a given week.

Mastery of unknown words score. This measure evaluates how well students acquire previously unknown words during a week of instruction and has several advantages over pretest-posttest gain scores (e.g., no ceiling effects, no need to adjust the difficulty level, higher accuracy). The mastery of unknown words score was calculated based on both the weekly pre- and posttests. First, all the words that the students spelled incorrectly on a pretest in a given week were classified into one of two outcomes based on the posttest: The words correctly spelled on the posttest (mastered during the week) and the words incorrectly spelled on the posttest (not mastered during the week). Then, the proportion of (a) the number of the words that were mastered to (b) the number of the words that were spelled incorrectly on

the pretest (both mastered and not mastered) was calculated by dividing (a) by (b) and multiplying by 100.

For the group data, the mean of this score was not calculated because the number of incorrectly spelled words on a pretest of a given week varied from student to student. Instead, the aggregated mastery score was used. To calculate it, the number of the words that were incorrectly spelled on the pretest, but correctly spelled on the posttest of each student, was counted and summed for all the students in the class. Then, the number of incorrectly spelled words on the pretest of each student was counted and summed for all the students. Finally, the first number was divided by the second number and multiplied by 100.

Loss of known words score. When examining the results of instruction on the basis of word acquisition (or mastery), there are four possible outcomes: Incorrectly spelled on the pretest, but correctly spelled on the posttest (mastered); incorrectly spelled on both the pre- and posttests (not mastered); correctly spelled

Table 1
Description of Measures

Measures	Test Used	Examines	Calculation
Pretest-posttest gain score	Pre- and posttests	Improvement in spelling accuracy in a given week	Percent correct on the posttest minus percent correct on the pretest
Mastery of unknown words score	Pre- and posttests	Mastery of unknown words as a result of a week of instruction	Proportion of correctly spelled words on the posttest out of all the incorrectly spelled words on the pretest
Loss of known words score	Pre- and posttests	Loss of known words in a given week	Proportion of incorrectly spelled words on the posttest out of all the correctly spelled words on the pretest
Generalization across other writing skills	Sentence dictation test	Accuracy of spelling when used in different tasks given mastery	Proportion of correctly spelled words out of all the mastered words appearing in a given sentence dictation test
Generalization over time (retention)	Retention test	Accuracy of spelling when used at a different occasion given mastery	Proportion of correctly spelled words out of all the mastered words appearing in a given retention test

on both the pre- and posttests (maintained); and correctly spelled on the pretest, but incorrectly spelled on the posttest (lost). The most desirable outcome in spelling instruction may be a 100% mastery, in which all the words unknown to the students prior to instruction (0% correct on the pretest) become known after instruction (100% correct on the posttest). If so, a teacher can and should use all resources to teach only the words unknown to the students based on pretest results. However, this would be the case only if the words correctly spelled on the pretest are not lost (i.e., correct on the pretest and incorrect on the posttest). To maximize students' learning, all words known to the students prior to instruction should be kept in their repertoire, while unknown words should come into the repertoire. Thus, this measure provided another way of examining the effects of instruction on increasing spelling accuracy.

The loss of known words score was calculated using the following procedures. First, all the words that were correctly spelled on a pretest in a given week were classified into one of two outcomes based on the posttest: The words correctly spelled on the posttest (maintained) and the words incorrectly spelled on the posttest (lost). Then, the proportion of (a) the number of the words that were lost to (b) the number of the words that were correctly spelled on the pretest (both the maintained and not maintained) was calculated by dividing (a) by (b), multiplied by 100. A smaller score indicates that students maintain more words. For the same reason as for the mastery of unknown words score, the mean was not used for the group data. Instead, an aggregated score was calculated.

Generalization to sentence dictation. In order to answer the third research question, "What was the magnitude and variation in students' generalization of mastered words on a sentence dictation task one week later?," the sentence dictation test was administered. To examine the generalization of words acquired only through peer tutoring, (a) the words on the test were unknown to the students prior to peer tutoring (misspelled on the pretest), (b) these words were known to the students after peer tutoring (correctly spelled on the posttest), and (c) only those words that satisfy both (a) and (b) should be evaluated for generalization. Thus, the percentage correct was given by dividing (a) the number of the correctly spelled mastered words by (b) the total number of mastered words that appeared on the sentence dictation test, multiplied by 100. This way, only the generalization of the words that were acquired during the week of instruction (mastered words) was included in the calculation, although the test might have contained the words that a particular student did not master.

Because the number of the mastered words on a given sentence dictation test differed student by student (i.e., the denominator of the equation), the mean percentage correct and standard deviation were not used for the group data; instead, the aggregated percentage correct was calculated using the following procedures. First, the number of the mastered words correctly spelled on the sentence dictation test was counted for each student and summed for all students in the class in a given week. Second, the number of the mastered words that appeared on the sentence dictation test was counted for each student and summed for all students. Third, the first sum was divided the second sum and multiplied by 100, which was the aggregated percentage correct for a class in a given week.

Generalization over time. In order to answer the fourth research question, "What was the magnitude and variation in students' retention of mastered words two weeks later?," retention tests were administered two weeks after instruction. The percentage correct on a given retention test was calculated in the same manner as the sentence dictation test: Only spellings of the mastered words (i.e., incorrectly spelled on the pretest but correctly spelled on the posttest) that appeared on a retention test were scored. The percentage correct was given by dividing the number of the correctly spelled mastered words by the number of the mastered words that appeared on the retention test, multiplied by 100. The aggregated percentage correct for a given week was calculated in the same manner as for the sentence dictation tests.

Student and teacher satisfaction. The students and teachers were given consumer satisfaction surveys at the end of the study to evaluate social acceptability, a subjective measure of effectiveness of peer tutoring. The student survey included 10 questions, such as "How much did you like peer tutoring?" and "If you were a teacher, would you let your students use peer tutoring?" The teacher read the questions aloud and asked students to circle the answer for each question. The teacher survey included questions regarding the teachers' experience and acceptability of peer tutoring as well as evaluation of assistance from the researcher.

Inter-observer agreement. Two staff members conducted inter-observer agreement checks. For the pre- and posttest, sentence dictation, and retention tests, one of the two personnel checked these tests independently from the researcher. The agreement was checked for 32.1% of the pretests, 32.1% of the posttests, 38.5% of the sentence dictation tests, and 50.0% of the retention tests for Ms. A's class. It was calculated by dividing the number of words where two observers agreed on the accuracy by the total number of words evaluated by both observers multiplied by 100. For Ms. A's class, the mean agreement was 97.0% for the pretests, 96.1% for the posttests, 97.2% for the sentence dictation tests, and 97.1% for the retention tests. For Mr. B's class, 28.6% of the pretests, 35.7% of the posttests, 30.8% of the sentence dictation tests, and 24.0% of the retention tests were checked; the mean agreement was 95.4%, 94.8%, 96.6%, and 96.0%, respectively.

In order to ensure that the peer tutoring procedures were not used during the baseline phases (treatment contamination), the researcher observed the teacher's instructions to see if he or she used any form of structured peer tutoring. No incidents of structured peer tutoring activities were observed during the baseline phases, although the teachers occasionally instructed the students to work in small groups and help each other. Inter-observer agreement was obtained for 28.6% of all the observed spelling periods in Ms. A's class and 30.0% of all the observed periods in Mr. B's class. The observers agreed 100%.

