NCEE 2011-4003

U.S. DEPARTMENT OF EDUCATION

The Impact of a Reading Intervention for Low-Literate Adult ESL Learners

The Impact of a Reading Intervention for Low-Literate Adult ESL Learners

December 2010

Larry Condelli
Stephanie Cronen
Johannes Bos
American Institutes for Research

Fannie Tseng
Jacklyn Altuna
Berkeley Policy Associates

Melanie Ali, Project Officer, Institute of Education Sciences

NCEE 2011-4003
U.S. DEPARTMENT OF EDUCATION

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

December 2010

This report was prepared for the Institute of Education Sciences under Contract No. ED-01-CO-0026/0025. The project officer was Melanie Ali in the National Center for Education Evaluation and Regional Assistance.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the reports.

To order copies of this report,

- Write to ED Pubs, Education Publications Center, U.S. Department of Education, P.O. Box 22207, Alexandria, VA 22304.
- Call in your request toll free to 1-877-4ED-Pubs. If 877 service is not yet available in your area, call 800-872-5327. Those who use a telecommunications device for the deaf (TDD) or a teletypewriter (TTY) should call 800-437-0833.
- Fax your request to 703-605-6794 or order online at http://www.edpubs.gov.

This report also is available on the IES website at http://ies.ed.gov/ncee/.

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-205-8113.

Acknowledgments

This study represents a collaborative effort of adult education programs, schools, teachers, researchers, students and the developers of the text *Sam and Pat.* We appreciate the willingness of the programs, schools, teachers and students to volunteer for the study, try new instructional materials and approaches, and respond to many requests for data and access to classrooms.

We were also fortunate to have the advice of an expert technical working group. Members included Russell Gersten, University of Oregon's Instructional Research Group; Daphne Greenberg, Georgia State University; Larry Hedges, Northwestern University; Robinson Hollister, Swarthmore College; Rebecca Maynard, formerly at the University of Pennsylvania, now at the Institute of Education Sciences (IES); Robin Scarcella, University of California-Irvine; and Anne Whiteside, City College of San Francisco. We also benefitted from the informed feedback on the report from Terry Salinger and Dan Sherman at the American Institutes for Research (AIR).

We would like to thank the *Sam and Pat* team who provided the teacher training and guidance during the study, including Jo Anne Hartel, Whit Hendon, Betsy Lowry, and John Strucker of World Education. We also thank the Mathematica team led by Susan Sprachman and Kathy Buek for leading the student pre- and post-testing for the study; John Sabatini, Jane Shore and Jennifer Lentini at the Educational Testing Service (ETS) for making sense of the test data and ensuring data quality; Anestine Hector-Mason, Dahlia Shaewitz, Suzannah Herrmann, and Lauren Amos for coordinating and conducting the classroom observations; Erin Hamilton, Andrea Harvey, and Erika Salomon for their support to the staff conducting data collection and processing; Castle Sinicrope at Berkeley Policy Associates (BPA) for analysis support; and all the AIR, BPA, Lewin Group, and Mathematica Policy Research, Inc. staff who worked with programs and schools to make the study happen: Mike Fishman, Mary Farrell, Karen Gardiner, Bret Barden, Rachel Augustin, Rachel Cook, Serena Retna, Sophie Shen, Lynne Blankenship, Julie Young, Anne Self, Chris Pefaure, and Savitha Moorthy.

Disclosure of Potential Conflicts of Interest[1]

The research team for this study consisted of a prime contractor, American Institutes for Research (AIR), and four subcontractors, BPA, ETS, the Lewin Group, and Mathematica Policy Research, Inc. None of these organizations or their key staff has financial interests that could be affected by findings from the Study of the Impact of a Reading Intervention for Low-Literate Adult ESL Learners. No one on the technical working group, convened by the research team approximately once per year to provide advice and guidance, has financial interests that could be affected by findings from the evaluation.

[1] Contractors carrying out research and evaluation projects for IES frequently need to obtain expert advice and technical assistance from individuals and entities whose other professional work may not be entirely independent of or separable from the particular tasks they are carrying out for the IES contractor. Contractors endeavor not to put such individuals or entities in positions in which they could bias the analysis and reporting of results, and their potential conflicts of interest are disclosed.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

Tables

Figures

THE IMPACT OF A READING INTERVENTION FOR LOW-LITERATE ADULT ESL LEARNERS

EXECUTIVE SUMMARY

According to the 2008 program year statistics from the U.S. Department of Education (ED), 44 percent of the 2.4 million students in the federally funded adult education program in the United States were English as a second language (ESL) students (ED, 2010). Of these, about 185,000 were at the lowest ESL level, beginning literacy. These students, many of whom face the dual challenge of developing basic literacy skills—including decoding, comprehending, and producing print—along with proficiency in English, represent a range of nationalities and cultural backgrounds. Although the majority of students come from Mexico and other Spanish-speaking countries, there are also students from Africa, India, the Philippines, China, Vietnam, and the Caribbean (Wrigley, Richer, Martinson, Kubo, & Strawn, 2003).

Adult basic education (ABE) and ESL programs, authorized by the Workforce Investment Act and also funded with state and local funds, are designed to assist students in their efforts to acquire literacy and language skills by providing instruction through local education agencies, community colleges, and community-based organizations. The content of instruction within ESL classes varies widely. It is often designed to assist students in their efforts to acquire literacy and language skills by providing a combination of oral language, competency-based work skills, and literacy instruction (Condelli, Wrigley, Yoon, Cronen, & Seburn, 2003). There is, however, little rigorous research that identifies effective instruction. A comprehensive review of published research studies on the effects of literacy interventions for ABE and adult ESL learners (Condelli & Wrigley, 2004) found that out of 17 adult education studies that used a rigorous methodology (i.e., quasi-experimental or randomized trials), only 3 included adult ESL learners (Diones, Spiegel, & Flugman, 1999; St. Pierre et al., 1995; St. Pierre et al., 2003). Furthermore, among the 3 studies that included adult ESL learners, only 1 presented outcomes for those learners, and that study experienced substantial methodological problems that limited the validity of the findings (e.g., a 40 percent overall attrition rate and different attrition rates in the intervention vs. control groups; Diones et al., 1999).

To help improve research-based knowledge of effective instruction for low-literate ESL learners, the National Center for Education Evaluation and Regional Assistance of ED's Institute of Education Sciences contracted with the American Institutes of Research (AIR) to conduct a Study of the Impact of a Reading Intervention for Low-Literate Adult ESL Learners. The intervention studied was the basal reader *Sam and Pat, Volume I,* published by Thomson-

Heinle (2006). The study team consisted of AIR, Berkeley Policy Associates (BPA), the Lewin Group, Mathematica Policy Research, Inc., Educational Testing Service (ETS), and World Education.

The goal of this study was to test a promising approach to improving the literacy skills of low-literate adult ESL students under real-world conditions. In their review of the research on ESL instruction in related fields, including adult second language acquisition, reading and English as a foreign language instruction, Condelli & Wrigley (2004) concluded that instruction based on a systematic approach to literacy development was a promising intervention for low-literate adult ESL learners that would be valuable to study (Brown et al., 1996; Cheek & Lindsay, 1994: Chen & Graves, 1995; Carrell, 1985; Rich & Shepherd, 1993; Roberts, Cheek & Mumm, 1994). Specifically, the factors identified as defining a systematic approach to literacy development included: (1) a comprehensive instructional scope that includes direct instruction in phonics, fluency, vocabulary development and reading comprehension, (2) a strategic instruction sequence, (3) a consistent instructional format, (4) easy-to-follow lesson plans, and (5) strategies for differentiated instruction.

Sam and Pat was selected as the focus of the study because it offers an approach to literacy development that is systematic, direct, sequential, and multi-sensory. It also includes multiple opportunities for practice with feedback. Consistent with characteristics identified as promising by Condelli & Wrigley (2004), *Sam and Pat* provides opportunities for cooperative learning, real world tasks, and an explicit focus on reading. In addition, the text was developed for and had been used by the developers with students similar to the study population (literacy level ESL learners).

The impact study used an experimental design to test the effectiveness of *Sam and Pat* in improving the reading and English language skills of adults enrolled in 66 ESL literacy classes at 10 sites. The study addressed three key research questions:

1. How effective is instruction based on the *Sam and Pat* textbook in improving the English reading and language skills of low-literate adult ESL learners compared to instruction normally provided in adult ESL literacy classes?

2. Is *Sam and Pat* effective for certain subgroups of students (e.g., native Spanish speakers)?

3. Is there a relationship between the amount of instruction in reading or English language skills and reading and English language outcomes?

This report describes the implementation of *Sam and Pat* at the study sites, compares the instruction and student attendance in *Sam and Pat* classes with that in the standard adult ESL classes, and examines the impact of *Sam and Pat* on reading and English language outcomes. In addition, the report examines the relationship between instruction, attendance, and student outcomes.

The study produced the following key results:

- ***More reading instruction was observed in Sam and Pat classes, while more English language instruction was observed in control classes.*** The *Sam and Pat* classrooms spent more time on reading development instruction (66 percent of observed intervals in *Sam and Pat* classrooms compared to 19 percent in control classrooms), and the difference was statistically significant. Conversely, the control classrooms spent more time on English language acquisition instruction (68 percent of observed intervals in control classrooms compared to 27 percent in *Sam and Pat* classrooms), and this difference was also statistically significant.

- ***Although students made gains in reading and English language skills, no differences in reading and English language outcomes were found between students in the Sam and Pat group and students in the control group.*** On average, students participating in the study made statistically significant gains in reading and English language skills over the course of the term (effect sizes of 0.23 to 0.40). However, there were no statistically significant impacts of *Sam and Pat* on the reading and English language outcomes measured for the overall sample.

- ***There were no impacts of Sam and Pat on reading and English language outcomes for five of six subgroups examined.*** For students with relatively lower levels of literacy at the start of the study, there was some suggestive evidence of a positive impact on reading outcomes.[2] Among students with lower levels of literacy at the beginning of the term, *Sam and Pat* group students scored higher on the Woodcock Johnson word attack (decoding) assessment than control group students (effect size = 0.16). Because this difference was not significant after adjusting for multiple comparisons, however, it is possible that the effect is due to chance alone.

[2] Lower literacy was defined as scoring at a Grade 2 equivalent or below on the Woodcock Johnson Letter-Word Identification and Word Attack subtests (raw scores of 31 and 9, respectively).

Summary of Study Design and Methods

The study was designed to estimate the impact of *Sam and Pat* relative to standard ESL instruction (i.e., the kind of instruction ESL students in study sites would receive in the absence of the study) on reading and English language outcomes.

The evaluation employed a randomized research design that included the following:

- 10 adult education program sites;
- 33 teachers;
- 66 classes; and
- 1,344 low-literate adult ESL learners.

The program sites were a purposive sample. From among the states with the largest adult ESL enrollments, we selected sites that had enrollments of adult ESL literacy learners large enough to support the study design, 2 or more classes for ESL literacy students that met at the same time and in the same location, and an enrollment process that would accommodate random assignment.

Within each site, teachers and students were randomly assigned to one of two groups:

- The *Sam and Pat* group, which was intended to include a minimum of 60 hours of *Sam and Pat*-based instruction per term, with any remaining class time being spent on the standard instruction provided by the program; and
- The control group, which consisted of the standard instruction provided by the program.

Teachers (or classes) within each program site were randomly assigned in pairs, so that, within each pair, the *Sam and Pat* and control class met at the same time, in the same or an adjacent building, and for the same number of hours. Data collection for the study occurred between September 2008 and May 2009 with two cohorts of students, one that attended in fall 2008 and the second in spring 2009. Students were tested on the study's battery of assessments, which included tests of reading and English language skills at the beginning of the term and after about 12 weeks of instruction. A description and schedule for the study's data collections are provided in Table ES.1.

The following tests were selected to measure the range of skills that could potentially be impacted by *Sam and Pat*–based instruction:

Reading Skills

- Woodcock-Johnson Letter-Word Identification (WJID; Woodcock, McGrew, & Mather, 2001)
- Woodcock-Johnson Passage Comprehension (WJPC; Ibid.)
- Woodcock-Johnson Word Attack (WJWA; Ibid.)
- SARA Decoding (SARA Dec; Sabatini & Bruce, in press)

English Language Skills

- Oral and Written Language Scales (OWLS; Carrow-Woolfolk, 1996)
- Receptive One-Word Picture Vocabulary Test (ROWPVT; Brownell, 2000)
- Woodcock-Johnson Picture Vocabulary Test (WJPV; Woodcock, McGrew, & Mather, 2001)

Table ES.1: Data Collection Schedule

Data Collection	Respondent	Summer 2008	Fall 2008	Spring 2009	Type of Data
Teacher Data Form (2008)	Teachers	X	X		Teacher background information
Teacher Data Form (2009)	Teachers			X	Descriptive information about instructional materials used and *Sam and Pat* implementation
Student Intake Form	Site Staff on Behalf of Students		X	X	Student background information
Reading and English Language Pre-Tests	Students		X	X	Pre-test data
Reading and English Language Post-Tests	Students		X	X	Outcomes data
Daily Student Attendance Sheets	Teachers		X	X	Dosage/exposure to instruction
Classroom Observations	Evaluation Staff		X	X	Descriptive information about instruction in both groups

The basic analytic strategy for assessing the impacts of *Sam and Pat* was to compare reading and English language outcomes for students who were randomly assigned to either the *Sam and Pat* or the control group, after controlling for student and teacher background characteristics (e.g., gender and ethnicity). The average outcome in the control group represents an estimate of the scores that would have been observed in the *Sam and Pat* group if they had not received the intervention; therefore, the difference in outcomes between the *Sam and Pat* and control groups provides an unbiased estimate of the impacts of *Sam and Pat*.

The Adult ESL Literacy Intervention: *Sam and Pat*

The *Sam and Pat* textbook (Hartel, Lowry, & Hendon, 2006) is described by the developers as a basal reader or textbook that tailors the methods and concepts of the Wilson and Orton-Gillingham reading systems developed for native speakers of English (Wilson & Schupack, 1997; Gillingham & Stillman, 1997) to meet the needs of adult ESL literacy level learners.[3] *Sam and Pat* was designed to incorporate the following components of the Wilson/Orton-Gillingham systems:

- A focus on moving students systematically and sequentially from simple to complex skills and materials;
- The use of multisensory approaches to segmenting and blending phonemes (e.g., sound tapping);
- An emphasis on alphabetics/decoding, fluency, vocabulary, and reading comprehension;
- The use of sound cards and controlled text (wordlists, sentences, stories) for practicing skills learned; and
- Continual review (cumulative instruction) of letters, sounds, and words already learned.

However, when writing *Sam and Pat,* the developers made variations on the base reading systems to make the text useful and relevant to the adult ESL literacy population for which the text was designed. Specifically, *Sam and Pat* differs from the base reading systems on four dimensions:

- The sequence in which the sounds of English are taught;
- The words chosen for phonics and vocabulary study;
- The simplification of grammar structures presented; and
- The added bridging of systematic reading instruction to ESL instruction.

---

[3] Although there is no available research on the effectiveness of *Sam and Pat*, the textbook and its accompanying training and technical support is based on these two reading systems (Wilson & Orton-Gillingham), which have shown promise in teaching struggling readers (Adams, 1991; Clark & Uhry, 1995; Kavenaugh, 1991; Torgesen et al., 2006).

Building on the components of the earlier reading systems, *Sam and Pat* was therefore designed to (1) sequence the teaching of English sound and spelling patterns to ESL students by moving from a focus on simple to complex literacy skills and materials, (2) provide a controlled basal that follows this sequence of patterns, (3) use a simplified grammar, (4) embed a controlled vocabulary that is relevant to the lives of this population of students, and (5) include a collection of stories that are based on simplified themes from daily life.

There are two volumes of *Sam and Pat,* and the Volume 1 literacy textbook was evaluated by this study. It is organized into a total of 22 multi-component lessons. The lessons follow what the developers consider to be an optimal sequence for introducing English phonics and high-frequency English sight words to non-native speakers of English. However, the sequence in which English vowels and consonant sounds are introduced has been modified from that usually used in approaches such as the Wilson and Orton-Gillingham reading systems. For example, like the Wilson System, *Sam and Pat* begins with the short-a sound, but short-a is followed several lessons later by short-u, rather than short-i. This modification was made to provide the maximum sound contrasts for the short vowel sounds that are notoriously challenging for English language learners to discriminate.