Inter-observer agreement of the fidelity of peer tutoring implementation was calculated for 35.1% of all the observed spelling periods in Ms. A's class and 36.7% of the periods in Mr. B's class during the peer tutoring phases. The inter-observer agreement, arrived at by dividing the number of agreement by the number of agreement and disagreement on the checklist multiplied by 100, ranged from 64.9 to 92.6, with a mean of 81.9% in Ms. A's class, and between 58.2 and 85.2, with a mean of 78.6%, in Mr. B's class.

Results

The results of this investigation are presented below by research question.

***Research Question 1*:** What was the magnitude and variation in fidelity of peer tutoring implementation?

Overall, the implementation fidelity of the CWPT procedures was low, as indicated by the fact that even the highest score on implementation fidelity did not exceed 90% in either class. In Ms. A's class, weekly fidelity ranged between 40.6 and 84.7 with a mean of 68.4% across all tutoring sessions. In Mr. B's class, it ranged between 47.5 and 77.6 with a mean of 65.2%. The most frequently missed steps were (a) students moved to their tutoring partners quickly and quietly, (b) students returned to their seats quickly after tutoring sessions, (c) tutors gave the next tutoring item immediately after the previous item, and (d) inappropriate behavior was low during the peer tutoring sessions (see Figures 1 and 2). When the original CWPT procedures were modified, all the steps related to the use of points were missed, further decreasing fidelity. The effort to link tokens to peers' point earning and tutoring fidelity appeared to have an initial increase in on-task behavior; however, the effects of tokens contingent upon the points did not last long. Also, peer tutoring sessions were conducted only twice a week, on average.

***Research Question 2*:** What was the magnitude and variation in students' increased spelling accuracy?

Weekly pretest-posttest gain scores are displayed in Figure 1. In Ms. A's class, the transition from the initial peer tutoring phase to the baseline phase resulted in a mean decrease of 12.5 points (difference in the mean percentage correct on a pretest and posttest in a given week) in accuracy in weekly spelling gain (from 30 points for peer tutoring to 17.5 points for baseline). A return to peer tutoring resulted in an increase of 7.4 points in weekly spelling gain compared to the baseline phase. Throughout the study, the mean percentage correct on the pretests ranged between 17.7% and 62.0% with a mean of 40.7% ($SD = 12.2$). Some of the lowest gains were due to a pretest ceiling effect during some weeks.

Implementation fidelity for spelling accuracy ranged between 40.6% and 84.7% with a mean of 68.9%. Although smaller gains would reasonably correspond to lower implementation fidelity and/or a smaller number of peer tutoring sessions in a given week, no clear co-variation was observed between gains in accuracy and implementation fidelity or the number of peer tutoring sessions. For example, even when the fidelity was relatively high during the first two weeks of the second peer tutoring phase, the pretest-posttest gain score was as low as the previous baseline. This may be because at such a low frequency of weekly peer tutoring, high fidelity may not influence learning outcome.

In Mr. B's class, initial implementation of peer tutoring resulted in a mean gain of 18.6 points in accuracy compared to the first baseline phase (from 20.5 points for baseline to 39.1 points for peer tutoring). With a return to the second baseline phase, mean gain in accuracy decreased by 15.8 points to 23.3 points, which increased by a mean gain of 11.5 points with the second implementation of peer tutoring. Although ceiling effects were less apparent in Mr. B's than in Ms. A's class, they occurred occasionally (e.g., weeks 13 and 14). Throughout the study, the mean

Figure 1. Pretest-posttest gain in spelling accuracy (left axis) and implementation fidelity (right axis) over weeks. Circles represent pretest-posttest gains and diamonds represent implementation fidelity. Open diamonds indicate percent implementation fidelity of peer tutoring when it was used once in a week. Closed diamonds indicate mean percentage fidelity when peer tutoring was used two or three times in a week.

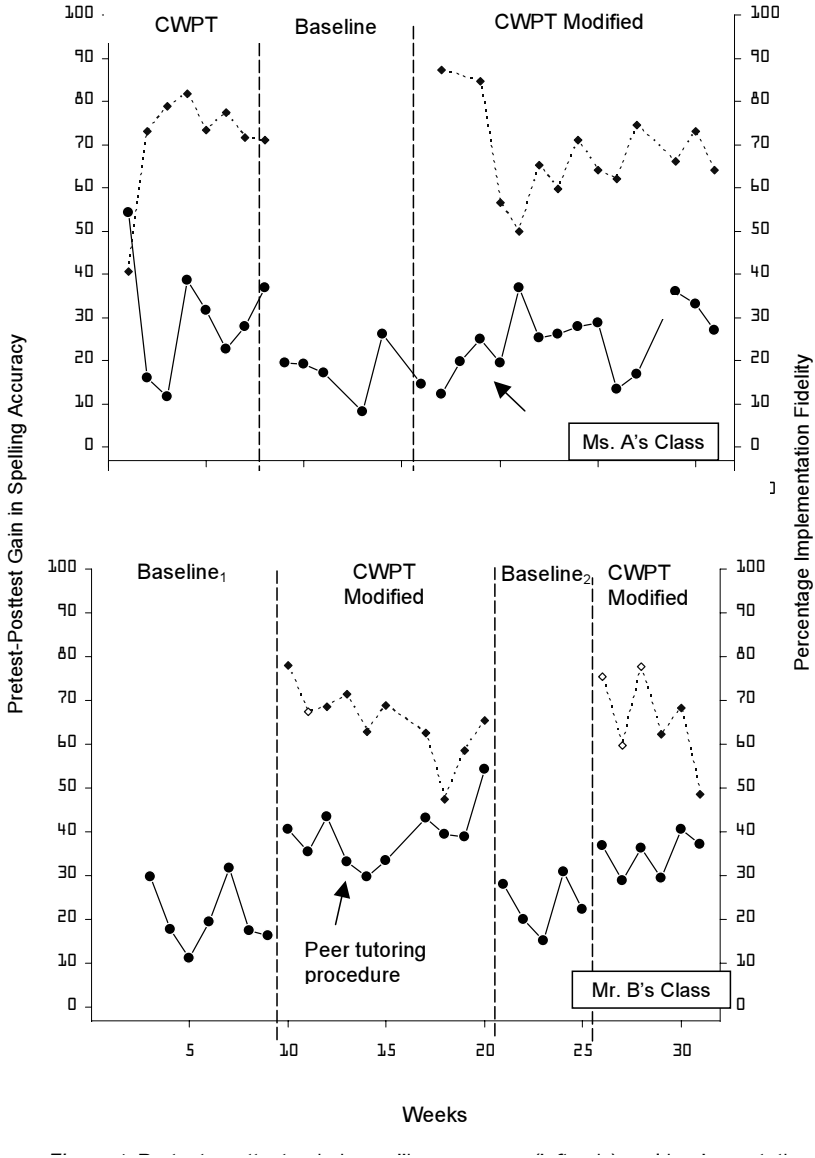
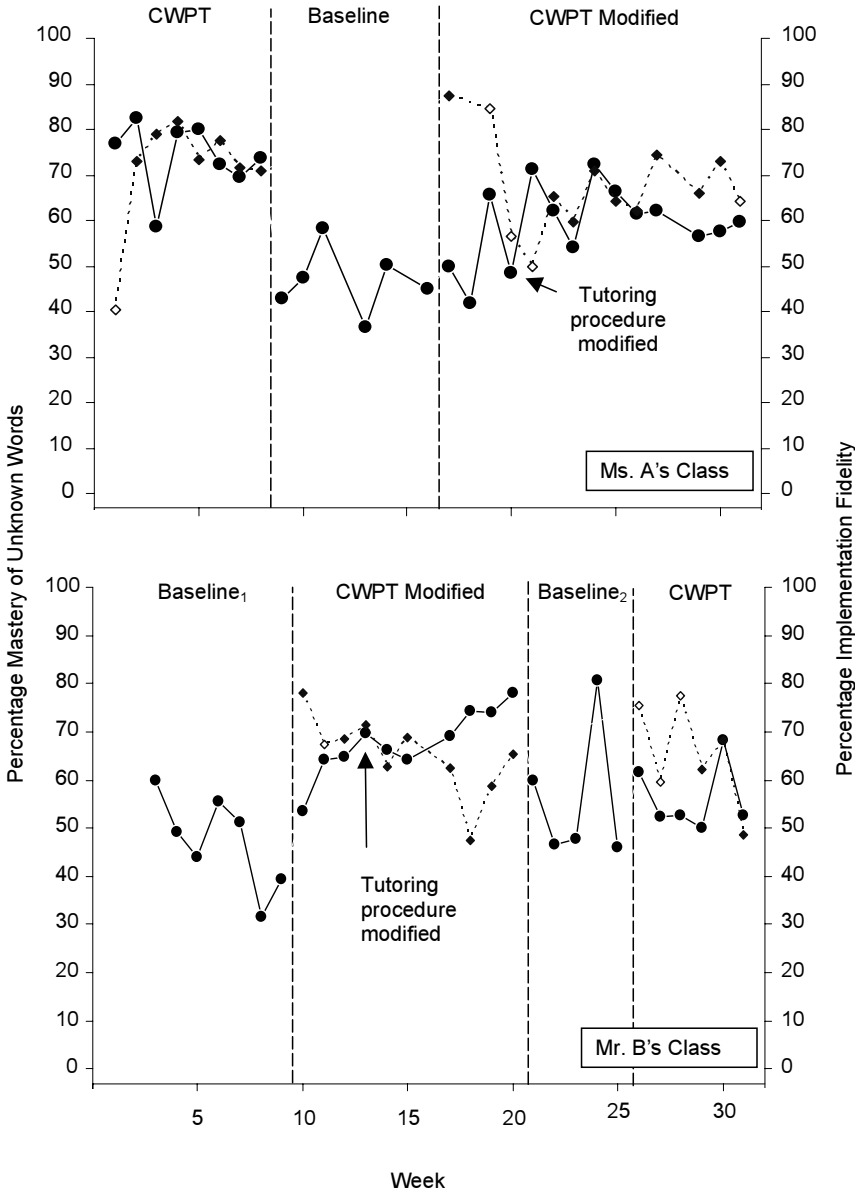


Figure 2. Percentage mastery of unknown words (circles) and implementation fidelity (diamonds) over weeks. Open diamonds indicate percent implementation fidelity of peer tutoring when it was used once in a week. Closed diamonds indicate mean percentage fidelity when peer tutoring was used two or three times in a week.



percentage correct on the pretests of Mr. B's class ranged between 18.4% and 57.3% with a mean of 39.0% (*SD* = 10.3). Again, no clear co-variation was seen between weekly gain and the implementation fidelity or frequency of use.

Based on a criteria for treatment failure (i.e., less than 10% improvement under peer tutoring relative to teacher procedures) used by Greenwood et al. (1987), the result of the current study may not be considered a treatment failure. The mean pretest-posttest gain scores during all the peer tutoring phases of the two classes was 31.3 points whereas that of all the baseline phases was 20.3 points. However, there was considerable variability within each phase and minimal improvements in posttest scores during the second peer tutoring phase in the two classes.

In Ms. A's class, the phase change from the first peer tutoring phase to the baseline phase resulted in an immediate and large drop in mastery, although interpretation of this change was hampered by the downward trend toward the end of the first peer tutoring phase. The second transition between the baseline phase and the second peer tutoring phase resulted in slight changes in the levels, and the trend changed. Overall, mastery during the peer tutoring phases was 18.0 points higher than that of the baseline phase.

In Mr. B's class, the initial implementation of peer tutoring resulted in a clear change in trends, with a slight initial increase. The return to the baseline resulted in an immediate drop and a downward trend. However, the re-introduction of peer tutoring did not replicate the same degree of impact as the first peer tutoring phase. The lack of a clear effect of peer tutoring during the second peer tutoring phase was similar to the changes demonstrated in the phase change in Ms. A's class. Nevertheless, overall, the mastery during peer tutoring phases was 12.5 points higher than that of the baseline phases. In summary, the data from the two classes suggested that the change of the instructional strategies exerted a moderate control over the students' spelling skill acquisition.

Loss of known words remained relatively low throughout all the phases and across classes, ranging from 8.1 to 9.0% in Ms. A's class and from 3.7% to 9.7% in Mr. B's class. Based on visual inspection, no clear difference was seen between the peer tutoring phases and the baseline phases (see Figure 3).

***Research Question 3*:** What was the magnitude and variation in students' generalization of mastered words on a sentence dictation task one week later?

Large variability was found within each phase throughout the study for generalization of spelling words to sentence dictation tests (see Figure 4). Because of missing data of weeks 10 through 16 in Ms. A's class, only a casual analysis in Mr. B's class was possible. Only small mean differences were noted between the peer tutoring and baseline phases. The aggregated percent correct for each phase was 61 for both the first and the second peer tutoring phases in Ms. A's class. In Mr. B's class, it was 45 during the first baseline phase, versus 58 during the first peer tutoring phase. When conditions reversed to baseline, the aggregated percentage declined to 42, but then recovered to 63 when peer tutoring was re-introduced. Thus, peer tutoring resulted in 13 to 18 percentage points higher accuracy in the generalized spelling than the teacher-led baseline instruction. Although data points were missing during the baseline in Ms. A's class due to a procedural change, comparisons between the peer tutoring phases of Ms. A's class and the baseline phases of Mr. B's class were pos-

Figure 3. Percentage loss of known words (circles) and implementation fidelity (diamonds) over weeks. Open diamonds indicate percent implementation fidelity of peer tutoring when it was used once in a week. Closed diamonds indicate mean percentage fidelity when peer tutoring was used two or three times in a week.