Although the current study was a large-scale effectiveness study, we took measures intended to facilitate the implementation of *Sam and Pat.* The *Sam and Pat* developers provided the teachers assigned to the *Sam and Pat* group with training and technical assistance on implementing *Sam and Pat.* The training was developed specifically for the study, and included a 3-day training before the start of the fall 2008 term and a 2-hour refresher webinar before the start of the winter 2009 term. The technical assistance provided to all *Sam and Pat* teachers included a site visit to observe and provide feedback early in the fall term, biweekly phone calls during the first 2 months of the fall term, and additional assistance as needed in response to phone calls and e-mails from teachers. The developers also provided 1 day of individualized assistance in person early in the winter term to teachers who appeared to be having difficulty implementing *Sam and Pat.*

Summary of Study Findings

Two-thirds of *Sam and Pat* Classes Observed Demonstrated Evidence of Implementing *Sam and Pat* as Intended

About two-thirds (65 percent) of the *Sam and Pat* classes observed met the study's instructional fidelity criteria regarding the use of *Sam and Pat* materials and engagement in reading instruction. More specifically, these teachers met the following criteria that were established in collaboration with the developers before the study began:

- *Sam and Pat* materials must be used for a minimum of 1 hour of instruction per class day;
- Each class day must include at least 1 hour of instruction in reading development; and
- Each class day, instruction should occur in at least three of the reading development instructional areas (e.g., phonics, fluency, reading comprehension).

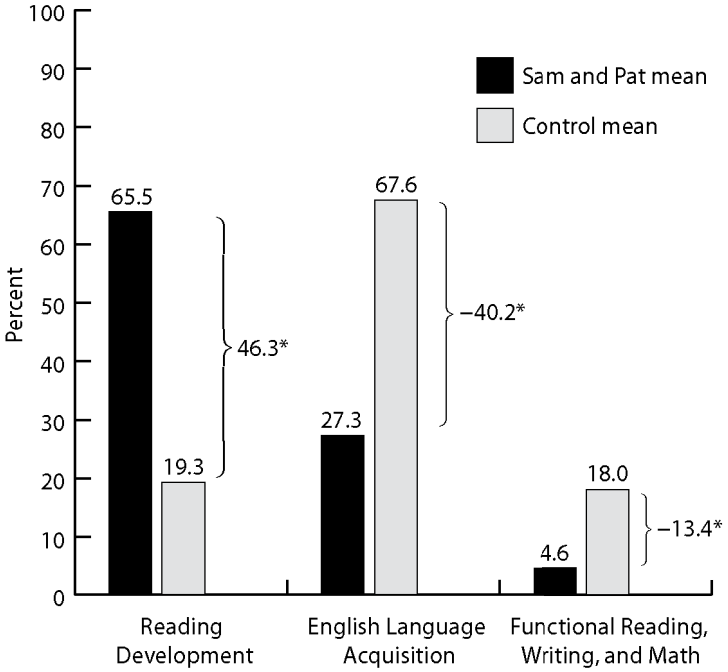
Because we did not observe all hours of instruction throughout the term, we cannot determine how many hours of *Sam and Pat* instruction were received by each student. However, students in the *Sam and Pat* group met for an average of 79 hours total over the course of the term (not shown in tables). The *Sam and Pat* developers recommended that the text be implemented for a minimum of 60 hours per term.

More Reading Instruction Observed in *Sam and Pat* Classes, While More English Language Instruction Observed in Control Classes

The *Sam and Pat* classrooms spent more time on reading development instruction than control classrooms (66 percent vs. 19 percent of observed time intervals, respectively), and the difference was statistically significant (Figure ES.1). Conversely, the control classrooms spent more time on English language acquisition instruction than did *Sam and Pat* classrooms (68 percent vs. 27 percent of observed time intervals, respectively), and this difference was statistically significant. The control classrooms also spent more time on functional reading, writing and math instruction (content related to English language acquisition instruction) than *Sam and Pat* classrooms (18 percent vs. 5 percent of observed time intervals, respectively).[4]

[4] We can only characterize implementation by reporting that (1) 65 percent of *Sam and Pat* classes met the study's fidelity criteria, and (2) significantly more reading instruction was delivered in these classes, as compared to the control group classes.

Figure ES.1: Percent of Observed Instructional Intervals Spent in Key Instructional Areas, by Group



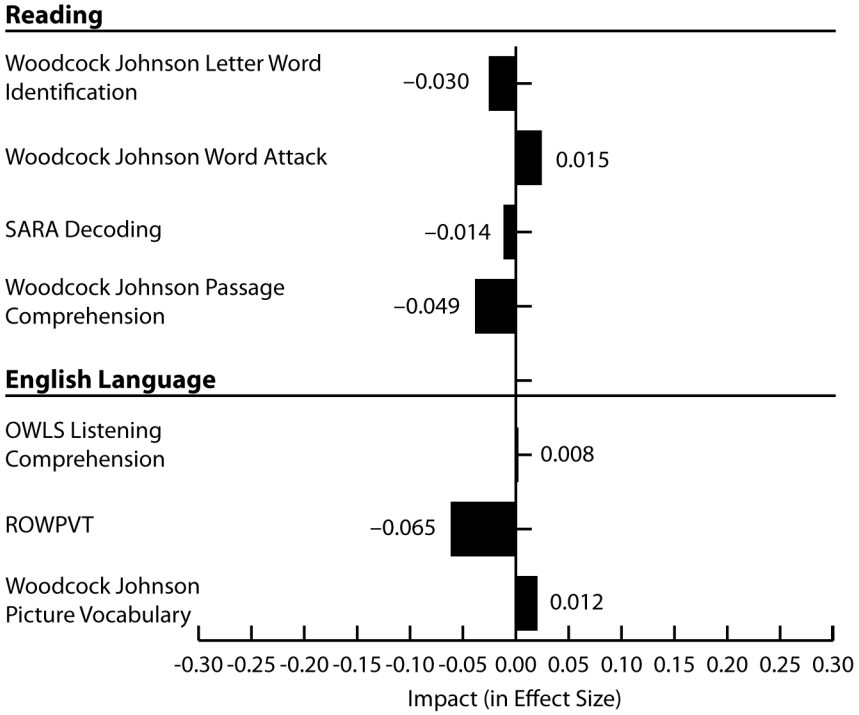
* Indicates a difference that is significant at the 0.05 level, based on a 2-tailed t-test.
Notes: N = 980 observation intervals for *Sam and Pat* group and 1,034 intervals for control group. Details may not sum to totals. Practices may be coded under multiple instructional areas during any one interval.
Source: Adult ESL Literacy Impact Study classroom observation protocol.

Students Made Gains, but There Were No Overall Impacts of *Sam and Pat* on Students' Reading and English Language Skills

On average, students participating in the study made statistically significant gains over the course of the term (effect sizes of 0.23 to 0.40). These gains are equivalent to 1 to 2 months of growth on the reading assessments, and 5 to 6 months of growth on the English language assessments.[5] However, there were no statistically significant impacts of *Sam and Pat* on the reading and English language outcomes measured for the overall sample (Figure ES.2). Effect sizes ranged from -0.06 to 0.01.

[5] It should be noted that publisher guidelines for the grade and age equivalent calculations used to determine months of gains are based on norming populations that differ from the study population. (The WJ assessments were normed on a nationally representative sample of U.S. residents aged 2 to 90+; the OWLS on a representative U.S. sample aged 3 to 21 years; and the ROWPVT on a representative U.S. sample aged 2 to 18 years.) No norming data exist for low-literate adult ESL learners. Additionally, the study used simplified or translated testing instructions when students did not appear to understand the tester's directions. For these reasons, the number of months of growth should be interpreted with caution.

Figure ES.2: Impact of *Sam and Pat* on Reading and English Language Skills: Differences Between *Sam and Pat* and Control Groups at the End of the Term



Notes: N = 580 for *Sam and Pat* group and 557 for control group. No impacts were statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

No Impacts of *Sam and Pat* on Reading and English Language Outcomes Found for Subgroups Based upon Student Native Language and Cohort

There were no statistically significant impacts found for students with a non-Roman-based alphabet background, native Spanish speakers, students from the first study cohort, or students from the second study cohort. Effect sizes ranged from –0.14 to 0.09.

Some Suggestive Evidence of a Positive Impact on Reading Outcomes for Lower Literacy Students

No statistically significant impacts were found for the students in the sample with relatively higher literacy levels (effect sizes ranged from –0.08 to 0.03). However, there was a suggestive finding for students who tested in the lower literacy score range at the beginning of the term. Within this subgroup, *Sam and Pat* group

students scored higher on the Woodcock Johnson word attack (decoding) assessment than control group students (effect size = 0.16). Because this difference was not statistically significant after adjusting for multiple comparisons, however, it is possible that the effect is due to chance alone. No impacts were found for the lower literacy students on the other reading and English language outcomes measured.

Student Exposure to Reading or English Language Instruction Unrelated to Most Reading and English Language Outcomes Measured, Although Weak Relationships Found Between Exposure to Instruction and One English Language Outcome

Student exposure to instruction was measured by the combination of reading and English language instruction provided in study classes and the number of hours students attended study classes. No statistically significant relationships were found between exposure to instruction and any of the reading outcomes measured and two of the three English language outcomes measured. However, the amount of exposure to English language instruction was positively and statistically significantly correlated with ROWPVT scores. The opposite pattern was found for reading instruction; exposure to reading instruction had a negative and statistically significant relationship with scores on the ROWPVT. However, the standardized coefficients in both cases were small (0.034 and –0.032, respectively). As an example, the 0.034 coefficient on the ROWPVT assessment indicates that, after controlling for total student attendance hours, an increase of 10 percent in the number of English language instruction intervals a student attended is associated with a 0.34 point increase on the test (which had a sample mean of 29). In addition, similar to the student attendance results, we cannot rule out the possibility that the statistically significant relationships were driven by other factors. Therefore, these findings should be interpreted with caution.

Generalizability of the Study Findings

The findings reported in this summary are limited to the specific intervention tested *(Sam and Pat, v. 1)* as implemented within the types of sites included in the study. For example, the study was implemented in sites large enough to offer at least 2 literacy level classes at the same time and location, within a subset of states that have the highest adult ESL enrollments. It is not known whether, or how, the results may generalize to other contexts.

CHAPTER 1:
INTRODUCTION AND OVERVIEW

According to the 2008 program year statistics from the U.S. Department of Education (ED), 44 percent of the 2.4 million students in the federally funded adult education program in the United States were English as a second language (ESL) students (ED, 2010). Of these, about 185,000 were at the lowest ESL level, beginning literacy. These students, many of whom face the dual challenge of developing basic literacy skills—including decoding, comprehending, and producing print—along with proficiency in English, represent a range of nationalities and cultural backgrounds. Although the majority of students come from Mexico and other Spanish-speaking countries, there are also students from Africa, India, the Philippines, China, Vietnam, and the Caribbean (Wrigley, Richer, Martinson, Kubo, & Strawn, 2003).

Adult basic education (ABE) and ESL programs, authorized by the Workforce Investment Act and also funded with state and local funds, are designed to assist students in their efforts to acquire literacy and language skills by providing instruction through local education agencies, community colleges, and community-based organizations. The content of instruction within ESL classes varies widely. It is often designed to assist students in their efforts to acquire literacy and language skills by providing a combination of oral language, competency-based work skills, and literacy instruction (Condelli, Wrigley, Yoon, Cronen, & Seburn, 2003). There is, however, little rigorous research that identifies effective instruction. A comprehensive review of published research studies on the effects of literacy interventions for ABE and adult ESL learners (Condelli & Wrigley, 2004) found that out of 17 adult education studies that used a rigorous methodology (i.e., quasi-experimental or randomized trials), only 3 included adult ESL learners (Diones, Spiegel, & Flugman, 1999; St. Pierre et al., 1995; St. Pierre et al., 2003). Furthermore, among the 3 studies that included adult ESL learners, only 1 presented outcomes for those learners, and that study experienced substantial methodological problems that limited the validity of the findings (e.g., a 40 percent overall attrition rate and different attrition rates in the intervention vs. control groups; Diones et al., 1999).

To help improve research-based knowledge on instruction for low-literate ESL learners, the National Center for Education Evaluation and Regional Assistance of ED's Institute of Education Sciences contracted with the American Institutes of Research to conduct a Study of the Impact of a Reading Intervention for Low-Literate Adult ESL Learners. The study is designed to evaluate the effectiveness of instruction based on a promising literacy textbook—*Sam and Pat*—using a random assignment design.

Selection of the Adult ESL Literacy Intervention

The goal of this study was to test a promising approach to improving the literacy skills of low literacy level adult ESL students under real-world conditions. In their review of the research on ESL instruction in related fields, including adult second language acquisition, reading and English as a foreign language instruction, Condelli & Wrigley (2004) concluded that instruction based on a systematic approach to literacy development was a promising intervention for low-literate adult ESL learners that would be valuable to study (Brown et al., 1996; Cheek & Lindsay, 1994: Chen & Graves, 1995; Carrell, 1985; Rich & Shepherd, 1993; Roberts, Cheek & Mumm, 1994). Specifically, the factors identified as defining a systematic approach to literacy development included: (1) a comprehensive instructional scope that includes direct instruction in phonics, fluency, vocabulary development and reading comprehension, (2) a strategic instruction sequence, (3) a consistent instructional format, (4) easy-to-follow lesson plans, and (5) strategies for differentiated instruction.

To select a literacy intervention for the study, an open competition was first held via a public solicitation for proposals. In addition to posting a solicitation for proposals in public forums such as discussion listservs, the study team conducted targeted outreach to 20 potential intervention providers. The potential intervention providers were identified through web searches as well as based upon the study team's knowledge of existing textbooks. When no proposals were received, follow-up calls to prospective intervention providers were made; the most common reason cited for not submitting a proposal was that the developer's existing intervention was not designed specifically for literacy level adult ESL students, and would require substantial revision. Study staff then contacted four additional intervention providers who had been recommended by experts in the field. Through a second round of proposals and curricula samples requested directly of these providers, the four providers' proposals received were found to be unacceptable by an external panel. *Sam and Pat* was recommended to IES and was subsequently selected as the focus of the study because it offers an approach to literacy development that is systematic, direct, sequential, and multi-sensory. It also includes multiple opportunities for practice with feedback. Consistent with characteristics identified as promising by Condelli & Wrigley (2004), *Sam and Pat* is designed to provide opportunities for cooperative learning, real world tasks, and an explicit focus on reading. In addition, the text was developed for and had been used by the developers in their own classrooms with students similar to the study population (adult literacy level ESL learners).

Research Questions

The study addressed three key research questions:

1. How effective is instruction based on the *Sam and Pat* textbook in improving the English reading and language skills of low-literate adult ESL learners compared to instruction normally provided in adult ESL literacy classes?

2. Is *Sam and Pat* effective for certain subgroups of students (e.g., native Spanish speakers)?

3. Is there a relationship between the amount of instruction in reading or English language skills and reading and English language outcomes?

As the research questions indicate, the purpose of the study was to test the effectiveness of a specific intervention *(Sam and Pat, v. 1)*. The findings from the study may not generalize to other literacy interventions for adult ESL learners.

Summary of Study Design

The study was designed to estimate the impact of *Sam and Pat*-based instruction and professional development, relative to standard ESL instruction (i.e., the kind of instruction ESL students in study sites would receive in the absence of the study).

The evaluation employed a randomized research design that included the following:

- 10 adult education program sites;
- 33 teachers;
- 66 classes; and
- 1,344 low-literate adult ESL learners.

The program sites were a purposive sample. From among the states with the largest adult ESL enrollments, we selected sites that had enrollments of adult ESL literacy learners large enough to support the study design, had 2 or more classes for ESL literacy students that met at the same time and in the same location, and had an enrollment process that would accommodate random assignment.

Within each site, teachers and students were randomly assigned to one of two groups:

- The *Sam and Pat* group, which was intended to include a minimum of 60 hours of *Sam and Pat*-based instruction per term, with any remaining class time being spent on the standard instruction provided by the program; and
- The control group, which consisted of the standard instruction provided by the program.

Teachers (or classes) within each program site were randomly assigned in pairs, so that, within each pair, the *Sam and Pat* and control class met at the same time, in the same or in an adjacent building, and for the same number of hours. Across the study sites, the total number of class hours varied and ranged from approximately 60 to 225 total hours per term, depending on the site's course schedule. Data collection for the study occurred between September 2008 and May 2009 with two cohorts of students, one that attended in fall 2008 and the second in spring 2009. Students were tested on the study's battery of assessments, which included tests of reading and English language skills, at the beginning of the term and after about 12 weeks of instruction.

Standard ESL Instruction: The Control Group

Adult ESL instruction encompasses a range of approaches and content, but its goal is to help students acquire facility with the English language and function in everyday life. Content includes oral language development, grammar, vocabulary, and cultural topics. ESL instruction may also include a life skills (functional) approach to language, such as learning how to complete forms, interpret labels, and negotiate tasks such as shopping and dealing with schools, doctors, and government agencies (Celce-Murcia, 2001; Crandall & Peyton, 1993).

Standard ESL instruction assumes that students are already literate in their first language; therefore, it does not usually focus on phonics or the other basic reading skills emphasized in *Sam and Pat* (Wrigley & Guth, 1992; Wrigley, Chisman, & Ewen, 1993). Although nationally representative data on adult ESL instruction or textbook use is not available, in a study of instruction of 38 adult ESL literacy classes in seven states, Condelli et al. (2003) found that ESL instruction focused on developing oral English language, vocabulary, and life skills. Of the 38 classes, 7 included reading instruction for more than half of the total class time, and 31 spent more than 40 percent of the class time on second language instruction—despite the fact that all of these classes were designated as "literacy level" (i.e., intended for low-literate students). Furthermore, across all classes, a majority of total class time (51 percent) was spent on second language instruction. When

reading instruction did occur, it was considered by the researchers to be unsystematic and of short duration (Condelli et al., 2003).