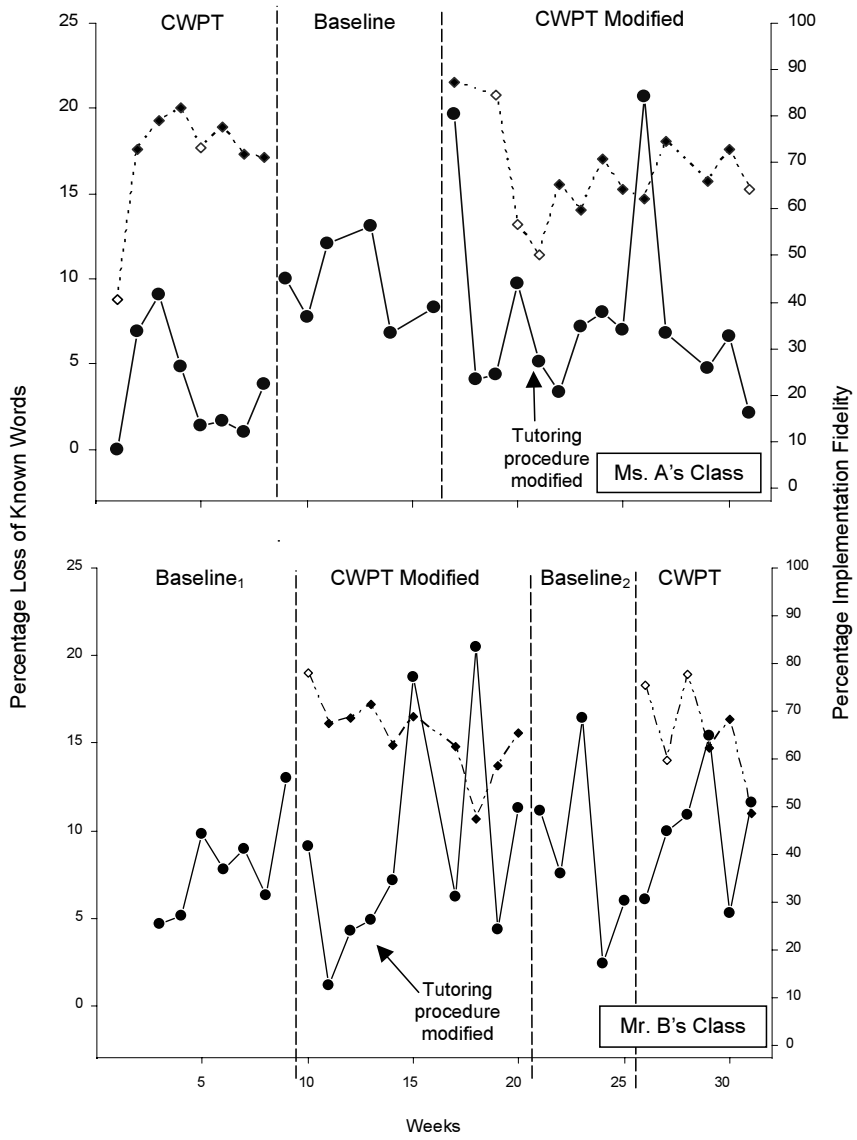
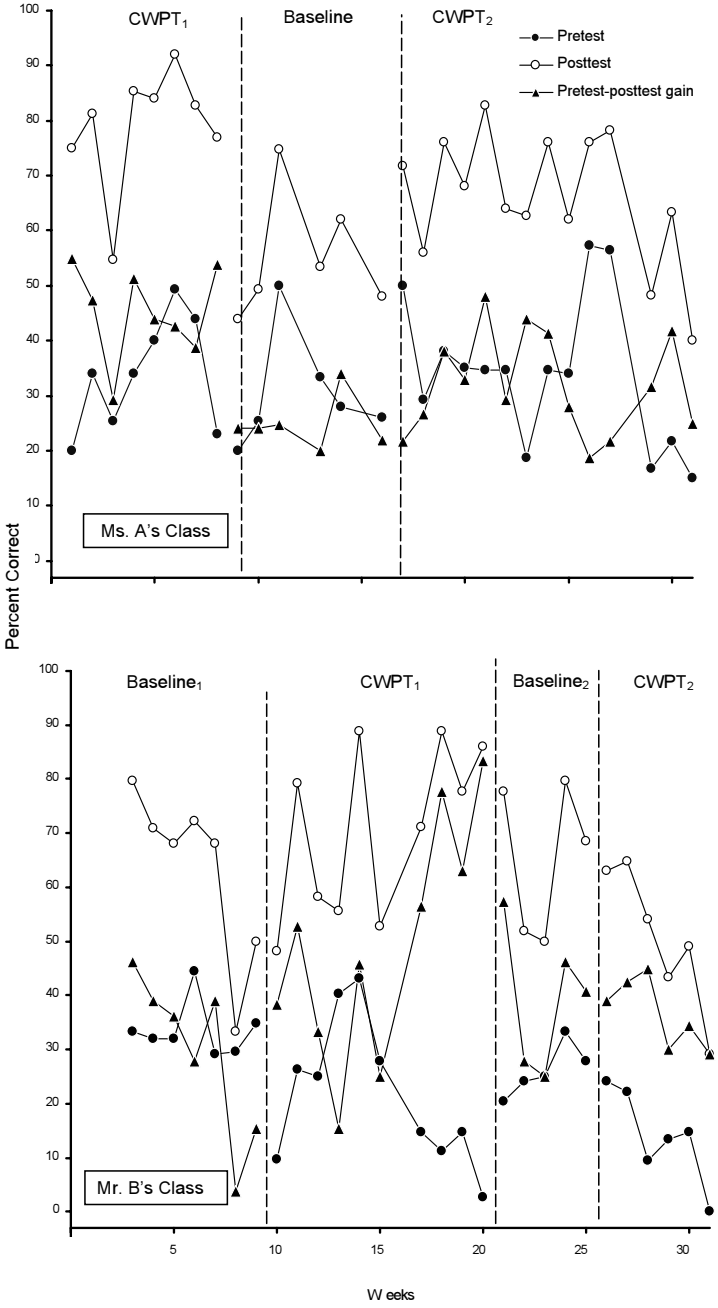


Figure 4. Students' generalization of mastered words on a sentence dictation task one week later.



sible. Across classrooms, the aggregated percentage correct during the first and second peer tutoring phases in Ms. A's was 16.1 and 19.4 points higher than those of the first and second baseline phases in Mr. B's class, respectively.

***Research Question 4*:** What was the magnitude and variation in students' retention of mastered words two weeks later?

Unlike the effects for generalization of spelling skills to sentence dictations, there was no clear evidence favoring peer tutoring over the baseline phases in Mr. B's class (see Figure 5).

Satisfaction Survey

Students. With regard to the question "How much did you like peer tutoring?," only less than one fourth of the students from both classes selected "a lot" (see Table 2).

However, most students responded in favor of using peer tutoring in subsequent questions such as "If you were a teacher, would you let your students use peer tutoring?" and "How much did peer tutoring help you spell better?" For the questions regarding interpersonal relationship with peers, not many students reported that they "[thought] that peers were more friendly during CWPT." On the question "Do you think some of the kids in your class feel that you are smarter now, because they have been your partner in peer tutoring?," 54% of Ms. A's students responded with "yes, for sure," compared to only 20% of Mr. B's students. Despite some differences like these across classes, the students moderately favored peer tutoring overall.

Teachers. Overall, Ms. A responded in favor of peer tutoring on most of the items on the questionnaire (see Table 3), selecting strongly agree or agree in responding to key questions such as "The CWPT procedures were helpful for students of all ability levels in my classroom," "[The] project staff provided the necessary assistance throughout the peer tutoring program," and "My students seemed to enjoy learning with the CWPT procedures." However, Ms. A pointed out two concerns when asked for comments. The first involved incorporating peer tutoring into her regular daily schedule. To the statement "The time for CWPT was easy to plan into my daily schedules," she selected "not sure." Her second concern involved students' inability to resolve conflicts during peer tutoring sessions (e.g., arguing over points, accusing other pairs of cheating). Finally, Ms. A expressed her dislike of the use of competing teams because of the negativity it appeared to have caused.

Figure 5. Students' retention of mastered words two weeks later.

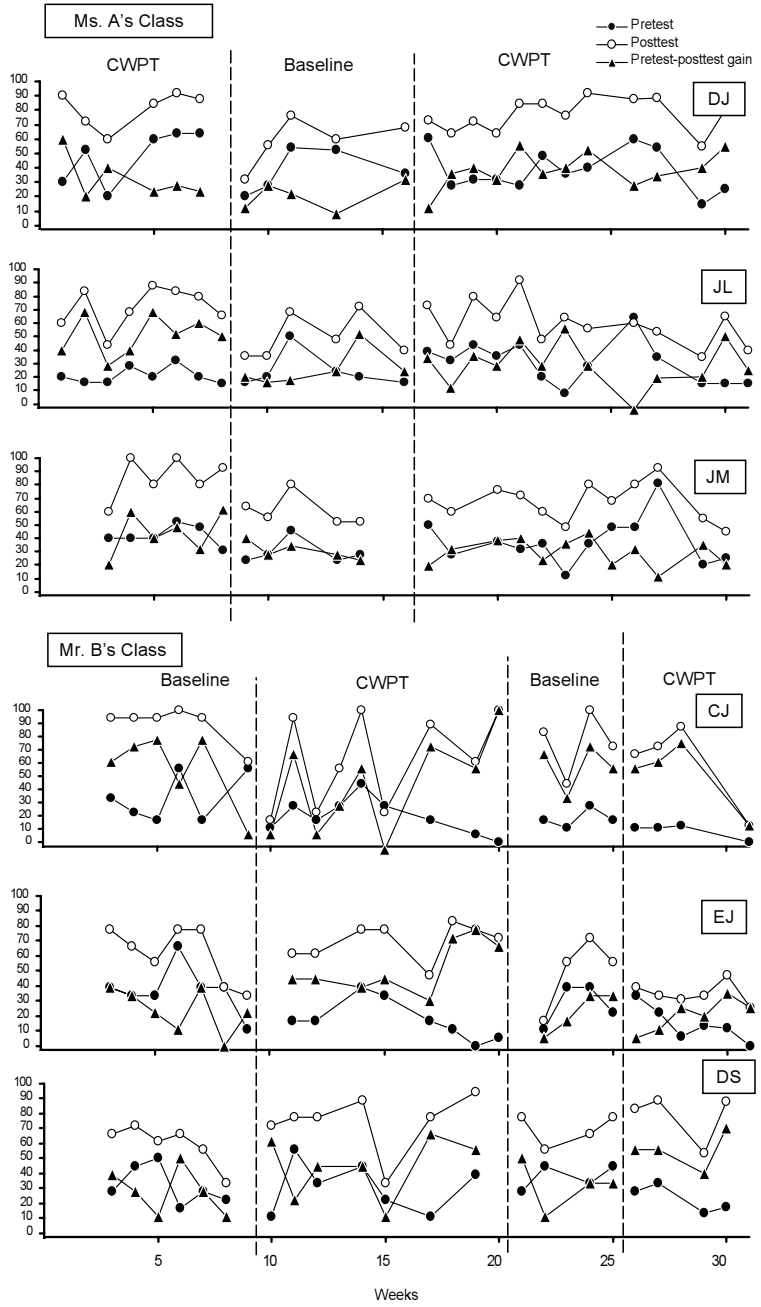


Table 2

Results of Student Survey

Questions	Ms. A's Class			Mr. B's Class		
1. How much did you like peer tutoring?	not at all 15.38	it was OK 61.54	a lot 23.08	not at all 20.00	it was OK 60.00	a lot 20.00
2. If you were a teacher, would you let your students use peer tutoring?	no way 0.00	maybe 30.77	for sure 69.23	no way 13.33	maybe 26.67	for sure 60.00
3. How much did peer tutoring help you spell better?	not at all 0.00	some 30.77	a lot 69.23	not at all 6.67	some 6.67	a lot 86.67
4. How much did you like being on a team in peer tutoring?	not at all 7.69	some 15.38	a lot 23.08	not at all 6.67	some 60.00	a lot 33.33
5. Do you think the kids in your class were more friendly to you during peer tutoring?	not really 46.15	maybe 23.08	yes, for sure 15.38	not really 40.00	maybe 46.67	yes, for sure 13.33
6. Do you feel that some of the kids in your class feel that you are smarter now, because they have been your partner in peer tutoring?	not really 15.38	maybe 30.77	yes, for sure 53.85	not really 20.00	maybe 60.00	yes, for sure 20.00
7. When you did CWPT, did you find out the kids in your class were smarter than you thought they were?	not really 23.08	maybe 23.08	yes, for sure 53.85	not really 33.33	maybe 20.00	yes, for sure 46.67
8. Would you like to do peer tutoring with your teacher next year?	not realy 23.08	it would be OK 7.69	yes 69.23	not really 33.33	it would be OK 20.00	yes 46.67
9. What CWPT role did you like best?	tutor 52.85	tutee 45.15		tutor 45.67	tutee 39.00	both 12.33
10. Do you like earning points during CWPT?	not at all 15.38	some 15.38	a lot 69.23	not at all	some 13.33	a lot 20.00

Table 3
Results of Teacher Survey

Questions	Ms. A	Mr. B
1. The Juniper Gardens Children's Project staff provided the necessary assistance throughout the peer tutoring program.	Strongly Agree	Agree
2. I felt the Project Staff explained the CWPT program in understandable terms.	Agree	Strongly Agree
3. Training sessions provided enough information to independently carry out the program effectively.	Strongly Agree	Disagree
4. The Project Staff offered helpful suggestions and feedback for implementing CWPT in my classroom.	Strongly Agree	Disagree
5. The Project Staff provided necessary instruction on how to prepare materials for daily tutoring sessions.	Agree	Strongly Disagree
6. The materials used with CWPT were beneficial.	Agree	Disagree
7. The procedure for training the students in tutoring involved a reasonable amount of time and were effective.	Agree	Disagree
8. The time allotted for tutoring (30 minutes for spelling and 40 minutes for reading) a day was necessary for improving my students' performance.	Agree	Disagree
9. The CWPT procedures were helpful for students of all ability levels in my classroom.	Agree	Strongly Disagree
10. The time for CWPT was easy to plan into my regular daily schedule.	Not Sure	Disagree
11. My students seemed to enjoy learning with the CWPT procedures.	Agree	Not Sure
12. The CWPT procedures helped students pay attention and get involved in instruction.	Agree	Disagree
13. Students were supportive and reinforcing to each other in the tutoring dyads.	Agree	Disagree
14. The students enjoyed seeing their scores as well as team scores on the point chart.	Agree	Agree
15. Students were capable of resolving conflicts that arose in their tutoring dyads.	Disagree	Strongly Disagree
16. The procedures of CWPT provided students with social skills necessary for good peer relations.	Agree	Strongly Disagree
17. CWPT was as effective as traditional approaches to teaching spelling/reading.	Agree	Agree
18. I plan to use CWPT in future class.	Agree	Not Sure
19. I would like to use CWPT next year for the following subjects.	Spelling	No Answer

Mr. B's responses differed from Ms. A's. He responded negatively to most statements. For example, he chose "strongly disagree" as response to "The CWPT procedures were helpful for students of all ability levels in my classroom" and "The procedures of CWPT provided students with social skills necessary for good peer relations." He chose "disagree" as his response to "The CWPT procedures helped students pay attention and get involved in instruction" and "The materials used with CWPT were beneficial."

***Research Question 5*:** What was the overall mean *ES* of the CWPT intervention across baseline and intervention phases?

***Research Question 5a*:** What were the *ES*s of the CWPT intervention on students' generalization of mastered vs. lost words on a sentence dictation task one week later?