The Adult ESL Literacy Intervention: *Sam and Pat*

Overview of Sam and Pat

The *Sam and Pat* textbook (Hartel, Lowry, & Hendon, 2006) is described by the developers as a basal reader or textbook that tailors the methods and concepts of the Wilson and Orton-Gillingham reading systems developed for native speakers of English (Wilson & Schupack, 1997; Gillingham & Stillman, 1997) to meet the needs of adult ESL literacy level learners.[6] *Sam and Pat* was designed to incorporate the following components of the Wilson/Orton-Gillingham systems:

- A focus on moving students systematically and sequentially from simple to complex skills and materials;
- The use of multisensory approaches to segmenting and blending phonemes (e.g., sound tapping);
- An emphasis on alphabetics/decoding, fluency, vocabulary, and reading comprehension;
- The use of sound cards and controlled text (wordlists, sentences, stories) for practicing skills learned; and
- Continual review (cumulative instruction) of letters, sounds, and words already learned.

However, when writing *Sam and Pat,* the developers made variations on the base reading systems to make the text useful and relevant to the adult ESL literacy population for which *Sam and Pat* was designed. Specifically, *Sam and Pat* differs from the base reading systems on four dimensions:

- The sequence in which the sounds of English are taught;
- The words chosen for phonics and vocabulary study;
- The simplification of grammar structures presented; and
- The added bridging of systematic reading instruction to ESL instruction.

Building on the components of the earlier reading systems, *Sam and Pat* was therefore designed to (1) sequence the teaching of English sound and spelling patterns to ESL students by moving from a focus on simple to complex literacy skills and materials, (2) provide a controlled basal that follows this sequence of

[6] Although there is no available research on the effectiveness of *Sam and Pat*, the textbook and its accompanying training and technical support is based on these two reading systems (Wilson & Orton-Gillingham), which have shown promise in teaching struggling readers (Adams, 1991; Clark & Uhry, 1995; Kavenaugh, 1991; Torgesen et al., 2006).

patterns, (3) use a simplified grammar, (4) embed a controlled vocabulary that is relevant to the lives of this population of students, and (5) include a collection of stories that are based on simplified themes from daily life.

There are two volumes of *Sam and Pat,* and the Volume 1 literacy textbook was evaluated by this study. It is organized into a total of 22 multi-component lessons. The lessons follow what the developers consider to be an optimal sequence for introducing English phonics and high-frequency English sight words to non-native speakers of English. However, the sequence in which English vowels and consonant sounds are introduced has been modified from that usually used in approaches such as the Wilson and Orton-Gillingham reading systems. For example, like the Wilson System, *Sam and Pat* begins with the short-a sound, but short-a is followed several lessons later by short-u, rather than short-i. This modification was made to provide the maximum sound contrasts for the short vowel sounds that are notoriously challenging for English language learners to discriminate.

Sam and Pat is also designed to introduce and build basic English speaking and reading vocabulary, as well as foundational skills in basic English grammar. Both the vocabulary and grammar components are focused on the functional needs of new immigrants in the domains of work, their children's school, shopping, family life, and interactions with the medical system.

Each lesson contains a chapter of an ongoing story that follows the daily lives and adventures of an immigrant family headed by the title characters. Like the basal readers written for English speaking adult beginning readers, the text is controlled; that is, it only contains words that follow phonics patterns that have been previously taught, as well as sight words that have also been taught. This is intended to give learners the opportunity to develop word reading skills and fluency in meaningful text, without encountering phonics patterns and sight words they have not been taught.

In addition, because *Sam and Pat* was created for ESL literacy students, the text has also been controlled for vocabulary and grammar content; learners only encounter word meanings and grammar patterns that have been previously introduced in accompanying oral and written activities. As the Introduction explains, "Only simple words that students might encounter in their daily lives are used in the stories. The stories are written with simplified grammar, since long sentences and complex structures can interfere with comprehension" (Hartel et al., 2006, p. v).

Intended Use of Sam and Pat

Sam and Pat was designed to provide learners with listening, speaking, reading, and writing activities that are sequenced and designed to reinforce each other. Each lesson is intended by the developers to include at least 1 day (approximately 2.5 hours) per week of pre-reading instruction and at least 1 day per week of decoding and reading comprehension instruction, with additional review and reteaching added as determined by the teacher.

The goal of the pre-reading instruction day is to explain, demonstrate, and provide practice opportunities for the new phonics, sight words, vocabulary, and grammar prior to reading each new chapter of *Sam and Pat.* The skill areas targeted on pre-reading instruction days include the following:

- Review/rereading a story for fluency;
- Review of names and sounds of letters learned previously, and introduction of new sounds;
- Pre-reading conversation, grammar, and/or vocabulary practice;
- Sight word instruction (review and new);
- Phonics instruction (review and new); and
- Pre-reading pictures for the upcoming story.

The skill areas targeted on decoding/reading comprehension instruction days include continued practice from the previous day as well as new activities:

- Review/rereading a story for fluency;
- Review of names and sounds of letters learned previously, and introduction of new sounds;
- Pre-reading review of conversation and vocabulary from previous day;
- Sight word instruction (review and new);
- Phonics instruction (review and new);
- Pre-reading review of pictures from the previous day;
- Reading the new story; and
- Written exercises based on text.

As implied by the inclusion of the target skill "conversation" during both days of instruction, literacy instruction based on *Sam and Pat* does not include reading and writing activities exclusively; speaking and listening activities also take place connected to the activities in the basal.

Several types of oral language activities, tied to the content, precede the story part of each chapter. For example, Lesson 1 begins with a line drawing of the characters *Sam and Pat* and the text, "This is Sam. This is Pat. They are *Sam and Pat.*" Before reading this chapter with the students, a teacher might conduct a

spoken language activity. For instance, she may write each learner's name on a place card. She would then point to a person and his place card and say, "This is Juan." Then she would point to another person and her card say, "This is Marie." After giving the class numerous opportunities to practice these phrases in different combinations and with each others' names, the teacher would next point to both learners and say, "They are Juan and Marie," followed by more practice as before.

The intended purpose of *Sam and Pat* is to provide ESL literacy learners with multiple opportunities for repetition, guided practice, and review. The developers report that when used correctly and in combination with appropriate spoken language activities, *Sam and Pat* requires teachers to spend about 7 class hours on each chapter of the book, including pre-reading and decoding/comprehension instruction, reteaching as necessary, and supporting oral language activities. The developers instructed the *Sam and Pat* teachers to implement the text for a total of approximately 60 hours per term, or 5 hours per week in a standard 12 week term. At that rate, an ESL literacy class would be expected to spend over a week on each chapter, and approximately 2 terms to complete the 22 chapters of *Sam and Pat,* Volume. 1. The *Sam and Pat* teachers were therefore expected to complete an average of 9 out of the 22 chapters each term. Teachers were told that they could implement more hours of *Sam and Pat;* however, the 5 hours per week recommendation was based on the developers' understanding of what is feasible given the amount of time classes met each week.

Teacher Training and Follow-Up Technical Assistance

Although the current study was a large-scale effectiveness study, we took measures intended to facilitate the implementation of *Sam and Pat.*[7] Prior to the fall 2008 term, the *Sam and Pat* developers provided the teachers assigned to the *Sam and Pat* group with 3 days of intensive training on the implementation of *Sam and Pat.* The training was developed specifically for the study, and included discussions about the origins and rationale for the approach, the unique characteristics of ESL literacy level learners based on current research, the structure and terminology of *Sam and Pat,* the components of reading and oral language instruction, the Lesson Plan template developed to support implementation, *Sam and Pat* reading and oral language instructional techniques and activities, and classroom organization and management. It also included

[7] The developers of *Sam and Pat* have not provided training to teachers implementing the text outside of the study, and therefore the training and technical assistance provided to teachers during the study represent possible differences from what teachers might receive from another source if implementing *Sam and Pat* in the field. However, since there are no data available on either the extent to which *Sam and Pat* is used in the field, or on the availability of other sources of training on the use of *Sam and Pat*, we cannot determine how representative the study conditions were of the national population of teachers using this text.

multiple opportunities for the teachers to reflect on their current ESL instructional practices, to observe and analyze videos in which the literacy textbook developers model *Sam and Pat* instruction[8], to engage in structured lesson planning with guidance and feedback from the trainers, and to self-assess what they are learning and evaluate the training activities to inform the pace and content of the workshop itself.

The *Sam and Pat* developers provided a refresher webinar training of about two hours early in winter 2009, before the start of the second term. The purpose of the webinar was to review the key principles of the training provided previously and provide more targeted training based upon teachers' experience during the first term. The agenda included sharing techniques teachers had found helpful as well as further training from the *Sam and Pat* developers on teaching phonics and engaging the more advanced students in the class in instruction.

The trainers also conducted one site visit to each of the teachers in the *Sam and Pat* group to observe instruction and provide feedback during the second or third week of the fall 2008 term. The trainers reviewed the classroom environment (e.g., the availability and use of specific instructional materials, the alignment of observed instruction with the *Sam and Pat* Lesson Plan template, and teacher practices), offered both oral and written feedback on the quality of instruction and suggestions for improvement, and provided other technical assistance to the *Sam and Pat* teachers as needed in response to e-mails or phone calls from the teachers.

Trainers also called each teacher in the *Sam and Pat* group biweekly during the first 2 months of each term. They asked the teachers how comfortable they were using *Sam and Pat,* if they required additional clarification on the activities or concepts, if they were having any difficulties with the materials, activities or lesson planning and if they would like additional technical assistance. Trainers referred teachers to relevant materials provided at the training, including the videos of *Sam and Pat* methods, to help refresh teachers on specific topics.

In addition, the trainers identified teachers who appeared to be having difficulty implementing *Sam and Pat* during their site visits and phone calls. The trainers provided 1-day individualized assistance in person to these teachers during the second week of the second term.

---

[8] *Sam and Pat* trainers gave a DVD to teachers that contained 23 instructional demonstration videos created by the developers for teachers' continued reference outside the training. Developers provided an additional video on phonics instruction after the refresher training.

Organization of the Report

This report describes the methodology and findings of the Impact of a Reading Intervention for Low-Literate Adult ESL Learners Study and is organized into five chapters. This chapter presented an overview of the study's conceptual background and research questions, summarized the design, and described the intervention. The remaining chapters are described below:

- Chapter 2: Study Design and Methods presents details on the study's recruitment and selection procedures and describes the random assignment methods used to assign students and teachers to groups. It also describes the study's assessment battery and other measures and data collection procedures.
- Chapter 3: Instruction and Attendance During the Study presents implementation data from classroom observations made by the research team. The research team observed each *Sam and Pat* and control class once per term using an observation instrument designed for this study. The observations allowed us to calculate measures of teachers' instructional fidelity to the *Sam and Pat* approach and also to describe other instructional activities of control and *Sam and Pat* teachers. This chapter also presents data on instructional and attendance service contrasts.
- Chapter 4: Impacts on Reading and English Language Skills presents findings from the impact analyses comparing students' post-test scores to estimate the impact of *Sam and Pat.*
- Chapter 5: Non-Experimental Analyses reports the correlational findings on the relationship between instruction, attendance, and outcomes.

There are six technical appendices to the report that provide greater detail on the assessments (Appendix A), study design (Appendix B), classroom observation methods (Appendix C), power analyses and impact estimation methods (Appendix D), and supplemental data analyses for chapters 3 and 4 (Appendices E and F).

CHAPTER 2:
STUDY DESIGN AND METHODS

This study employed individual random assignment of students and teachers to either the *Sam and Pat* group or the control group. The power analysis conducted for the study's design report established the site, class, and student numbers we used as targets in our recruiting effort (Condelli et al., 2009). Based on this analysis, we estimated that the study required about 1,800 students and 40 classes from 10 adult ESL sites to have sufficient statistical power to detect differences in reading and language outcomes between the *Sam and Pat* and control group.[9] In this chapter we describe our site selection and recruitment methods and the random assignment procedures used. We also present baseline data on students and teachers and the data collection summary and schedule.

Selection of Adult ESL Programs and Sites

Study staff identified adult ESL programs and screened them for study eligibility through a multi-step process. First, data from the U.S. Department of Education (ED, 2007) were used to identify states with the largest adult ESL enrollments. These states were California, New York, Texas, Florida, Illinois, Minnesota, Washington, New Jersey, and North Carolina. Evaluation staff contacted the state directors of adult education in each state, explained the study, and asked them to identify programs in their state that might be eligible for the study according to the following selection criteria:

- A managed enrollment policy or enrollment history in which a majority of learners enter during the first two weeks of the term;
- A history of high student retention rates (approximately 70% or more students remaining in class until the end of the term);
- Enrollments of adult ESL literacy learners large enough to support the study design (i.e., able to enroll about 90 low-literate ESL students per term in study classes);
- A sufficient number of adult ESL literacy instructors to support the evaluation's requirements (at least three instructors per site in the low-literate ESL student classes);[10] and
- Two or more classes for ESL literacy students that met at identical days and times and were located in the same or adjacent buildings.

[9] Appendix D provides power calculations using the study's actual sample sizes.

[10] We wanted at least two classes taught by different instructors to allow for a *Sam and Pat* and control group. A third instructor was needed as a backup in the event one of the *Sam and Pat* teachers became unable to complete the semester. Backup teachers received *Sam and Pat* training but were never needed at any site.

In addition, the site could not already be offering instruction based on *Sam and Pat.*

The state adult education directors identified 130 programs based on the specifications above and provided us with contact information. We then contacted the program directors to gauge their interest in participating in the evaluation, and to learn more about the types of students they served and the number of classes they provided. From these interviews, we found that 67 programs served low literacy students and had enough students and classes to participate in the study. We conducted follow-up screening via telephone conferences with program directors to verify information and to obtain additional information to ascertain the program's study eligibility. We sought explanation and clarification on enrollment policy; students' prior education and literacy levels; student attrition rates; class schedules, sizes, and locations; any barriers or concerns site staff had about the study; and the interest of staff in participating in the study.

Of the 67 programs contacted, 32 programs appeared to meet the selection criteria and had program directors who expressed an interest in participating in the study. The program directors of the 32 programs were contacted a second time to confirm their interest in participating and to verify information regarding their program's eligibility for the study. Evaluation staff also provided the program directors with more information about the study, including details about random assignment. Seven programs declined to participate in the study. Among the remaining 25 programs, 12 were interested in participating and appeared to meet the study criteria, and 13 expressed interest but did not meet the study criteria upon further discussion. From a close screening of the remaining 12 programs' enrollment policies, student attrition rates, teacher training and qualifications, and class schedules and location, we selected 8 programs that offered 13 instructional sites (i.e., multiple sites within some programs) to visit for further consideration. During the visits evaluation staff again verified that the site conformed to study criteria and that teachers and site staff were willing to participate.

After site visits, three programs were either no longer interested in participating in the study or had insufficient numbers of adult ESL literacy students. Within the remaining five programs, there were 10 sites eligible for the study. These sites were recruited to participate. The sites were located in California, Texas, Florida,

and Illinois.[11] Within these sites, we identified all pairs (n = 17) of adult ESL literacy classes that met at the same time and location and included all pairs in the study. Table 2.1 shows the number of class pairs and students at each site. The classes were scheduled to meet for 5 to 17.5 hours per week, for a period of 8 to 18 weeks. The total number of hours that class pairs were scheduled to meet each term was 74 to 245 hours in the fall term and 65 to 210 hours in the spring term (not shown in tables).

Table 2.1: Number of Classes and Students in the Study, Overall and by Site

Site	Number of Classes	Number of Students Randomly Assigned
Site A	8	222
Site B	8	54
Site C	6	109
Site D	8	86
Site E	4	72
Site F	8	349
Site G	4	61
Site H	4	98
Site I	12	205
Site J	4	88
Total	**66**	**1,344**

Source: Project database used for random assignment.

Recruitment and Random Assignment of Teachers and Students

Teachers

During site visits, staff from the program site identified the classes and teachers who would participate in the study. Only teachers with at least one year of experience teaching adult ESL literacy students were eligible. Study staff explained to the teachers that the study's purpose was to evaluate *Sam and Pat,* a literacy intervention for low literate ESL learners that had a focus on basic

[11] The nature of the study's random assignment requirements (e.g., the need to have at least 2 literacy level classes at the same time and place) and the targeted recruitment from the states with the largest ESL enrollments may have implications for the generalizability of the study's results. It is possible that the study sites are somehow different from sites that did not meet the requirements; however, we cannot address that possibility with the data that are available. There are no nationally representative data that we can compare our site characteristics against, and no descriptive data were collected at the site level during the study.

reading skills and phonics. Study staff also explained to teachers that they would be assigned randomly to either teach with *Sam and Pat* or to teach as usual if they were assigned to the control group. Students would also be assigned randomly to attend one of the two classes and would be assessed shortly after starting class and then at the end of the term. Teachers were given a brief, simple explanation of random assignment and why it was being used in the study. They were also told that study staff would observe them teaching at least once per term. Staff explained that teachers in the intervention group would be required to attend the training on *Sam and Pat* and that control group teachers would teach as usual and receive no training.