***Research Question 5b*:** What were the *ES*s of CWPT on students' retention of mastered words two weeks later?

Effect sizes of the CWPT intervention were estimated using Cohen's (1988) *d*, the standardized mean difference. The basic formula is as follows:

$$d = \frac{X_1 - X_2}{(\overline{SD_1 + SD_2})/2}$$

For single-case studies, the pooled estimate of standard deviation is recommended (Coe, 2000). The basic formula is as follows:

$$SD_{\text{pooled}} = \sqrt{\frac{(N_E\text{-}1)\ SD_E^2 + (N_c\text{-}1)SD_c^2}{N_E + N_c\ \text{-}2}}$$

As seen in Table 4, the *ES*s of the CWPT intervention for Ms. A and Mr. B were 1.09 and 2.38, respectively. The large *ES* of the CWPT intervention for Ms. A's class indicated that the mean was at the 84th percentile of the baseline group. For Mr. B's class, the large *ES* indicated that the mean was beyond the 97.7 percentile of the baseline group. The *ES*s for mastery words for Ms. A and Mr. B in the CWPT intervention were 1.93 and 1.16, respectively. The large *ES* of mastered spelling words in Ms. A's classroom indicated that the mean was at the 97.1 percentile of the baseline group. Similarly, the large *ES* of lost words in Mr. B's classroom indicated that the mean was at the 86th percentile of the baseline group.

Table 4
Effect Sizes of CWPT Intervention in Two Classrooms

Phase	Statistic	Ms. A	Mr. B
Baseline	Mean	17.52	21.64
	Pooled Standard Deviation	5.98	6.87
CWPT	Mean	26.75	37.48
	Pooled Standard Deviation	10.25	6.40
	Effect Size	1.099	2.38

Table 5
Effect Sizes of Mastered Spelling Words in CWPT Intervention in Two Classrooms

Phase	Statistic	Ms. A	Mr. B
Baseline	Mean	46.72	50.99
	Pooled Standard Deviation	7.30	12.34
CWPT	Mean	64.69	63.47
	Pooled Standard Deviation	10.93	8.85
	Effect Size	1.93	1.16

Table 6
Effect Size of Lost Spelling Words in CWPT Intervention

Phase	Statistic	Ms. A	Mr. B
Baseline	Mean	9.68	8.25
	Pooled Standard Deviation	2.51	3.89
CWPT	Mean	6.28	9.19
	Pooled Standard Deviation	5.21	5.42
	Effect Size	-0.83	0.19

The *ES*s for the loss of spelling words in the CWPT intervention in Ms. A and Mr. B's classrooms were -0.83 and 0.19, respectively. The negative *ES* for the loss of spelling words in Ms. A's classroom indicated that CWPT was not a powerful intervention for mastery of spelling words. The result of the analysis for the *ES* of CWPT on the loss of words in Mr. B's classroom indicated that the mean was at the 97th percentile of the baseline group. In Mr. B's classroom, the *ES* for the CWPT intervention on the generalization of words on a sentence dictation task one week later was 1.52, indicating that the mean was at the 93.3 percentile of the baseline group. The *ES* for the CWPT intervention on the retention of mastered words two weeks later in Mr. B's classroom was 0.16, indicating that the mean is at the 84th percentile of the baseline group.

Table 7
Effect Sizes of Generalization and Retention of Spelling Words in CWPT Intervention in Mr. B's Classroom

Phase	Statistic	Generalization	Retention
Baseline	Mean	42.78	56.18
	Pooled Standard Deviation	11.86	15.28
CWPT	Mean	60.81	59.04
	Pooled Standard Deviation	12.30	19.24
	Effect Size	1.52	0.16

Discussion

The existing CWPT literature lacks a body of spelling intervention research on skill generalization. Indeed, studies of spelling interventions employing generalization measures often have not taken initial word mastery into account. Therefore, the current study attempted to evaluate the effectiveness of CWPT on generalization measures given initial word mastery as well as fidelity of peer tutoring implementation. Despite modified peer tutoring implementation, most notably the decreased number of sessions and modifications teachers made from standard CWPT procedures, a modest superiority of CWPT procedures compared to teacher-led instruction was apparent in weekly pretest-posttest gain scores, mastery scores, and generalization to the sentence dictation task, but not in retention.

While the majority of meta-analytical studies have focused on reading and math (e.g., Gersten & Baker, 2000), the current single-case study of CWPT investigated the overall mean *ES*s of CWPT in spelling across baseline and intervention phases in two classrooms. The findings revealed that the *ES*s were large, indicating that this instructional strategy was effective in teaching spelling words to third graders. In addition, the *ES*s of mastery vs. lost words was large in one classroom where teacher and student variables did not interact and affect the variation and magnitude in pre-posttest gain scores, suggesting that CWPT was a powerful intervention providing the instructional components for at-risk students to master spelling words.

Implications for Future Research

Fidelity of CWPT implementation. In general, the results of the analysis of CWPT fidelity of implementation were low throughout the study due to modifications introduced by teachers in their efforts to control the students' off-task and disruptive behaviors during the tutoring sessions. Additionally, compared to the recommended use of CWPT procedures (i.e., four sessions per week), the number of tutoring sessions in a week was small because of potentially conflicting teaching requirements (e.g., testing, parent-teacher conferences) and the school's scheduling of extracurricular activities throughout the week.

In the existing CWPT literature, a reduced number of opportunities to participate in CWPT sessions, low implementation fidelity, and a high level of percentage correct on the pretest (ceiling effect) have been considered threats to a successful outcome with the CWPT program as measured by weekly pretest-posttest gain scores (e.g., Greenwood et al., 1987; Greenwood, Delquadri, & Hall, 1989; Greenwood et al., 1992).

An anecdotal observation showed that peer tutoring for students in this study was not a strong reinforcer and that the teachers were not open to using back-up reinforcement procedures in an effort to improve its effectiveness as a generalized conditioned reinforcer. For only a few students, earning points for correct spellings was highly reinforcing as they consistently worked at a rapid rate. However, for the majority, earning points and being the winning team were not reinforcing. Further, over-reporting of points earned (cheating) during peer tutoring sessions was not corrected as indicated in the CWPT manual (Greenwood et al., 1997) or punished by teachers. In addition, the tokens delivered by the teachers were not effective

because they were awarded non-contingently on students' behavior outside of the peer tutoring sessions, which may have generally weakened the reinforcing value of tokens.

Spelling accuracy. A better implementation of CWPT had been planned, and greater effects in students' spelling learning were anticipated based on past research. Yet, even in the face of the implementation problems just discussed, some modest effects favoring learning spelling words with CWPT were observed. Thus, on the weekly pretest-posttest gain scores and mastery of unknown words scores, a slight superiority of CWPT procedures over teacher-led instruction was observed.

No comparable CWPT study in the literature has addressed the mastery of unknown words prior to instruction. In the studies of the self-correction procedure by McGuffin, Martz, and Heron (1997) and by Wirtz, Gardner, Weber, and Bullara (1996), the mastery of the words taught via the self-correction procedures was about 30 points higher than for words taught by conventional strategies. However, whether or not the cause of this difference between the current study's findings and studies by McGuffin et al. (1997) and by Wirtz et al. (1996) may be attributed to procedural differences or the reduced number of weekly CWPT sessions in the current study is not clear.

In terms of the differences between the mastery of unknown words scores and pretest-posttest gain scores, the advantages of the mastery scores were apparent in some weeks. For example, the level of mastery was twice as large as pretest-posttest gain scores. This implies that the mastery indices reflected academic gains more directly than pretest-posttest gain scores, which might be suppressed by ceiling effects. For example, in the second week in Ms. A's class, the mean percentage correct was 62 for the pretest and 78 for the posttest; thus, the difference between the posttest was 16 points whereas the mastery of unknown words score was 83. Compared to the standard focus on pre- and posttest differences, mastery may be a better measure of students' improvement in peer tutoring.