Shortly before the training, study staff randomly assigned teachers to group and informed them of their assignment. Control group teachers received no training and taught their classes as they usually did during the study period. Teachers using the *Sam and Pat* book and materials were instructed not to share them with control group teachers during the study's data collection period.

Random assignment of teachers to group occurred during the summer of 2008, with teachers maintaining their assignment across both terms.

Students

Prior to the beginning of the fall 2008 and spring 2009 terms, students registered for classes as they normally did. If the site staff determined through their standard procedures that a student belonged in a literacy level class, the student was identified as eligible for the study classes. Intake staff then referred the students to a site or study staff member to recruit them into the study. Staff explained the following in students' native languages[12]:

- The school was trying a new way to teach literacy-level classes to see if it is better at helping students learn to read and speak English than the school's standard instruction;
- The class, as well as the other class that met at the same time, were part of the study;
- Students would have the opportunity to participate in the study and those who chose to participate would be assigned to one of two classes;
- Students would be assigned to class using a chance process like the lottery and they would have a 50-50 chance of being in either class;

[12] Before the start of data collection, study staff identifed languages students were likely to speak based on program data on students who attended the previous semester. Study staff produced audio recordings of the information about the study in languages where it was expected that no staff speaking those languages would be available on site at the time of intake.

- ❖ Students would participate in a pre- and post-test; and
- ❖ Students would receive $40 at the end of the term for their participation in data collection.

Students were also assured of confidentiality of all data.

Staff then answered any questions and asked whether students were willing to participate in the study. Students who agreed received and signed an informed consent form that was written in their native language. Staff read the form aloud for students who could not read. Of 1,430 students referred, 86 (6 percent) declined to participate, leaving a total of 1,344 students in the study (see Table 2.1). During the first week of class, study staff were on site and randomly assigned participating students into a *Sam and Pat* or control class using the study's Web-based data system, which had a built-in randomization function.[13] During the second week of instruction, site intake staff performed this function. All random assignment occurred within the first two weeks of class. Any student entering the class after that time was allowed to attend class but was not randomly assigned and did not participate in any data collection activities.

Students who chose not to participate in the study were assigned to a class by the site according to the site's standard procedures. Those students were not included in data collection activities, although they may have attended study classes.

The sample selection process described in this chapter is summarized in Figure 2.1. Although we undertook a purposeful selection of programs and sites, the teachers and students at those sites, once selected to be in the study, were then randomly assigned to a treatment or control group. This ensures that the resulting impact estimates are internally valid. The description of the program and site selection process is intended to help readers understand the population to which the findings generalize.

[13] Some students in the study expressed a desire to be assigned to a class with a friend or family member. These students were randomly assigned as a "pod." There were 61 pods in the sample, and 139 students participated as a member of a pod (67 students in the *Sam and Pat* group and 72 students in the control group).

Figure 2.1: Sample Selection Flowchart

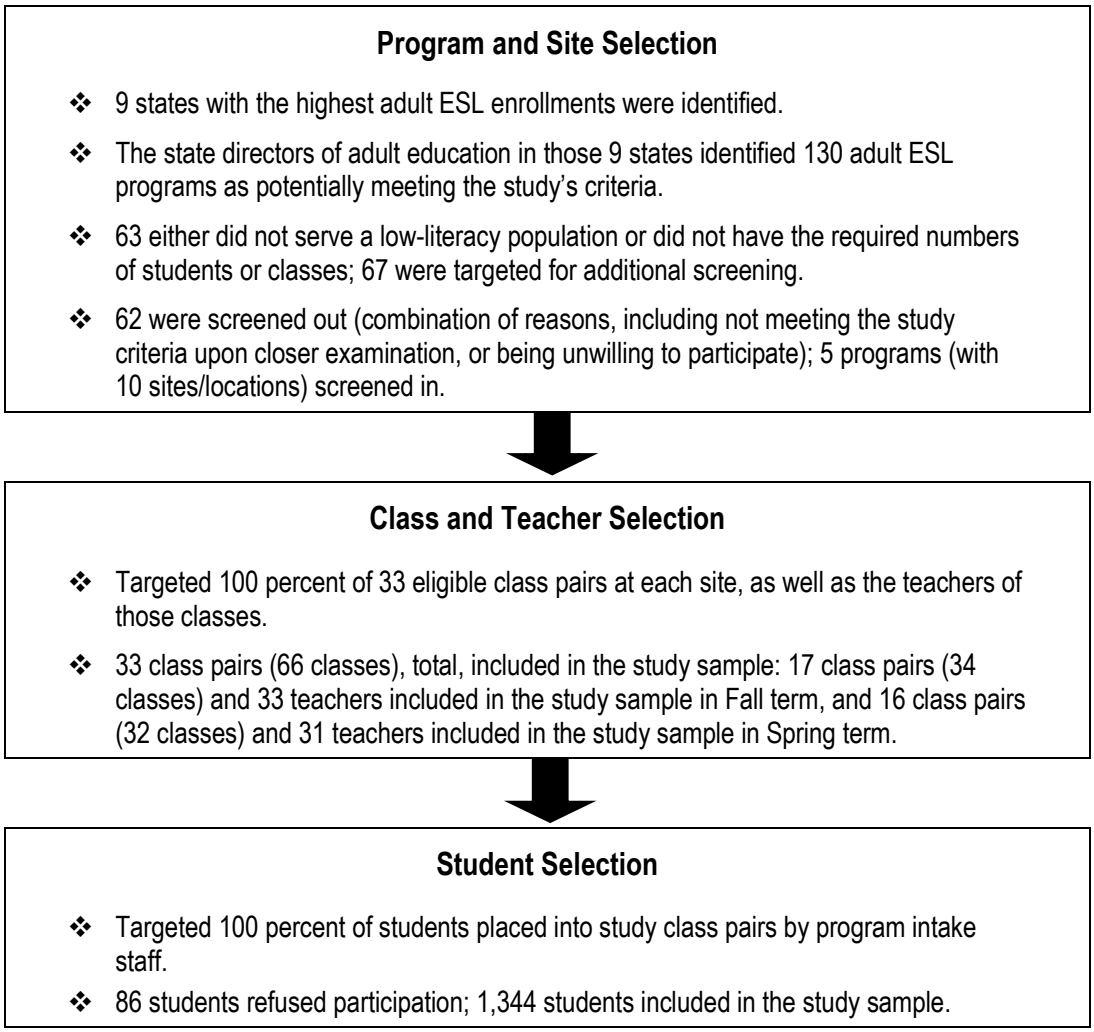


Figure 2.2 summarizes the recruitment, random assignment, and flow of data collection for the study.

Data Collection

Data were collected from teachers and students who participated each term. The data collections included the following:

- two brief teacher data forms;
- a student intake form filled out by site staff;
- a student pre- and post-test assessment battery;
- daily student attendance forms; and
- a classroom observation instrument.

These collections are described in more detail in the following sections. The data collection schedule is summarized in Table 2.2, and response rates overall and by group are provided in Table 2.3.

Teacher Data Form (2008)

The 2008 Teacher Data Form was used to collect background information about the study teachers, including teacher credentials, educational background, years of overall and ESL teaching experience, and demographics. The form was administered via a combination of in-person and mail survey during the summer and fall of 2008. All teachers responded.

Teacher Data Form (2009)

There were two versions of the 2009 Teacher Data Form—one specific to *Sam and Pat* teachers and one for the control group teachers. The 2009 data form served as a follow-up survey to collect information from all teachers on the instructional materials used throughout the year and to ask *Sam and Pat* teachers a variety of questions about their use of *Sam and Pat* (e.g., time spent preparing for lessons, final lesson number covered each term). The form was administered via a mail survey at the conclusion of the spring 2009 term, with a response rate of 87 percent.

Figure 2.2: Study Procedural Flow Chart

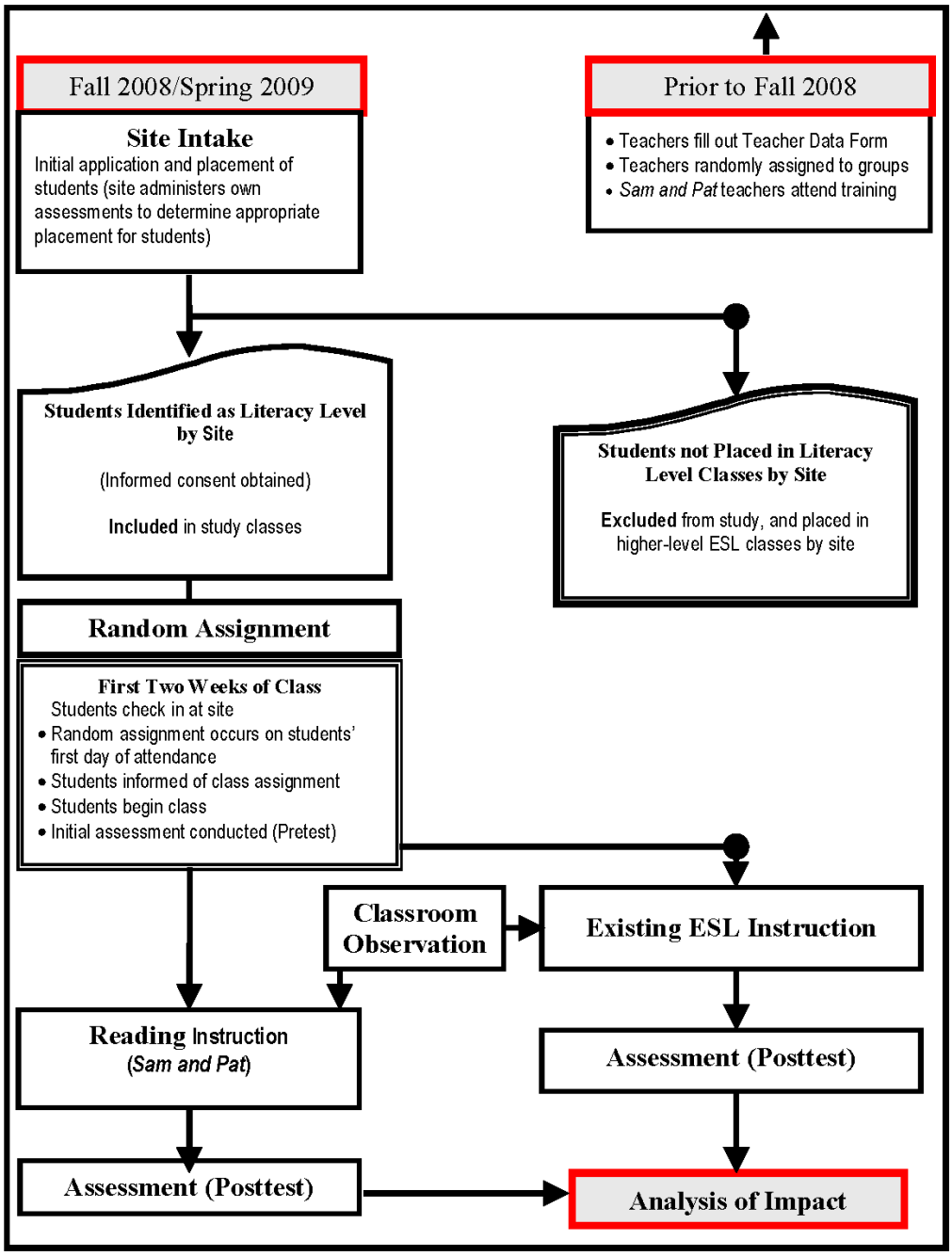


Table 2.2: Data Collection Schedule

Data Collection	Respondent	Summer 2008	Fall 2008	Spring 2009	Type of Data
Teacher Data Form (2008)	Teachers	X	X		Teacher background information
Teacher Data Form (2009)	Teachers			X	Descriptive information about instructional materials used and *Sam and Pat* implementation
Student Intake Form	Site Staff on Behalf of Students		X	X	Student background information
Reading and English Language Pre-Tests	Students		X	X	Pre-test data
Reading and English Language Post-Tests	Students		X	X	Outcomes data
Daily Student Attendance Sheets	Teachers		X	X	Dosage/exposure to instruction
Classroom Observations	Evaluation Staff		X	X	Descriptive information about instruction in both groups

Table 2.3: Percentage of Teachers, Classes, and Students Participating in Data Collections, by Group

	Overall	*Sam and Pat*	Control	Difference	P-Value
Teacher Collections					
Teacher data form (2008)	100.0	100.0	100.0	0.0	†
Sample Size (Teachers)	**33**	**16**	**17**		
Teacher data form (2009)	87.1	86.7	87.5	0.8	0.94
Sample Size (Teachers)	**31**	**15**	**16**		
Class Collections					
Attendance records	100.0	100.0	100.0	0.0	†
Classroom observations	97.0	93.8	100.0	6.2	0.80
Sample Size (Classes)	**66**	**33**	**33**		
Student Collections					
Student intake form	100.0	100.0	100.0	0.0	†
Pre-test battery	94.3	94.7	94.0	0.6	0.62
Post-test battery	84.6	86.1	83.1	2.9	0.14
Sample Size (Students)	**1,344**	**674**	**670**		

† Not applicable.
Note: A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.

Student Intake Form

The Student Intake Form was used by site staff during registration to collect basic background information about students (names, contact information, years of prior education, etc.), and it represents the kind of information typically collected by programs. For the purposes of the study, site staff entered this information into the study's Web-based MIS. Data were obtained for every student.

Student Assessments

Each student was assessed with a battery of standardized reading and English language (i.e., speaking/listening) tests. The study assessment battery included pre- and post-tests that measure the reading and English language skills that were the primary outcomes for the study. The following assessments were administered:[14]

[14] See Appendix A for a discussion of the assessment selection and administration procedures.

Reading Cluster

- ❖ Woodcock Johnson III Tests of Achievement (WJ: Woodcock, McGrew, & Mather, 2001)

    - Letter-Word Identification (WJID)—measures students' word identification skills as indexed by pronunciation of familiar printed words.
    - Word Attack (WJWA)—measures skills in applying phonic and structural analysis skills as indexed by pronunciation of unfamiliar words.
    - Passage Comprehension (WJPC)—students read a short phrase or passage, then choose or supply missing words that make sense in the context.

- ❖ ETS SARA—Decoding (SARA-Dec) and Letter Naming (SARA-LN). The Educational Testing Service (ETS) developed the Study Aid and Reading Assistant (SARA; Sabatini & Bruce, in press) assessment battery for research purposes to measure English reading skills. The Decoding subtest from the battery measures skills in applying phonic and structural analysis skills as indexed by pronunciation of unfamiliar words. The Letter Naming subtest measures knowledge of the alphabet by asking students to name letters.

English Language Cluster

- ❖ Woodcock Johnson Picture Vocabulary (WJPV: Woodcock, McGrew, & Mather, 2001). Students are shown images and asked to identify the relevant words. This assessment measures oral expressive vocabulary.
- ❖ Oral and Written Language Scales (OWLS: Carrow-Woolfolk, 1996)—Listening subtest. The examiner reads aloud a verbal stimulus and the student points to one of four pictures. The OWLS is designed to measure the construct of listening comprehension (understanding continuous oral text, from simple items, such as a request to identify the picture representing a particular characteristic, to more complex items, such as a request to interpret something a character in the picture has said, "What did that mean?").
- ❖ Receptive One-Word Picture Vocabulary Test (ROWPVT: Brownell, 2000). The examiner says a word and the student must point to one of four pictures that represents the object named. The ROWPVT is designed to measure the construct of receptive (hearing) vocabulary.
- ❖ We administered the SARA Letter Naming subtest only at the beginning of the term (pre-test) to capture variability in basic knowledge of the alphabet, and the SARA Decoding subtest only at the end of the term

(post-test) to provide more discrimination within the expected range of decoding ability after one term. The WJ Picture Vocabulary test was also only administered at post-test. All other assessments were administered at the beginning and end of the term. Response rates were 94 percent on the pre-test battery and 85 percent on the post-test battery. Table 2.4 summarizes the pre- and post-test administration times. For the post-test battery, test reliabilities ranged from 0.81 to 0.96 (see Appendix A for details).

Daily Student Attendance Sheets

Daily student attendance sheets were filled out by teachers for each class period in order to provide the study with a measure of instructional "dosage." It was also used to track class entry/exit and any "crossover" of students between *Sam and Pat* and control classes. A complete set of attendance sheets was received from 6 out of 10 sites during the fall term and all 10 sites during the spring term. For the fall term, attendance records were missing for 1 week of attendance in 5 classes spread across 4 sites.