With the pretest-posttest gain score measure, the mean percentage correct on the pretest should fall between 20% and 40% in order to avoid a ceiling effect (Greenwood et al., 1997). However, it has been reported that teachers often fail to achieve this criterion (Greenwood et al., 1992). An anecdotal report based on interviews with teachers suggested the following reasons.

First, teachers often follow the curriculum sequences of textbooks that are approved by the school or district policy. Thus, they are reluctant to adjust the difficulty levels of pretest scores to make the mean pretest scores fall between 20% and 40% by using different sets of words, as it would mean departing from the spelling sequence of the textbook. Second, some teachers think that no matter how easy or difficult the words being taught are, the majority of students cannot master them. Therefore, they refuse to change their spelling instructional procedures. Finally, some teachers are reluctant to adjust pretest scores because low scores may negatively affect students' self-esteem and their motivation to learn new words. Although these concerns may or may not be actualized, convincing

teachers to achieve the criterion by changing the difficulty level to meet the needs of a researcher is often difficult. By employing the mastery score, a more accurate measure may be achieved even when the pretest scores are higher than the desired 20% to 40% correct range.

The loss of known words score examined the effects of CWPT on the maintenance of words known to the students prior to instruction. Regardless of the intervention condition, the proportion of words lost remained low, suggesting that if students correctly spell the words on the pretests, they are likely to spell the same words correctly on the posttests. However, because the words correctly spelled on the pretests were also practiced during the instructional sessions in the current study, it remains unclear how many of the words known to the students prior to instruction might be maintained if no instruction were given. The more appropriate question may be how many of the words that are correctly spelled on the pretests would be maintained without being practiced.

Generalization across writing activities. The current study took mastery of spelling words into account when evaluating the generalization of spelling skills across writing activities. Other CWPT studies have not isolated generalization of words acquired prior to instruction from those acquired only through CWPT (i.e., those unknown prior to instruction). The results suggested that CWPT is only slightly more effective than teacher-led instruction in terms of generalization of spelling skills to sentence dictation tasks.

Although no comparable study in the CWPT and non-CWPT literature exists, a study by Diaz, McLaughlin, and Williams (1990) has some implications for the current study. These authors analyzed the percentage of spelling words correct on sentence dictation tests mastered via the Add-A-Word program, in which a new word was added as students master a word through the Cover-Copy-Compare procedures (i.e., students look at a model, cover the model, write the word from memory, and compare the written word with the model). In one condition, words were taught by the Add-A-Word program only. In another condition, sentence practice was in place in addition to the Add-A-Word program. The results indicated that the Add-A-Word program combined with additional sentence practice resulted in a mean percentage correct of 80 on the sentence dictation tests, while the Add-A-Word program alone resulted in a mean percentage correct of 65.

Because Diaz et al. (1990) did not employ teacher-led instruction as a baseline, it remains unknown whether the Add-A-Word program alone was superior to teacher-led instruction in the generalization of spelling skills to sentence dictation tasks. However, the results of the study by Diaz et al. suggested that sentence practice may be important for students to use words in sentences once words are taught. This may explain the smaller than expected effectiveness of CWPT in sentence dictation tests of the current study. That is, to achieve generalization, sentence practice may be necessary.

Retention. There was no apparent difference between CWPT and baseline phases in terms of the percentage correct on retention measures. Thus, the implication is that, regardless of the instructional strategies used (either peer tutoring or teacher-led instruction), generalization of spelling skills over time may not differ.

Although these results cannot be compared with prior CWPT studies because mastery was not considered (i.e., Harper et al., 1993; Maheady & Harper, 1987; Mallette et al., 1991), a few non-CWPT studies have taken mastery into account. For example, in a study by Wirtz et al. (1996), the bi-weekly retention tests resulted in 69 to 84% correct for the words taught via self-correction procedures versus 40 to 74% correct for the words taught via traditional instruction. Although this study demonstrated the superiority of the self-correction procedure on retention, when looking at the data of individual students, the effects varied among participants. In fact, one participant scored slightly higher on the retention test when the words were taught via traditional instruction. This suggests that the generalization effects of self-correction may or may not be more effective than conventional instruction, depending on the individual.

In summary, the current findings suggested that the introduction of CWPT may not result in a significant improvement in the generalization of spelling skills across other writing activities and over time compared to teacher-led instruction. Stokes and Baer (1977) recommended that researchers act as if there is no such thing as "free" generalization. Thus, teachers should employ other strategies such as a systematic review of materials and sentence practice besides and in addition to peer tutoring in order to ensure that generalization occurs.

Social validity. The majority of the students reported that they were moderately in favor of using peer tutoring procedures. This finding is similar to the results of previous studies in CWPT (e.g., Greenwood et al., 1987). In terms of teacher satisfaction, while Ms. A reported satisfaction with peer tutoring on most questions, Mr. B expressed many concerns. ranging from teacher and student training for peer tutoring, to student engagement during peer tutoring, to scheduling conflict between CWPT and other activities, and conflict resolution between students during peer tutoring sessions. This difference in opinions happened even though the review of the standard CWPT procedures for the teachers, the training for the students, and the assistance by the researcher were provided in approximately the same manner and degree across the two classes.

These differences in the two teachers' satisfaction reports might stem from differences in their students' behavior. Based on informal observation, Mr. B had much more difficulty in keeping students on task, and this difficulty in class management may have affected his satisfaction, resulting in a negative response overall. In fact, in a class with challenging behavior, any instructional strategy lacking strong behavior management procedures may lead to unsatisfactory results. Thus, it is recommended to always establish a strong stimulus control for CWPT and to ensure a high quality of implementation even when a class is managed poorly outside the peer tutoring in further research.

Limitations and Recommendations for Further Research

In addition to the challenges to peer tutoring implementation previously discussed, the current study had several limitations. First, changes were made in generalization testing procedures because of student attrition. This may have resulted in inaccurate measurements. Second, the quality of students' handwriting was often poor, resulting in words being considered incorrect because of careless writing even though it appeared that the students knew the words. A third concern relates to the

mastery score. The procedures used to calculate the mastery measure (e.g., tracking the history of each word for each student) were complex and time consuming. Without extraneous support for this measurement, teachers would have little reason to use this complex measure than simple pretest-posttest gain. Future research may wish to examine the utility of the mastery score under various conditions.

Fourth, other possibilities exist for measuring generalization. For generalization across other writing activities, sentence dictation tests were used. A true evaluation of generalization should be based on self-directed free writing by students (e.g., McNeish, Heron, & Okyere, 1992; Pratt-Struthers, Bartalamy, Williams, & McLaughlin, 1989; Pratt-Struthers, Struthers, & Williams, 1983). However, collecting the students' writing works from other classes and checking for all the spelling words taught was considered too time consuming given present resources. Additionally, the students might prefer using already known, familiar words compared to newly acquired but relatively unfamiliar words; hence, not all the words taught could be examined for generalization. Thus, as Harper et al. (1993) mentioned, their use of a sentence dictation test was a compromise. A future study might use a different time interval and other measures of generalization. An examination of the combined effect of peer tutoring and supplemental sentence practice might also be fruitful because peer tutoring alone was not found to be effective enough to drastically improve the generalization of spelling skills on sentence dictation tests.

For retention, the time interval between the end of instruction and the retention tests could have been longer. Using a longer time span may give a truer measure of generalization, because students should be able to use words in sentences any time once they have acquired the words. At the same time, use of a longer time interval increases the chance that the students re-encounter those words before the test. A future study might vary the time span to examine long-term retention and employ additional procedures to supplement peer tutoring to ensure the retention of the spelling skills.

In summary, the current study used various measures to examine the effects of CWPT, including pretest-posttest gain scores, mastery of unknown words scores, loss of known words scores, and generalization measures. However, because of low implementation fidelity and a reduced number of peer tutoring sessions per week, these effects probably underestimate those possible in well-implemented CWPT programs. Future research is needed to address this issue. Despite the limitations, however, the current study added uniquely to the body of peer tutoring literature by analyzing the mastery of unknown words, the generalization of spelling skills over time and across writing activities with consideration being given to mastery of words.

References

Busk, P. I., & Serlin, R. C. (1992). Meta-analyis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Campbell, J. M. (2006). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*, 234-246.

Coe, R. (2000). *What is an 'effect size'?: A guide for users.* Retrieved July 6, 2007, from http:// www.cemcentre.org

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, J. O., Heron, T. E., & Heward, W. L. (1987). *Applied behavior analysis.* Columbus, OH: Merrill.

Cooper, J. O., Valentine, J. C., & Charlton, K. (2000). The methodology of meta-analysis. In R. Gersten, E. P. Schiller, & S. Vaughn (Eds.), *Contemporary special education research: Syntheses of the knowledge base on critical instructional issues* (pp. 263-280). Hillsdale, NJ: Lawrence Erlbaum Associates.

Diaz, N. D., McLaughlin, T. F., & Williams, R. L. (1990). The effects of practicing words in sentences on generalization of spelling to written work with mildly mentally handicapped students. *Psychology in the Schools, 27*, 347-353.

Education Sciences Reform Act of 2002 (P.L. 107-279). Retrieved September 3, 2003, from http: //www.ed.gov/news/pressreleases/2002/11/11062002a.html *Gersten, R., & Baker, S. (Summer, 2000).* What we know about effective instructional practices for English-Language Learners. *Exceptional Children, 66(4)* 454-470.

Greenwood, C. R. (2006). *Research to improve the social and academic achievement of children and youth in an urban, poverty neighborhood: A personal perspective.* Retrieved July 6, 2007, from http:// www.umich.edu.

Greenwood, C. R., Arreaga-Mayer, C., Utley, C. A., Gavin, K. M., & Terry, B. J. (2001). Classwide peer tutoring learning management system: Applications with elementary-level English language learners. *Remedial and Special Education, 22*(1), 34-47.

Greenwood, C. R., Delquadri, J. C., & Carta, J. J. (1997). *Together we can!: Classwide peer tutoring to improve basic academic skills.* Longmont, CO: Sopris West.

Greenwood, C. R., Delquadri, J., & Hall, R. V. (1989). Longitudinal effects of classwide peer tutoring. *Journal of Educational Psychology, 81*, 371-383.

Greenwood, C. R., Dinwiddie, G., Bailey, V., Carta, J. J., Dorsey, D., Kohler, F. W., & Nelson, C. (1987). Field replication of classwide peer tutoring. *Journal of Applied Behavior Analysis, 20*, 151-160.

Greenwood, C. R., Terry, B., Arreaga-Mayer, C., & Finney, R. (1992). The classwide peer tutoring program: Implementation factors moderating students' achievement. *Journal of Applied Behavior Analysis, 25*, 101-116.

Harper, G. F., Mallette, B., Maheady, L., Parkes, V., & Moore, J. (1993). Retention and generalization of spelling words acquired using a peer-mediated instructional procedure by children with mild handicapping conditions. *Journal of Behavioral Education, 3*, 25-38.

Hershberger, S. L., Wallace, D. D., Green, S. B., & Marquis, J. G. (1999). Meta-analysis of single-case designs. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 109-132). Newbury Park, CA: Sage.

Kalkowski, P. (2001). *Peer and cross-age tutoring.* Retrieved July 5, 2007, from http://www.nwrel.org/scpd/sirs/9/c018.html

Kunsch, C. A., Jitendra, A. K., & Sood, S. (2007). The effects of peer-mediated instruction in mathematics for students with learning problems: A research synthesis. *Learning Disabilities Research & Practice, 22*(1), 1-12.

Leung, C. K., Marsh, H. W., & Craven, R. G. (2002). Are peer tutoring programs effective in promoting academic and self-concept in educational settings? A meta-analytical review. Retrieved 2007 from http://www.aare.edu.au

Maheady, L., & Harper, G. F. (1987). A class-wide peer tutoring program to improve the spelling test performance of low-income, third- and fourth-grade students. *Education and Treatment of Children, 10*, 120-133.

Mallette, B., Harper, G. F., Maheady, L., & Dempsey, M. (1991). Retention of spelling words acquired using a peer-mediated instructional procedure. *Education and Training in Mental Retardation, 2*6, 156-164.

Mathes, P. (1994). The efficacy of peer tutoring in reading for students with mild disabilities: A best-evidence synthesis. *School Psychology Review, 23*(1), 59-80.

McGuffin, M. E., Martz, S. A., & Heron, T. E. (1997). The effects of self-correction versus traditional spelling on the spelling performance and maintenance of third grade students. *Journal of Behavioral Education, 7*, 463-476.

McNaughton, D., Huges, C. A., & Clark, K. (1994). Spelling instruction for students with learning disabilities: Implications for research and practice. *Learning Disability Quarterly, 17*, 169-185.

McNeish, J., Heron, T. E., & Okyere, B. (1992). Effects of self-correction on the spelling performance of junior high school students with learning disabilities. *Journal of Behavioral Education, 2*, 17-27.

Mortweet, S. L., Utley, C. A., Walker, D., Dawson, H. L., Delquadri, J. C., Reddy, S., & Greenwood, C. R. (1999, Summer). Classwide peer tutoring: An effective spelling instructional procedure for students with educable mental retardation and their typical peers. *Exceptional Children, 65*(4), 524-536.

No Child Left Behind Act of 2001, 20 U.S.C. 70 § 6301 *et seq*. 2002.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education, 40*, 194-204.

Pratt-Struthers, J., Bartalamy, H. R., Williams, R. L., & McLaughlin, T. F. (1989). Effects of the add-a-word spelling program on spelling accuracy during creative writing: A replication across two classrooms. *B. C. Journal of Special Education, 13*, 151-158.

Pratt-Struthers, J., Struthers, T. B., & Williams, R. L. (1983). The effects of the add-a-word Spelling Program on spelling accuracy during creative writing. *Education and Treatment of Children, 6*, 277-283.

Rusch, F., Rose, T., & Greenwood, C. R. (1988). *Behavior analysis in special education.* Englewood Cliffs, NJ: Prentice Hall.

Sideridis, G. D., Utley, C., Greenwood, C. R., Delquadri, J., Dawson, H., Palmer, P., & Reddy, S. (1997). Classwide peer tutoring: Effects on the spelling performance and social interactions of students with mild disabilities and their typical peers in an integrated instructional setting. *Journal of Behavioral Education, 7*, 435-462.

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis, 10*, 349-367.

Swanson, H. L., & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research, 3*, 277-321.

Utley, C. A., Mortweet, S. L., & Greenwood, C. R. (1997). Peer-mediated instruction and interventions. *Focus on Exceptional Children.* Denver, CO: Love Publishing.

Wirtz, C. L., Gardner, R. III, Weber, K., & Bullara, D. (1996). Using self-correction to improve the spelling performance of low-achieving third graders. *Remedial and Special Education, 17*, 48-58.

Submitted: March 2, 2007
Revised: April 22, 2007
Accepted: May 12, 2007