Table 2.4: Assessment Administration Schedule, by Test

Assessment	Administered at Pre-Test (Beginning of Term)	Administered at Post-Test (End of Term)
WJ Letter-Word Identification	X	X
WJ Word Attack	X	X
SARA Letter Naming	X	
SARA Decoding		X
WJ Passage Comprehension	X	X
OWLS	X	X
ROWPVT	X	X
WJ Picture Vocabulary		X

Classroom Observations

Classroom observations were conducted by trained study staff once per term in each study class using a structured observation guide. The observation guide was designed to capture the content of the *Sam and Pat* and control teachers' instruction. It was also designed to record the instructional materials used in both groups, allowing us to document the use of *Sam and Pat* in study classrooms.

Staff received 1.5 days of training and two practice observations with feedback before going into the field. They also received a 2-hour retraining and feedback as needed after the first observation. Each term, approximately 10 percent of the observations were conducted by two staff so that inter-rater reliability could be

determined. Observer agreement ranged from 0.86 to 0.95 in fall and 0.90 to 0.98 in spring. More information and a copy of the observation instrument is provided in Appendix C.

Integrity of Random Assignment

Baseline Equivalency of Sam and Pat and Control Groups

To verify that random assignment succeeded in creating two equivalent study groups, we used teacher and student data collected at the beginning of the term to compare characteristics across groups. For discrete outcomes we calculated χ^2 statistics; for continuous outcomes we calculated t-statistics. As shown in Tables 2.5 and 2.6, there were no statistically significant differences between the two groups on the characteristics measured. This was true for both teachers and students.

In addition, we compared students' assessment scores as of the beginning of the term to determine whether there were any pre-existing differences between groups. As Table 2.7 shows, there were also no significant test score differences between groups at the beginning of the term.

Student Movement Between Groups

A potential threat to the validity of study findings is movement between groups after random assignment (e.g., students assigned to the control group subsequently attended a *Sam and Pat* class). To monitor the extent to which random assignment to group was preserved during the study, we kept track of the classes students attended throughout the term. This was accomplished through review of attendance records and by maintaining communication with site staff. If we learned that a student had attended a class that was not his or her assigned class, we discussed the case with the site staff to determine whether the student could be encouraged to return to the assigned class. In all such cases, movement was between paired study classes (*Sam and Pat* and control). When movement occurs between study groups, it is referred to as *crossover.* We have defined crossover students as those students who, at any point during the term, attended a study class to which they were not randomly assigned. Table 2.8 shows the total number of crossover students, and the percent of the sample categorized as crossover by type. The overall number of students who attended a class to which they were not randomly assigned was 13 (1 percent of the total student sample). Nine of these students (0.7 percent), attended an unassigned study class throughout the entire term, and 4 students (0.3 percent) attended both their assigned and an unassigned study class at some point in the term. All crossovers were treated as members of their randomly assigned groups for the purposes of impact analyses (i.e., we followed an "intent to treat" (ITT) analysis approach).

Table 2.5: Teacher Background Characteristics, by Group (Percentages)

	All Teachers	*Sam and Pat*	Control	Difference	P-Value
Gender					0.353
Male	60.6	68.8	52.9	–15.8	
Female	39.4	31.3	47.1	15.8	
Race/Ethnicity					0.978
White	39.4	37.5	41.2	3.7	
Black or African American	30.3	31.3	29.4	–2.0	
Hispanic or Latino	24.2	25.0	23.5	–1.5	
Teacher Credentials					0.923
ESL or TESL	24.2	25.0	23.5	–1.5	
State certification	30.3	25.0	35.3	10.3	
State certification with additional credential*	27.3	31.3	23.5	–7.7	
No certification, or accreditation other than state*	18.7	18.7	18.7	0	
Highest Education Level Completed					0.261
Bachelor's	42.4	56.3	29.4	-26.8	
Master's	48.5	43.8	52.9	9.2	
Sample Size (Teachers)	**33**	**16**	**17**		
Years of Experience Teaching Adult ESL					0.693
1–3 years	26.7	26.7	26.7	0.0	
4–7 years	40.0	46.7	33.3	–13.3	
8 years or more	33.3	26.7	40.0	13.3	
Sample Size (Teachers)	**30**	**15**	**15**		

* Additional credential or accreditation includes, for example, an ESL or adult education certification that is awarded upon completion of additional units of study in a topic area.

Notes: Calculations are based on the full sample of teachers. A two-tailed χ^2 or t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.

Source: Adult ESL Literacy Impact Study teacher data form (2008).

Table 2.6: Student Background Characteristics, by Group (Percentages)

	All Students	*Sam and Pat*	Control	Difference	P-Value
Gender					0.417
Male	41.0	39.9	42.1	2.2	
Female	59.0	60.1	57.9	–2.2	
Race/Ethnicity					0.612
Asian/Pacific Islander, Native-Hawaiian and Other	13.5	12.6	14.3	1.7	
Black or African American	12.4	13.4	11.5	–1.9	
White	24.9	24.9	24.8	–0.1	
Hispanic or Latino	46.4	45.8	47.0	1.2	
Missing/Unknown	2.8	3.3	2.4	–0.9	
Total Years of Schooling					0.110
3 years and under	25.7	23.3	28.2	4.9	
4–8 years	30.0	31.5	28.5	–2.9	
9 years or more	44.3	45.3	43.3	–2.0	
Number of Years in the U.S.					0.155
3 years and under	63.1	61.6	64.6	3.1	
4–8 years	3.4	4.3	2.5	–1.8	
9 years or more	33.5	34.1	32.8	–1.3	
Sample Size (Students)	**1,344**	**674**	**670**		
First Language					0.281
Armenian	23.2	23.8	22.7	–1.1	
Chinese	8.6	7.7	9.5	1.8	
Haitian Creole	13.5	15.0	12.1	–2.9	
Spanish	46.4	45.5	47.3	1.8	
Vietnamese	2.8	3.3	2.2	–1.0	
Other	5.4	4.7	6.1	1.4	
Sample Size (Students)	**1,343**	**673**	**670**		

Table continued, next page.

Table 2.6: Student Background Characteristics, by Group (Percentages) (Continued)

	All Students	*Sam and Pat*	Control	Difference	P-Value
Age*					0.723
18–25 years	19.28	20.00	18.55	–1.45	
26–30 years	11.90	11.43	12.37	0.94	
31–40 years	23.72	23.46	23.98	0.52	
41–50 years	18.90	17.74	20.06	2.32	
51–60 years	14.53	15.64	13.42	–2.22	
61–70 years	8.73	9.17	8.30	–0.88	
70 years and above	2.94	2.56	3.32	0.76	
Sample Size (Students)	**1,328**	**665**	**663**		

* Mean age for overall sample = 40.37 years; *Sam and Pat* Group = 40.46 years; Control Group = 40.29. Range = 18 to 84 years.

Notes: Percentages are unadjusted, and based on all students for whom intake data were available. A two-tailed χ^2 or t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.

Source: Adult ESL Literacy Impact Study student intake forms collected at the beginning of each term (fall 2008 and spring 2009).

Table 2.7: Mean Student Assessment Scores at Beginning of Term, by Group

Outcome	All Students	*Sam and Pat*	Control	Difference	P-Value
Reading Assessments					
SARA Letter Naming	21.73	21.63	21.84	0.20	0.550
Woodcock Johnson Letter/Word Identification Scale	403.24	405.09	401.38	–3.71	0.546
Woodcock Johnson Word Attack Scale	431.28	433.97	428.57	–5.40	0.382
Woodcock Johnson Passage Comprehension Scale	402.40	403.23	401.56	–1.67	0.760
English Language Assessments					
OWLS	13.36	13.34	13.37	0.02	0.965
ROWPVT	21.51	21.38	21.64	0.25	0.769
Sample Size (Students)	**1,344**	**674**	**670**		

Notes: Scores are unadjusted, and based on the full sample of students. Missing values were set to 0 and flagged with a missing value dummy variable code. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.

Source: Adult ESL Literacy Impact Study assessments administered during the first two weeks of each term (fall 2008 and spring 2009).

Table 2.8: Number and Percent of Students Who Attended Unassigned Study Classes, by Crossover Type

Type of Crossover	Number of Students	Percent of Students (N = 1,344)
Students who attended both *Sam and Pat* and control classes at some point in the term	4	0.30
Students who attended an unassigned study class throughout the entire term	9	0.67
Total number of students who attended an unassigned study class	13	0.97

Source: Adult ESL Literacy Impact Study Attendance Database.

CHAPTER 3: INSTRUCTION AND ATTENDANCE DURING THE STUDY

To describe instruction across study classes during the study and document the implementation of *Sam and Pat* materials and related instructional practices, members of the study team conducted structured observations of study classes once per cohort, at approximately 6 weeks into the beginning of each term. The observation instrument used for this purpose was designed in collaboration with the intervention developers, and was designed to capture instruction used in adult ESL classrooms as well as instruction in the key reading content or skill areas of *Sam and Pat* instruction. Trained observers coded instruction from the following categories:

Reading Development	
• Phonics • Learning vocabulary to reinforce reading instruction • Reading comprehension	• Writing and spelling for phonics reinforcement • Fluency and accuracy in reading
Writing Unrelated to Reading Activities	
• Subskills and practice • Free writing	• Guided composition
English Language Acquisition	
• Oral communication skills—listening • Grammar (understanding how English works) • Socio-cultural knowledge	• Oral communication skills—speaking • Vocabulary and idioms (not related to reading activities)
Functional Reading, Writing, and Math	
• Text based • Graphic literacy	• Alphabet based • Working with numbers and math
Making Links Between What Is Learned in Classroom and the Outside World	
Use of Students' Native Language	
Other (Uncodable) Instruction and Breaks	

Specific instructional practices were documented within each of these instructional areas, including practices that *Sam and Pat* teachers were either explicitly trained to employ or were expected to use in the course of covering the content of *Sam and Pat*.

For each 5-minute interval of the class observed, observers circled any practice used during instruction (see Appendix C for a copy of the observation guide). Therefore, multiple practices could be coded in one or more instructional areas during each interval observed.

The observers also documented any materials used during instruction, using the categories below:

***Sam and Pat* Materials**	
• *Sam and Pat* workbook or worksheets (including overheads of these pages or blown up copies of text book pictures)	• *Sam and Pat* key word (sound/symbol) cards • Wilson letter cards • *Sam and Pat* phonetic word grids
Other Materials	
• Other commercial text or worksheets (including overheads of these pages or blown up copies of text book pictures)	• Blackboard/whiteboard • Other (specify)

Like instructional practices, materials were coded during each interval and could be coded under multiple instructional areas, depending on the practices observed.

In this chapter, we present a description of the *Sam and Pat* teachers' instruction, including the extent to which the intervention materials and practices were used in *Sam and Pat* classes and the proportion of *Sam and Pat* classrooms that met the study's implementation fidelity criteria. We also present findings from the spring *Sam and Pat* teacher survey to give the reader context for the implementation results. Finally, we provide a comparison of both the instruction and student attendance observed in *Sam and Pat* and control classrooms to demonstrate the "service contrast" between the study's two groups.

Description of Instruction in *Sam and Pat* Classrooms

General Class Duration

The average *Sam and Pat* classroom observation lasted 31.6 intervals, or approximately 158 minutes (not shown in tables).[15] Observations ranged from 21 to 43 intervals (105 to 215 minutes). To determine the extent to which this class time focused on *Sam and Pat* instruction, we report on the proportion of observed instructional intervals that included the use of *Sam and Pat* materials and practices in sections below.

[15] One interval is equivalent to 5 minutes, although it should be noted that instruction in a content area did not have to occur during all 5 minutes in order to be coded as occurring.

Proportion of Instructional Intervals Incorporating Sam and Pat Materials

Overall, observers documented the use of *Sam and Pat* materials during 44 percent of intervals observed in *Sam and Pat* classes (Table 3.1). Given the average observation length of 31.6 intervals, the materials were therefore used for an average of 13.9 intervals, or about 70 minutes per observation (not shown in tables). Given that the average number of days study classes met per week was 3.5 (not shown in tables), this indicates that *Sam and Pat* teachers were using *Sam and Pat* materials for an average of approximately 70 minutes X 3.5 days = 245 minutes per week. This is less than the 5 hours (300 minutes) that the *Sam and Pat* teachers were asked to spend on *Sam and Pat*; however, given that the developers trained the *Sam and Pat* teachers to provide instruction using materials beyond the *Sam and Pat* workbook (e.g., chalkboard, realia and index cards), it is possible that teachers were delivering instruction based on *Sam and Pat* for additional time that is not captured in Table 3.1. The following section provides information on instructional practices used by *Sam and Pat* teachers, including instruction utilizing the *Sam and Pat* workbook as well as other materials.

Table 3.1: Percent of Instructional Intervals During Which *Sam and Pat* Materials Were Used in *Sam and Pat* Classrooms

Instructional Area	Cohort 1	Cohort 2	Overall
Total, Any Content Area	**44.9**	**43.0**	**44.0**
Reading Development	**42.1**	**31.8**	**37.0**
Pre-reading (print directionality, etc.)	3.4	2.3	2.9
Phonics	13.8	10.0	11.9
Writing and Spelling for Phonics Reinforcement	7.2	7.7	7.4
Learning Vocabulary to Reinforce Reading Instruction	5.0	2.3	3.7
Fluency and Accuracy in Reading	11.6	11.4	11.5
Reading Comprehension	10.6	8.1	9.4
English Language Acquisition	**4.0**	**5.0**	**4.5**
Oral Communication Skills—Listening	0.6	0.2	0.4
Oral Communication Skills—Speaking	1.2	3.1	2.1
Grammar	1.8	1.7	1.7
English Vocabulary & Idioms	0.8	0	0.4
Sociocultural Knowledge	0	0	0
Functional Reading, Writing, and Math	**0**	**0**	**0**
Sample Sizes: 16 Cohort 1 observations (499 intervals); 15 Cohort 2 observations (481 intervals); 31 observations, total (980 intervals). One missing observation per cohort.			

Notes: Details may not sum to totals. Materials may be coded under multiple instructional areas during any one interval. Percents are unadjusted, and based on all *Sam and Pat* classes for which data were available.
Source: Adult ESL Literacy Impact Study classroom observation protocol.

Use of Instructional Practices in Support of Sam and Pat

Table 3.2 provides a list of the practices that observers could expect to see used in support of *Sam and Pat* implementation, as identified in collaboration with the *Sam and Pat* developers during the planning stage of the study. Overall, these activities were documented during 53.5 percent of intervals observed in *Sam and Pat* classes, which is equivalent to 16.9 intervals on average, or approximately 85 minutes. This implies that during some intervals, these practices were being used independently of *Sam and Pat* materials. For example, the activities could take place by using the board, index cards, other types of materials (e.g., realia), or no materials (e.g., air writing).

Table 3.2: Percent of Instructional Intervals During Which *Sam and Pat* Teachers Engaged in Practices in Support of *Sam and Pat*

Instructional Area and Practice	Cohort 1	Cohort 2	Overall
Total, All Content Areas	**59.3**	**47.7**	**53.5**
Pre-literacy	**5.0**	**4.0**	**4.5**
Recognizing individual letters and working with the names of letters	4.4	2.7	3.6
Working with upper vs. lower case letters of the alphabet	1.2	1.2	1.2
Phonics	**19.8**	**16.4**	**18.2**
Explains, describes, or demonstrates sound-symbol pattern or decoding rule	16.0	12.1	14.1
Uses multi-sensory approaches to emphasize phonemic correspondences	6.2	4.2	5.2
Practices sound-symbol correspondence either independently or guided by teacher	18.2	15.4	16.8
Writing and Spelling for Phonics Reinforcement	**9.6**	**14.1**	**11.8**
Matching/labeling pictures with phonetically regular words	2.4	3.5	3.0
Writing letter(s) that represent a phoneme	1.8	1.0	1.4
Circling the phonetically regular word	2.0	1.0	1.5
Taking dictation of phonetically regular words	1.8	2.7	2.2
Oral spelling of phonetically regular words	3.2	4.0	3.6
Copying/writing phonetically regular words	1.4	7.1	4.2
Learning Vocabulary to Reinforce Reading Instruction	**18.2**	**15.0**	**16.6**
Introduces a small number (8 or fewer) of vocabulary words or reviews old vocabulary words related to the class readings	12.4	7.5	10.0
Writes words on board, reads aloud, students repeat	5.8	4.4	5.1
Dictates vocabulary words to students	1.2	3.3	2.2

Table continued, next page.

Table 3.2: Percent of Instructional Intervals During Which *Sam and Pat* Teachers Engaged in Practices in Support of *Sam and Pat* (Continued)

Instructional Area and Practice	Cohort 1	Cohort 2	Overall
Air-writes or traces words with their finger while spelling out loud	4.6	1.9	3.3
Matches vocabulary words (orally or physically) to pictures or realia	3.0	1.2	2.1
Labels pictures (in writing) with vocabulary words	0.4	0.2	0.3
Sorts cards with vocabulary words or pictures into topics	0	0.6	0.3
Writes vocabulary words on flash cards or in notebooks	4.2	6.9	5.5
Does a cloze exercise to fill in new vocabulary	1.4	1.5	1.4
Fluency and Accuracy in Reading	**12.6**	**13.7**	**13.2**
Reads text aloud to students before having them read	5.6	6.4	6.0
Reads text aloud, listen to others and read along, or take turns reading	10.2	12.3	11.2
Practices reading parts of sentences	0	0	0
Follows along during reading by tracing under the words with an eraser or finger	2.4	1.9	2.1
Reading Comprehension	**12.4**	**11.6**	**12.0**
Previews the text and/or pictures before reading	4.8	3.3	4.1
Interacts with students to elicit storyline and/or understanding of new words in readings before reading	4.4	1.9	3.2
Activates or builds students' background knowledge related to the reading	3.8	2.7	3.3
Asks students direct recall questions	2.2	3.1	2.7
Asks students inferential questions after reading	1.0	1.2	1.1
Previews the text and/or pictures before reading guided by or independent of teacher	4.6	2.3	3.5
Makes predictions about aspects of the story, predicts the ending of sentences or readings, or asks questions relevant to the text during reading	0.8	3.1	1.9
Matches sentences from the reading to pictures	0.4	1.0	0.7
Acts out a story	0.8	0	0.4
Sample Sizes: 16 Cohort 1 observations (499 intervals); 15 Cohort 2 observations (481 intervals); 31 observations, total (980 intervals). One missing observation per cohort.			

Notes: Details may not sum to totals. Practices may be coded under multiple instructional areas during any one interval. Percents are unadjusted, and based on all *Sam and Pat* classes for which data were available.
Source: Adult ESL Literacy Impact Study classroom observation protocol.

Two-thirds of Sam and Pat Classes Observed Met Fidelity Criteria

Study staff worked with the developers during the design of the classroom observation instrument to operationally define fidelity to the intervention in the *Sam and Pat* classrooms. The following criteria were established:

- *Sam and Pat* materials must be used for a minimum of 1 hour of instruction per class day—the equivalent of approximately 12 observation intervals;
- Each class day must include at least 1 hour (12 intervals) of instruction in reading development; and
- Each class day, instruction should occur in at least three of the reading development instructional areas (e.g., phonics, fluency,

Table 3.3 shows that about two-thirds (65 percent) of the *Sam and Pat* classes observed met our three fidelity criteria: (1) *Sam and Pat* materials were used during 12 or more intervals; (2) instruction in reading development took place during at least 12 intervals; and (3) instruction occurred in at least three of the reading development instructional areas during the observation. The number of classes meeting the fidelity criteria (10) is a constant across cohorts (not shown in tables). These results indicate that while not all teachers implemented *Sam and Pat* to the full extent intended by the intervention's developers, all fidelity criteria were met in approximately two-thirds of the *Sam and Pat* classes observed.

Table 3.3: Percent of Observations During Which All Fidelity Criteria Were Met in *Sam and Pat* Classes

	Cohort 1	Cohort 2	Overall
All Fidelity Criteria Met	62.5	66.7	64.5
Sample Sizes: 16 Cohort 1 observations; 15 Cohort 2 observations; 31 observations, total. One missing observation per cohort.			

Note: Percents are unadjusted, and based on all *Sam and Pat* classes for which data were available.
Source: Adult ESL Literacy Impact Study classroom observation protocol.

Context of Implementation in *Sam and Pat* Classrooms

To better understand the context of *Sam and Pat* implementation, including factors that may have facilitated or limited implementation, we analyzed data from the spring *Sam and Pat* teacher survey. The types of information collected from teachers included the following:

- The number of times teachers reported accessing various supports for *Sam and Pat* implementation;
- The number of minutes teachers reported preparing for their *Sam and Pat* classes each week, on average;
- The final *Sam and Pat* lesson number each teacher reported covering (out of 22); and
- Teacher reports on the frequency of using a variety of materials other than *Sam and Pat* during instruction.

Accessing Implementation Supports

The *Sam and Pat* developers invited teachers to access a variety of supports during the study that were designed to facilitate implementation. Teachers were asked at the end of the spring term how many times they had accessed those supports throughout the year: phone calls, video (i.e., watching developers model instruction), or e-mail support—all provided by the developers. On average, teachers reported speaking to developers by phone 3.4 times during the year, watching modeling videos 5.8 times, and accessing support by e-mail 2.5 times during the year (Table 3.4). Most teachers (85 percent) reported accessing each support for implementation. Exceptions included two teachers who did not access phone support, one teacher who did not use the instructional modeling videos, and one teacher who did not exchange e-mails with the developers (not shown in tables).

Table 3.4: Number of Times Each Support for *Sam and Pat* Was Accessed, as Reported by *Sam and Pat* Teachers

Support	Mean	Std	Min	Max
Phone call support from *Sam and Pat* developers	3.4	3.8	0	15
Video support from *Sam and Pat* developers (e.g., clips of instructional modeling via CD-ROM, DVD, or online)	5.8	7.2	0	25
E-mail support from *Sam and Pat* developers	2.5	2.2	0	8
Sample Size: 13 teachers.				

Note: Means are unadjusted, and based on all *Sam and Pat* teachers for whom data were available.
Source: Adult ESL Literacy Impact Study spring 2009 *Sam and Pat* teacher data form.

Preparation Time for Sam and Pat Lessons

The *Sam and Pat* developers estimated that teachers would need to spend approximately 2 hours preparing for their *Sam and Pat* classes each week. An average reported preparation time that exceeds developers' estimates could indicate a greater than expected burden on teachers, while less preparation time reported by an individual teacher could indicate that a teacher was not adequately prepared to implement *Sam and Pat* instruction. To determine how much time teachers spent preparing for their *Sam and Pat* classes, we asked teachers to report the average number of minutes spent preparing each week during both terms. As shown in Table 3.5, teachers reported spending 133 minutes (or 2.2 hours) preparing for their classes each week, on average, which is consistent with the developers' expectations. There were six teachers, however, who reported spending less than 2 hours preparing for their classes each week during both cohorts, and three of these teachers spent 30 minutes or less per week preparing (not shown in tables).

Table 3.5: Average Number of Minutes Per Week Spent Preparing to Teach Study Class, as Reported by *Sam and Pat* Teachers

	Mean	Std	Min	Max
Overall	**133.1**	**116.1**	**10**	**420**
For Cohort 1 Class	138.8	112.4	14	360
For Cohort 2 Class	127.4	124.0	10	420
Sample Size: 13 teachers.				

Note: Means are unadjusted, and based on all *Sam and Pat* teachers for whom data were available.
Source: Adult ESL Literacy Impact Study spring 2009 *Sam and Pat* teacher data form.

Lesson Number Completed

The *Sam and Pat* text includes 22 lessons. The developers stated that teachers should move through the book at a pace that works for their students, and that teachers of literacy learners should not be expected to make it through the entire book in one term. Therefore, we expected to see a range of responses from teachers on their final lesson covered, and that is what we found (Table 3.6). Overall, the final lesson number covered ranged from 3 to 22, with an average of 13 (Table 3.6). An additional descriptive table on the distribution of final *Sam and Pat* lesson numbers covered (Table E.1) is provided in Appendix E.

Table 3.6: Final *Sam and Pat* Lesson Number Covered in Class, as Reported by *Sam and Pat* Teachers

	Mean	Std	Min	Max
Overall	**13.2**	**5.4**	**3**	**22**
For Cohort 1 Class	11.7	5.3	3	22
For Cohort 2 Class	14.7	5.3	6	22
Sample Size: 22 classes (10 missing cases).				

Note: Means are unadjusted, and based on all *Sam and Pat* classes for whom data were available.
Source: Adult ESL Literacy Impact Study spring 2009 *Sam and Pat* teacher data form.

Use of Materials Other than Sam and Pat

Teachers were expected to supplement *Sam and Pat* with other materials (e.g., ESL texts), based on their program's standards and ESL curricula. Teachers could also reinforce *Sam and Pat* with materials they made themselves, such as handouts related to a *Sam and Pat* lesson. Table 3.7 shows that most teachers did use additional materials beyond *Sam and Pat* during the study, to varying extents. For example, 46 percent of teachers reported using material from a second workbook or text three or more times per month, and 77 percent of teachers reported using worksheets that they or another teacher created three or more times per month.

Table 3.7: Percent of Teachers Who Reported Supplementing *Sam and Pat* Instruction With Other Materials During the Study, by Frequency of Use, as Reported by *Sam and Pat* Teachers

Materials	Used Less Than Three Times Per Month	Used Three or More Times Per Month
A second or third workbook or text, or handouts from those workbooks or texts	53.8	46.2
Teacher-created worksheets	23.1	76.9
Dictionaries or picture dictionaries	61.5	38.5
Other (stories or paragraphs, computer software, e-mail, Web pages, or video or audio recordings)	61.5	38.5
Sample Size: 13 teachers.		

Notes: Details may not sum to totals due to rounding. Means are unadjusted, and based on all *Sam and Pat* teachers for whom data were available.
Source: Adult ESL Literacy Impact Study spring 2009 *Sam and Pat* teacher data form.

Sam and Pat and Control Group Differences in Instruction and Student Attendance

More Reading Instruction Observed in Sam and Pat Classes, while More English Language Instruction Observed in Control Classes

This section summarizes the instructional service contrast across all *Sam and Pat* and control classrooms. The purpose of this analysis was to compare instruction in the *Sam and Pat* classes to that in the control classes. In total, 64 classroom observations were conducted—33 during Cohort 1 and 31 during Cohort 2 (Table 3.8).

Table 3.8: Number of Classroom Observations, by Cohort and Group

Cohort	*Sam and Pat*	Control
1	16	17
2	15	16
Total	**31**	**33**

Table 3.9 presents the number of intervals observed for each instructional area by group. It is important to note that multiple instructional areas could have been coded within each interval. As a result, percentages do not sum to 100. The primary instructional areas associated with *Sam and Pat* are the following reading development codes:

- Phonics;
- Writing and spelling for phonics reinforcement;
- Learning vocabulary to reinforce reading instruction;
- Fluency and accuracy in reading; and
- Reading comprehension.

Therefore, we expected to see a higher percentage of *Sam and Pat* classrooms' observation intervals spent in those five areas as compared to control classrooms. In control classrooms, we expected to see a greater percent of observation intervals spent in the English Language Acquisition codes, specifically with respect to the following areas:

- Oral communication skills—Listening;
- Oral communication skills—Speaking;
- Grammar;
- English vocabulary and idioms; and
- Sociocultural knowledge (cultural facts, life skills, etc.).

Table 3.9: Percent of Instructional Intervals Spent in Key Instructional Areas, by Group

	Sam and Pat mean	Control mean	Difference	P-Value of difference
Reading Development	**65.5**	**19.3**	**46.3**	**0.000***
Pre-literacy	7.3	2.2	5.1	0.001*
Phonics	19.5	5.8	13.7	0.000*
Writing & Spelling for Phonics Reinforcement	11.8	2.5	9.3	0.000*
Learning Vocabulary to Reinforce Reading Instruction	19.9	4.1	15.8	0.000*
Fluency & Accuracy in Reading	15.5	8.1	7.4	0.026*
Reading Comprehension	13.8	3.7	10.1	0.000*
English Language Acquisition	**27.3**	**67.6**	**–40.2**	**0.000***
Oral Communication—Listening	2.9	10.2	–7.3	0.005*
Oral Communication—Speaking	8.8	25.8	–17.0	0.000*
Grammar: Understanding How English Works	17.3	33.3	–16.0	0.004*
English Vocabulary & Idioms	1.9	24.5	–22.7	0.000*
Sociocultural Knowledge	0	4.6	–4.6	0.040*
Other Instructional Areas:				
Writing Unrelated to Reading Activities	2.9	10.8	–7.9	0.054
Functional Reading, Writing, & Math	4.6	18.0	–13.4	0.008*
Other Instruction and Breaks	8.7	11.1	–2.4	0.191
Links to Outside World	1.3	7.4	–6.1	0.071
Use of Students' Native Language	20.7	43.9	–23.2	0.000*
Sample Sizes: Number of Intervals	980	1034		
Number of Observations	31	33		

* Indicates a difference that is significant at the 0.05 level, based on a 2-tailed t-test.
Notes: Details may not sum to totals. Practices may be coded under multiple content areas during any one interval. Estimates are based on pooled observation intervals that have been regression-adjusted by dummy variables representing the sites at which instruction occurred. Calculations used data from all classes for which data were available.
Source: Adult ESL Literacy Impact Study classroom observation protocol.

As the data show, the *Sam and Pat* classrooms did in fact experience a higher percentage of instructional intervals in the five reading development content areas discussed above (66 percent in *Sam and Pat* classrooms compared to 19 percent in control classrooms), and the difference was statistically significant. Conversely, the control classrooms experienced a higher percentage of instructional intervals in English language acquisition content areas (68 percent in control classrooms compared to 27 percent in *Sam and Pat* classrooms), and this difference was also statistically significant.

Other instructional areas that showed statistically significant differences between study groups were Functional Reading, Writing, and Math (18 percent in control classrooms compared to 5 percent in *Sam and Pat* classrooms), and Use of Students' Native Language (44 percent in control classrooms compared to 21 percent in *Sam and Pat* classrooms). No other significant differences were found among other instructional areas measured.

No Group Differences in Hours of Class Attended

To document hours attended and provide further contextual information for the impact findings, we collected attendance records from each participating class. This section compares the number of class hours attended by students in *Sam and Pat* and control classrooms.

The student persistence differences were estimated using a two-level hierarchical model identical to the model used to estimate student outcome impacts. The model is described in Chapter 4.

As shown in Table 3.10, the difference between the mean hours of attendance in the *Sam and Pat* group (79 hours) and in the control group (72 hours) was not statistically significant.[16]

An additional descriptive table on the distribution of student attendance hours (Table E.2) is provided in Appendix E.

---

[16] Because we did not observe all hours of instruction throughout the term, we cannot determine whether the 79 hours of *Sam and Pat* attendance included the 60 hours of *Sam and Pat* instruction recommended by the developers of the text. We can therefore only characterize implementation by reporting that (1) 65 percent of *Sam and Pat* classes met the study's fidelity criteria, and (2) significantly more reading instruction was delivered in these classes, as compared to the control group classes, as described in this chapter.

Table 3.10: Hours of Attendance, by Group

Outcome	*Sam and Pat* Group	Control Group	Diff.	P-Value for Difference
Hours of Class Attended	79.4	71.9	7.5	0.337
Sample Size: 1,344 Students (674 *Sam and Pat* and 670 control)				

Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students and background characteristics of teachers. Calculations used data for the full student sample. A two-tailed t-test was applied to the difference between the *Sam and Pat* and control groups. The difference was not statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study attendance database, student intake forms (fall 2008 and spring 2009), and fall 2008 teacher data form.

CHAPTER 4:
IMPACTS ON READING AND ENGLISH LANGUAGE SKILLS

To test the impacts of *Sam and Pat* on reading and English language outcomes, each student was administered a battery of assessments prior to and following the term-long intervention (pre- and post-test batteries). These assessments were selected to measure the range of skills that could potentially be impacted by *Sam and Pat*–based instruction. In this chapter we present the results of the impact analyses for the following assessments:

Reading Skills Assessments	
• Woodcock-Johnson Letter-Word Identification (WJID)	• Woodcock-Johnson Passage Comprehension (WJPC)
• Woodcock-Johnson Word Attack (WJWA)	• SARA Decoding (SARAdec)
English Language Skills Assessments	
• Oral and Written Language Scales (OWLS)	• Receptive One-Word Picture Vocabulary Test (ROWPVT)
• Woodcock-Johnson Picture Vocabulary Test (WJPV)	

We first present the overall impacts by comparing scores on each assessment for students in the *Sam and Pat* versus control groups. We then present impacts for special subgroups of students, such as students with lower reading scores at the beginning of the term.

Estimation Model

The basic analytic strategy for assessing the impacts of *Sam and Pat* was to compare reading and English language outcomes for students who were randomly assigned to either the *Sam and Pat* or the control group.[17] The average outcome in the control group represents an estimate of the scores that would have been observed in the *Sam and Pat* group if they had not received the intervention; therefore, the difference in outcomes between the *Sam and Pat* and control groups provides an unbiased estimate of the impacts of *Sam and Pat*.

[17] In sites with only one pair of study classes (and therefore one *Sam and Pat* teacher), the threat to internal validity caused by the confounding of the teacher and the intervention was dealt with by (1) confirming the integrity of the random assignment (described in Chapter 2) by testing the baseline equivalence of the *Sam and Pat* and control groups on a range of teacher characteristics; and (2) statistically controlling for teacher characteristics in the impact analyses, as well as controlling for the site in which the teachers were located.

Given the nested structure of the data, impacts were estimated using a two-level regression model where the first level was the student and the second level was the teacher.[18] In each regression equation, the dependent variable was the post-test score on each assessment. The independent variables included a *Sam and Pat*-control group dummy variable, a site dummy variable, the pre-test score for each assessment, student-level covariates, and teacher-level covariates. The full list of student- and teacher-level covariates and additional details on the estimation model are included in Appendix D.

Impacts on Students' Reading and English Language Skills

No Impacts on Reading Outcomes

The impacts on the four reading outcomes were not statistically significant. Effect sizes ranged from –0.05 to 0.01 (Table 4.1). This pattern of results indicates that instruction incorporating the *Sam and Pat* intervention was not more effective at raising reading scores on the reading skills measured than the study sites' business-as-usual instruction.[19]

---

[18] There were several reasons for not also including site or class (i.e., cohort within teacher) levels in the model. First, it was determined that a site level was unnecessary. A Wald test of the differences in impacts across sites was not statistically significant ($p = 0.397$; see Figures F.1-F.7 in Appendix F for results by site). Therefore, we pooled the sample across sites and accounted for site differences by including sites as fixed effects in the impact model. Second, when we tried including both teacher and class levels in the model, we experienced problems obtaining stable estimates, and the likelihood function would not converge. Maximum likelihood estimation convergence problems like these usually indicate that there is insufficient independent information in the data to estimate random effects at each level. Similarly, it was not possible to account for both site and class pair level random assignment blocking in the model; the two blocking factors overlapped completely in 4 study sites (i.e., there was only one class pair at each site).

[19] As a data quality check, ten percent of the WJ Letter Word Identification, WJ Word Attack, and SARA Decoding assessment scores were randomly selected and rescored by staff with expertise on the assessment. Details on the rescoring methods can be found in Appendix A. A combined measure of the rescored and original scores was used in the impact analyses presented in Table 4.1, whereby the original scores were replaced with the revised scores for the ten percent of the scores that were rescored. Additional details on the rescoring methods can be found in Appendix A. A table of impacts based on the original scores of the assessments is provided in Appendix F (Table F.1).

No Impacts on English Language Outcomes

For the English language assessments OWLS and ROWPVT, we measured impacts using raw scores.[20] As with the results for reading, none of the impacts on the English language outcomes measured were statistically significant. Effect sizes ranged from –0.06 to 0.01 (Table 4.1).

Table 4.1: Impact of *Sam and Pat* on Reading and English Language Outcomes

Outcome	*Sam and Pat* Group	Control Group	Diff.	Effect Size	P-Value for Difference
Reading Assessments					
Woodcock Johnson Letter Word Identification Scale (Rescored)	440.611	442.223	–1.612	–0.030	0.477
Woodcock Johnson Word Attack Scale (Rescored)	466.495	465.893	0.602	0.015	0.732
SARA Decoding (Rescored)	13.230	13.383	–0.153	–0.014	0.753
Woodcock Johnson Passage Comprehension Scale	432.740	433.626	–0.885	–0.049	0.226
English Language Assessments					
OWLS	17.870	17.788	0.081	0.008	0.892
ROWPVT	28.490	29.614	–1.124	–0.065	0.106
Woodcock Johnson Picture Vocabulary Scale	431.545	431.311	0.234	0.012	0.806
Sample Size: 1,137 students (580 *Sam and Pat;* 557 control).					

Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Putting the Findings into Context

Table 4.1 shows that, overall, no impacts on reading and English language were found for *Sam and Pat*. In this section, we examine students' gains between the beginning of the term (pre-test) and the end of the term (post-test) to provide context for these findings. For example, it is possible that no impacts were found

[20] The OWLS and ROWPVT assessments are scaled to measure language skills at ages 12 and 17. However, because the scaled scores exhibited floor effects, raw scores were used instead. A table showing the impacts based on scaled scores is provided in Table F.2 of Appendix F. In addition, a table showing the impacts based on Woodcock-Johnson raw scores is provided in Table F.3.

because students in both groups, on average, did not make gains on any of the assessments. The opposite could be true as well; it may be that both groups made gains, but the magnitude of the gains was similar between the two groups. Because we have already tested the group differences in outcomes while controlling for pre-tests, we can assume similar gains were made, on average, by each group.[21] The real question that needs to be addressed, therefore, is whether or not students (overall) made gains in their reading and English language skills during the study.

First, we take a look at the change in scores from pre- to post-test for each of the assessments administered at both the beginning and end of the term. Table 4.2 shows that the mean gains (difference between pre- and post-test scores) across all reading and English language assessments were statistically significant (effect sizes of 0.23 to 0.40).

Table 4.2: Mean Pre- vs. Post-Test Scores on Reading and English Language Assessments

Outcome	Mean Pre-Test Score	Mean Post-Test Score	Overall Mean Gain (Diff.)	Effect Size	P-Value for Difference
Reading Assessments					
Woodcock Johnson Letter Word Identification Scale (Rescored)	428.315	442.122	13.808	0.260	0.000*
Woodcock Johnson Word Attack Scale (Rescored)	457.400	466.503	9.103	0.227	0.000*
Woodcock Johnson Passage Comprehension Scale	427.061	433.780	6.719	0.364	0.000*
English Language Assessments					
OWLS	14.239	18.075	3.836	0.383	0.000*
ROWPVT	22.898	29.285	6.387	0.399	0.000*
Sample Size: 1,113 students (567 *Sam and Pat*; 546 control).					

*Indicates that difference is significant at 5 percent level, based on 2-tailed dependent t-tests.
Notes: These figures are not regression-adjusted. Only assessments administered at both pre- and post-testing were included in this table. Calculations used data for all students for whom both pre- and post-test data were available.
Source: Adult ESL Literacy Impact Study assessments administered at the beginning and end of each term (fall 2008 and spring 2009).

These data indicate that students made statistically significant gains from the beginning of the term to the end of the term. To help interpret the magnitude of these gains, we converted pre- and post-test mean scores on each test to a grade

[21] A table of the gains made by group is available in Table F.4 of Appendix F. There were no statistically significant group differences in the gains made.

level equivalent (GLE) for the reading scores and an age equivalent (AE) for the English language scores.[22]

Table 4.3 presents the GLEs and AEs for each assessment administered at the beginning and end of the term. Students' reading scores were equivalent to grades K.9 to 2.2 (i.e., 9 months into grade K and 2 months into grade 2, respectively) at the beginning of the term and 1.0 to 2.4 at the end of the term, which corresponds to 1 to 2 months of growth in the reading skills measured.

For the English language assessments, mean scores were equivalent to ages 2 years and one month of age (ROWPVT) and 2 years and 4 months of age (OWLS) at the beginning of the term, and 2 years and 7 months of age (ROWPVT) to 2 years and 9 months of age (OWLS) at the end of the term. This translates into approximately 5 to 6 months of growth in the English language skills measured.

Based on these results, and those presented earlier in this chapter, students made statistically significant gains on reading and English language outcomes, although at the end of the term there were no statistically significant differences in outcomes between the *Sam and Pat* and control group.

---

[22] The GLEs and AEs are based on test publisher guidelines (Brownell, 2000; Carrow-Woolfolk, 1996; Woodcock, McGrew, & Mather, 2001). The GLE is a .0 to .9 scale based on a 10 month school year (September to June), where each tenth would translate to approximately 1 month of gains. The AE is based on age, which for the OWLS and ROWPVT starts at approximately 2-0 years (2 years and 0 months) of age. It should be noted that publisher guidelines for GLE and AE calcuations are based on norming populations that differ from the study population. (The WJ assessments were normed on a nationally representative sample of U.S. residents aged 2 to 90+; the OWLS on a representative U.S. sample aged 3 to 21 years; and the ROWPVT on a representative U.S. sample aged 2 to 18 years.) No norming data exist for low-literate adult ESL learners. Additionally, the study used simplified or translated testing instructions when students did not appear to understand the tester's directions (see Appendix A for a summary of these adaptations). For these reasons, GLEs and AEs should be interpreted with caution.

Table 4.3: Grade Level or Age Equivalents (GLEs/AEs) for Pre- and Post-Test Means

Outcome	Mean Pre-Test GLE/AE	Mean Post-Test GLE/AE	Gain (in Months)
Reading Assessments (in GLEs)			
Woodcock Johnson Letter Word Identification Scale	2.0	2.2	2
Woodcock Johnson Word Attack Scale	2.2	2.4	2
Woodcock Johnson Passage Comprehension Scale	K.9	1.0	1
English Language Assessments (in AEs)			
OWLS	2-4	2-9	5
ROWPVT	2-1	2-7	6
Sample Size: 1,113 students (567 *Sam and Pat;* 546 control).			

Notes: Only assessments administered at both pre- and post-testing were included in this table. Calculations used data for all students for whom both pre- and post-test data were available.
Source: Adult ESL Literacy Impact Study assessments administered at the beginning and end of each term (fall 2008 and spring 2009).

Subgroup Analyses

Overview of Subgroups

Although there were no overall impacts on the reading and English language skills tested, those results may mask underlying variation among special subpopulations. To test whether *Sam and Pat* was effective for any of our groups of interest, we also estimated impacts for the following subgroups:

- Native language group
  - Non-Roman alphabet background
  - Native Spanish speakers
- Higher and lower literacy level
- Cohorts 1 and 2

Non-Roman alphabet background. Adult learners whose native language is not based on the Roman alphabet (e.g., Chinese, Arabic, etc.) may encounter difficulties learning the English alphabet, even if they have some literacy in their native language (Birch, 2002). Some of these learners (e.g., Arabic literacy learners) may also need to become accustomed to the directionality (left-to-right) of the English language (Sherow, 2006; NCFL, 2004).

While the research suggests that students with a non-Roman alphabet background may face challenges translating their native language into reading and writing English, the developers of *Sam and Pat* reported designing the intervention to be appropriate for this group. Therefore, it was of interest to test impacts among students with a non-Roman alphabet background.

Native Spanish speakers. Although no national statistics exist on the distribution of language groups among low-literate adult ESL learners attending adult English language classes, anecdotal evidence and smaller scale studies suggest that Spanish speakers comprise the majority of this category of learners (e.g., Condelli et al., 2003). What works to improve English literacy and language skills for this group is therefore of key interest to practitioners and policymakers. In addition, an earlier study on adult ESL literacy learners found that students with a Spanish language background responded to instruction differently than the other language groups represented in the study (Condelli et al., 2003). Therefore, we investigated impacts for Spanish speakers.

Literacy level at the beginning of the term. The language and reading development of adult ESL learners can be predicted by a student's literacy level prior to formal English language instruction (Condelli et al., 2003). Therefore, we analyzed the intervention's effects separately for students who tested at relatively lower ("lower literacy") and higher ("higher literacy") reading levels at the beginning of the term. Lower literacy was defined as scoring at a Grade 2 equivalent or below on the Woodcock Johnson Letter-Word Identification and Word Attack subtests (raw scores of 31 and 9, respectively).[23] Higher literacy was defined as scoring above those scores, although it should be noted that these students are still categorized as literacy level by their ESL programs.

Cohort 1 or 2. The *Sam and Pat* intervention was implemented by the same group of teachers over the course of two consecutive terms (Cohorts 1 and 2). As a result, teachers gained experience teaching the *Sam and Pat* curriculum during the first term, which we hypothesized would benefit the *Sam and Pat* instructors' teaching quality during the second term. Therefore, we examined impacts separately for each cohort.

[23] See Table F.7 for the percent of students defined as lower literacy at each site.

No Impacts on Reading and English Language Outcomes Found for Subgroups Based upon Native Language and Cohort

There were no statistically significant impacts found for students with a non-Roman-based alphabet background, native Spanish speakers, students from the first study cohort, or students from the second study cohort. (Tables 4.4–4.7).

Some Suggestive Evidence of A Positive Impact on Reading Outcomes for Lower Literacy Students

No statistically significant impacts were found for higher literacy level students (Table 4.6; bottom panel). However, there was a suggestive finding for students who tested in the lower literacy score range at the beginning of the term (Table 4.6; top panel). Within this subgroup, *Sam and Pat* group students scored higher on the WJ word attack assessment than control group students (445 and 439, respectively; effect size = 0.16).

The WJ word attack assessment tests students' decoding skills; the pattern found for the lower literacy students on this measure is consistent with the focus on decoding instruction in *Sam and Pat* classrooms (phonics and writing for phonics reinforcement, as shown in Chapter 3; Table 3.9). In addition, students in the lower literacy subgroup would have the most to gain in that skill area (reading) if targeted by instruction. Because the difference between the *Sam and Pat* and control groups was not statistically significant after correcting for multiple comparisons, however, it is possible that the effect is due to chance alone.[24]

[24] We corrected for multiple comparisons using the Benjamini-Hochberg Procedure. Details about how this procedure was employed are provided in Appendix D.

Table 4.4: Impact of *Sam and Pat* on Reading and English Language Outcomes Among Students With a Non-Roman-based Alphabet Background

Outcome	*Sam and Pat* Group	Control Group	Diff.	Effect Size	P-Value for Difference
Reading Assessments					
Woodcock Johnson Letter Word Identification (Rescored)	431.648	432.030	–0.382	–0.008	0.895
Woodcock Johnson Word Attack Scale (Rescored)	464.172	462.252	1.920	0.047	0.527
SARA Decoding (Rescored)	10.879	11.499	–0.620	–0.062	0.437
Woodcock Johnson Passage Comprehension Scale	432.619	434.919	–2.300	–0.137	0.062
English Language Assessments					
OWLS	16.868	17.161	–0.294	–0.029	0.668
ROWPVT	23.100	24.170	–1.070	–0.091	0.188
Woodcock Johnson Picture Vocabulary Scale	426.298	428.253	–1.955	–0.113	0.221
Sample Size: 434 non-Roman alphabet students	212	222			

Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom there were post-test data and data on the variable that defined the subgroup. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Table 4.5: Impact of *Sam and Pat* on Reading and English Language Skills Among Spanish Speaking Students

Outcome	*Sam and Pat* Group	Control Group	Diff.	Effect Size	P-Value for Difference
Reading Assessments					
Woodcock Johnson Letter Word Identification (Rescored)	457.626	462.125	–4.499	–0.084	0.334
Woodcock Johnson Word Attack Scale (Rescored)	479.766	479.603	0.163	0.005	0.955
SARA Decoding (Rescored)	16.835	18.569	–1.019	–0.097	0.285
Woodcock Johnson Passage Comprehension Scale	438.329	439.724	–1.394	–0.090	0.152
English Language Assessments					
OWLS	20.001	19.771	0.230	0.022	0.766
ROWPVT	35.565	37.326	–1.761	–0.087	0.233
Woodcock Johnson Picture Vocabulary Scale	437.905	438.588	–0.684	–0.034	0.673
Sample Size: 503 Spanish-speaking students	252	251			

Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom there were post-test data and data on the variable that defined the subgroup. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Table 4.6: Impact of *Sam and Pat* on Reading and English Language Skills Among Students With Lower and Higher Literacy Levels at the Beginning of the Term

Outcome	*Sam and Pat* Group	Control Group	Diff.	Effect Size	P-Value for Difference
Students with Lower Literacy at Beginning of Term					
Reading Assessments					
Woodcock Johnson Letter Word Identification (Rescored)	408.280	407.566	0.714	0.017	0.827
Woodcock Johnson Word Attack Scale (Rescored)	445.430	439.111	6.320	0.156	0.047*
SARA Decoding (Rescored)	7.179	6.407	0.772	0.103	0.200
Woodcock Johnson Passage Comprehension Scale	422.026	422.386	–0.360	–0.023	0.747
English Language Assessments					
OWLS	12.794	13.068	–0.274	–0.033	0.637
ROWPVT	20.288	22.837	–2.549	–0.211	0.109
Woodcock Johnson Picture Vocabulary Scale	420.903	421.488	–0.585	–0.033	0.690
Sample Size: 502 lower literacy students	248	254			
Students with Higher Literacy at Beginning of Term					
Reading Assessments					
Woodcock Johnson Letter Word Identification (Rescored)	465.049	467.670	–2.621	–0.059	0.426
Woodcock Johnson Word Attack Scale (Rescored)	483.071	485.275	–2.204	–0.083	0.257
SARA Decoding (Rescored)	17.799	18.592	–0.793	–0.083	0.258
Woodcock Johnson Passage Comprehension Scale	441.090	441.804	–0.714	–0.051	0.368
English Language Assessments					
OWLS	21.471	21.426	0.045	–0.004	0.961
ROWPVT	34.598	35.005	–0.407	–0.022	0.672
Woodcock Johnson Picture Vocabulary Scale	439.393	438.771	0.622	0.035	0.606
Sample Size: 635 higher literacy students	332	303			

*Indicates that impact is significant at 5 percent level, based on 2-tailed t-tests. No impacts were significant after adjusting for multiple comparisons.

Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom there were post-test data and data on the variable that defined the subgroup.

Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Table 4.7: Impact of *Sam and Pat* on Reading and English Language Skills Among Students in Cohort 1 and Cohort 2

Outcome	*Sam and Pat* Group	Control Group	Diff.	Effect Size	P-Value for Difference
Cohort 1					
Reading Assessments					
Woodcock Johnson Letter Word Identification (Rescored)	434.876	434.372	0.303	0.006	0.914
Woodcock Johnson Word Attack Scale (Rescored)	461.325	460.295	1.030	0.025	0.654
SARA Decoding (Rescored)	11.637	11.841	–0.204	–0.019	0.745
Woodcock Johnson Passage Comprehension Scale	430.998	432.064	–1.066	–0.058	0.231
English Language Assessments					
OWLS	17.683	16.824	0.858	0.086	0.399
ROWPVT	26.921	28.154	–1.233	–0.075	0.142
Woodcock Johnson Picture Vocabulary Scale	430.253	430.135	0.119	0.006	0.921
Sample Size: 684 Cohort 1 students	345	339			
Cohort 2					
Reading Assessments					
Woodcock Johnson Letter Word Identification (Rescored)	448.884	452.455	–3.571	–0.064	0.378
Woodcock Johnson Word Attack Scale (Rescored)	473.270	473.660	–0.389	–0.010	0.891
SARA Decoding (Rescored)	15.294	15.723	–0.429	–0.040	0.598
Woodcock Johnson Passage Comprehension Scale	435.056	435.922	–0.866	–0.050	0.400
English Language Assessments					
OWLS	18.263	18.818	-0.555	–0.051	0.393
ROWPVT	30.869	31.615	-0.746	–0.040	0.646
Woodcock Johnson Picture Vocabulary Scale	433.593	432.518	1.075	0.051	0.512
Sample Size: 453 Cohort 2 students	235	218			

Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom there were post-test data and data on the variable that defined the subgroup. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

CHAPTER 5:
NON-EXPERIMENTAL ANALYSES

In Chapter 4, we presented evidence that, overall, the specific intervention being tested did not have statistically significantly impacts on reading or English language outcomes. It is still possible, however, that reading and language instruction—regardless of whether it is delivered in a *Sam and Pat* or a control classroom—is related to reading and English language outcomes.

As explained in Chapter 2, we conducted one classroom observation per class. On average, we recorded thirteen 5-minute intervals spent on reading instruction, and 17 intervals spent on English language instruction. While one may expect that the number of reading instruction intervals would have a positive relationship with student reading assessment scores and that a similar relationship would exist between English language instruction and language outcomes, a negative relationship is also possible. For example, teachers of students who are struggling with reading might be more likely to provide reading instruction, whereas teachers of students who are more proficient in reading to start with might provide less reading instruction. The result would be a negative relationship between the amount of reading instruction and reading performance on the post-test.

In this chapter, we present results from our non-experimental analysis of the relationship between reading or English language instruction and the outcomes measured in those domains, pooling data across the *Sam and Pat* and control groups. We interpret the results with caution because the analyses are correlational in nature; this means that the results can provide information on relationships between the instruction and outcomes measured, but cannot be used to infer causation.

We also test the relationship of the combination of instruction and attendance hours in predicting outcomes, because the relationship between instruction and outcomes may depend on the amount of exposure students have to the instruction. Specifically, for each outcome, we explored the role of instruction and attendance by testing the following predictors: (1) percent of intervals of instruction in each of the key reading and English language content areas, (2) total attendance hours, and (3) the interaction of instruction and attendance (exposure to reading instruction) and the interaction of English language instruction and attendance (exposure to English language instruction).[25] In contrast to the impacts presented in Chapter 4, these analyses are exploratory in nature and are designed to describe

[25] The regression model used in these analyses included the same covariates and site indicator variables included in the impact models, as described in Appendix D.

patterns and interactions between variables. Any patterns or relationships found should not be taken as causal.

No Direct Relationship Between Reading or English Language Instruction and Outcomes

Table 5.1 presents the relationship between the percent of observational intervals that included reading or English language instruction and student performance on the assessments, without taking into account hours of attendance. Coefficients for the instructional variables ranged from –5.14 to 6.0 and were not statistically significant, meaning that the instructional variables were not predictive of outcomes on their own.[26]

Table 5.1: Relationship Between Reading and English Language Instruction and Outcomes

Outcome	Percent Reading Composite Coefficient (Standardized)	P-Value	Percent Language Composite Coefficient (Standardized)	P-Value
Reading Assessments				
Woodcock Johnson Letter Word Identification (Rescored)	–5.137	0.229	5.898	0.232
Woodcock Johnson Word Attack Scale (Rescored)	2.225	0.498	–2.116	0.577
Woodcock Johnson Passage Comprehension Scale	–0.669	0.592	0.326	0.821

Table continued next page.

[26] We also tested the overall joint significance of the regression coefficients in the nonexperimental models using a chi-squared test, which is the F-test equivalent in a mixed model. For all of the nonexperimental models presented in Chapter 5, the chi-squared values were significant at $p < .05$. This indicates that although none of the instructional composite coefficients were significant (Table 5.1), and only some of the instructional exposure coefficients were significant (Table 5.3) whereas all the attendance coefficients were significant (Table 5.2), the overall model that included the variables of interest as well as the covariates was significantly predictive of post-test scores.

Table 5.1: Relationship Between Reading and English Language Instruction and Outcomes (Continued)

Outcome	Percent Reading Composite Coefficient (Standardized)	P-Value	Percent Language Composite Coefficient (Standardized)	P-Value
English Language Assessments				
OWLS	–0.226	0.780	0.634	0.496
ROWPVT	–1.659	0.172	2.293	0.102
Woodcock Johnson Picture Vocabulary Scale	1.428	0.445	–2.795	0.211
Sample Size: 1,137 Students (587 *Sam and Pat* and 557 control).				

Notes: Estimates are based on multilevel models, whereby assessment outcomes are regressed on the percent of reading and English language instruction intervals per class, pre-test scores, and teacher and student demographic variables, and control for clustering at the class level. Calculations used data for all students for whom post-test data were available. The displayed p-values are the independent variable coefficient p-values. They indicate the probability of obtaining a z-score (coefficient divided by the standard error) at least as large as the one measured. No z-tests were statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study classroom observation protocol, student intake forms, assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Positive (although Weak) Relationship Between Attendance and Reading and English Language Outcomes

Table 5.2 presents findings on the relationship between student attendance hours and reading and language assessment scores, without accounting for type of instruction. For these analyses, assessment scores were regressed on attendance hours and all of the covariates included in the impact analyses. The attendance hours coefficients indicate that attendance hours have a positive and statistically significant relationship with all student outcomes, holding other factors constant, but the magnitude of each coefficient is small (0.03 to 0.10). For example, the 0.07 coefficient for the Woodcock Johnson Word Attack assessment indicates that a 10-hour increase in the number of hours attending class is associated with a 0.70 score increase on that assessment (for which the sample mean was 467). One must also consider the possibility that the observed relationship between attendance and assessment scores is not causal; the relationship may be due to an unobserved factor, such as motivation, which is correlated with attendance. For example, students with higher attendance may have more motivation to learn and therefore

perform better on the student assessments than other students.[27] In the current study, there is no way to disentangle program attendance from such potential unobserved factors as motivation.

Table 5.2: Relationship Between Hours of Attendance and Outcomes

Outcome	Attendance Hours Coefficient	P-Value
Reading Assessments		
Woodcock Johnson Letter Word Identification (Rescored)	0.104*	0.000
Woodcock Johnson Word Attack Scale (Rescored)	0.071*	0.000
Woodcock Johnson Passage Comprehension Scale	0.043*	0.000
Oral Language Assessments		
OWLS	0.027*	0.000
ROWPVT	0.028*	0.000
Woodcock Johnson Picture Vocabulary Scale	0.056*	0.000
Sample Size: 1,137 Students (587 *Sam and Pat* and 557 control).		

*Indicates that the coefficient is significant at the 0.05 level, based on 2-tailed z-tests.
Notes: Estimates are based on multilevel models, whereby assessment outcomes are regressed on the total number of attendance hours, pre-test scores, and teacher and student demographic variables, and control for clustering at the class level. Calculations used data for all students for whom post-test data were available. The displayed p-values are the independent variable coefficient p-values. They indicate the probability of obtaining a z-score (coefficient divided by the standard error) at least as large as the one measured.
Source: Adult ESL Literacy Impact Study attendance database, student intake forms, assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Student Exposure to Reading or English Language Instruction Unrelated to Most Reading and English Language Outcomes Measured, Although Weak Relationships Found Between Exposure to Instruction and One English Language Outcome

Finally, we examined the relationship between the interaction of student attendance hours and instruction (i.e., exposure to reading or English language instruction) and student outcomes. The results in Table 5.3 indicate no statistically significant relationships between exposure to instruction and any of the reading outcomes measured and two of the three English language outcomes measured.

[27] Moreover, if the relationship between attendance and student assessment scores is due to an unobserved factor, the direction (positive or negative) of the coefficient bias is unclear. For example, it is also plausible that students with higher attendance need more instruction and are less likely to score highly on the student assessments.

However, the amount of exposure to English language instruction, measured by the combination of English language instruction and attendance hours, was positively and statistically significantly correlated with ROWPVT scores. The opposite pattern was found for reading instruction; exposure to reading instruction had a negative and statistically significant relationship with scores on the ROWPVT. However, the standardized coefficients in both cases were small (0.034 and –0.032, respectively). As an example, the 0.034 coefficient on the ROWPVT assessment indicates that, after controlling for total student attendance hours, an increase of 10 percent in the number of English language instruction intervals a student attended is associated with a 0.34 point increase on the test (which had a sample mean of 29). In addition, similar to the student attendance results, we cannot rule out the possibility that the statistically significant relationships were driven by other factors. Therefore, these findings should be interpreted with caution.

Table 5.3: Relationship Between Exposure to Instruction and Outcomes

Outcome	Reading Composite: Attendance Interaction Coefficient	P-Value	Language Composite: Attendance Interaction Coefficient	P-Value
Reading Assessments				
Woodcock Johnson Letter Word Identification (Rescored)	-0.088	0.062	0.048	0.328
Woodcock Johnson Word Attack Scale (Rescored)	0.000	0.985	-0.015	0.686
Woodcock Johnson Passage Comprehension Scale	-0.025	0.067	0.012	0.397
English Language Assessments				
OWLS	-0.006	0.524	0.005	0.587
ROWPVT	-0.032*	0.018	0.034*	0.016
Woodcock Johnson Picture Vocabulary Scale	0.004	0.862	-0.028	0.191
Sample Size: 1,137 Students (587 *Sam and Pat* and 557 control).				

*Indicates that the coefficient is significant at the 0.05 level, based on 2-tailed z-tests.
Notes: Estimates are based on multilevel models, whereby assessment outcomes are regressed on the interaction of percent of reading or English language instruction units per class and total number of attendance hours, pre-test scores, and teacher and student demographic variables, and control for clustering at the instructor level. Calculations used data for all students for whom post-test data were available. The displayed p-values are the independent variable coefficient p-values. They indicate the probability of obtaining a z-score (coefficient divided by the standard error) at least as large as the one measured.
Source: Adult ESL Literacy Impact Study classroom observation protocol, attendance database, student intake forms, assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

REFERENCES

Adams, M. J. (1991). *Beginning to Read*. Cambridge: MIT Press.

Benjamini, Y., & Hochberg, Y. (1995). "Controlling the false discovery rate: A new and powerful approach to multiple testing." *Journal of the Royal Statistical Society, Series B, 57,* 1289–1300.

Birch, B. M. (2002). *English L2 reading: Getting to the bottom*. Mahwah, NJ: Erlbaum.

Bloom, H. S. 1984. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review*, Vol. 8.pp. 225 – 246.

Brown, R., Pressley, M., Van Meter, P., & Schuder, T. (1996). "A quasi-experimental validation of transactional strategies instruction with previously low-achieving, second-grade readers." *Journal of Educational Psychology*, 88, 18–37.

Brownell, R. (Ed). (2000). Expressive One-Word Picture Vocabulary Test manual (3rd ed). Novato, CA: Academic Therapy Publications.

Carrell, P. L. (1985). "Facilitating ESL reading by teaching text structure." *TESOL Quarterly*, 19, 727–752.

Carrow-Woolfolk, E. (1996). Oral Written and Language Scales. Princeton, NJ: Pearson Education.

Celce-Murcia, M. (2001). "Language teaching approaches: An overview." In M. Celce-Murcia (Ed.), *Teaching English as a Second or Foreign Language* (3rd ed.). Boston: Heinle & Heinle.

Cheek, E. H., & Lindsey, J. D. (1994). "The effects of two methods of reading instruction on urban adults' word identification and comprehension abilities." *Journal of Vocational and Technical Education,* 11(1), 14–19.

Chen, H. C., & Graves, M. (1995). "Effects of previewing and providing background knowledge." *TESOL Quarterly*, 29, 663–682.

Clark, D., & Uhry, J. (1995). *Dyslexia Theory and Practice of Remedial Instruction* (2nd ed.). Baltimore, MD: York Press.

Condelli, L., & Wrigley, H. (2004). *Identifying Promising Literacy Interventions for Adult ESL Literacy Students: A Review of the Literature*. Washington DC: American Institutes for Research.

Condelli, L., Wrigley. H., Yoon. K., Cronen, S., & Seburn, M. (2003). *What Works Study for Adult ESL Literacy Students*. Washington DC: U.S. Department of Education (draft, not released).

Condelli, L., Bos, H., Tseng, F., Cronen, S., Altuna, J., Fishman, M., Farrell, M., and Gardiner, K. (2009). *The Adult ESL Literacy Impact Study: Design Report*. Washington DC: U.S. Department of Education.

Crandall, J. A., & Peyton, J. K. (eds.) (1993). *Approaches to Adult ESL Literacy Instruction.* Washington, DC & McHenry, IL: Center for Applied Linguistics and Delta Systems. (ERIC Document Reproduction Service No. ED 364 127.)

Diones, R., Spiegel, S., & Flugman, B. (1999). *Crossroads Café Impact Evaluation Study*. New York: City Univ. of New York. Center for Advanced Study in Education.

Gillingham, A., & Stillman, B. (1997). *The Gillingham Manual* (8th ed.). Cambridge: Educators Publishing Service.

Hartel, J., Lowry, B., & Hendon, W. (2006). *Sam and Pat* Boston: Thompson/Heinle.

Hopkins, K. H. (1998). *Educational and Psychological Measurement and Evaluation* (8th ed.). Boston, MA: Allyn and Bacon.

Kavenaugh, J. (1991). *The Language Continuum From Infancy to Literacy.* Baltimore: York Press.

The National Center for Family Literacy and The National Center for ESL Literacy Education. (2004). *Practitioner Toolkit: Working with Adult English Language Learners*. Louisville, KY and Washington, DC.

Rich, R., & Shepherd, M. J. (1993). "Teaching text comprehension strategies to adult poor readers." *Reading and Writing*, 5, 387–402.

Roberts, R., Cheek, E., & Mumm, R. (1994). "Group intervention and reading performance in a medium-security prison facility." *Journal of Offender Rehabilitation*, 20(3/4), 97–116.

Sabatini, J. P. & Bruce, K. (in press). Study and Reading Aid (SARA): Technical Guide. Princeton, NJ: Educational Testing Service.

St. Pierre, R., Swartz, J., Gamse, B., Murray, S., Deck, D, & Nickel, P. (1995). *National Evaluation of the Even Start Family Literacy Program: Final Report*. Cambridge, Massachusetts: Abt Associates Inc.

St. Pierre, R., Ricciuti, A., Tao, F., Creps, C., Swartz, J., Lee, W., Parsad, A., & Rimdzius, T. (2003). *Experimental Research Design Study of the Even Start Family Literacy Program: Final report*. Cambridge, Massachusetts: Abt Associates Inc.

Schochet, P. Z. (2007). *Guidelines for Multiple Testing in Experimental Evaluation of Education Intervention* Princeton: Mathematica Policy Research.

Sherow, S. M. (2006). *PA Literacy Corps: Applying Research to Practice: Teaching and Learning Strategies, Second Language Acquisition and English as a Second Language.* Pennsylvania State University.

Stemler, S. E. (2004). "A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability." *Practical Assessment, Research and Evaluation*, 9, 1–14.

Thissen, D., Steinberg, L., & Kuang, D. (2002). "Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons." *Journal of Educational and Behavioral Statistics*, 27 (1), 77–83.

Torgesen, J., et al. (2006). *National Assessment of Title I Interim Report—Volume II: Closing the Reading Gap: First Year Findings from A Randomized Trial of Four Reading Interventions for Striving Readers.* Retrieved from Institute of Education Sciences, U. S. Department of Education Web site: http://www.ed.gov/rschstat/eval/disadv/title1interimreport/index.html

U.S. Department of Education. (2010). Statistics retrieved from http://wdcrobcolp01.ed.gov/CFAPPS/OVAE/NRS on January 28, 2010.

U.S. Department of Education. (2007). Statistics retrieved from http://wdcrobcolp01.ed.gov/CFAPPS/OVAE/NRS on October 17, 2007.

Van de Craats, I., Kurveis, J., & Young-Scholten, M. (2006). *Research on Low-Educated Second Language and Literacy Acquisition Low-Educated Second Language and Literacy Acquisition.* Proceedings of the Inaugural Symposium.

Wilson, B., & Schupack, H. (1997). *Reading, Writing and Spelling: The Multisensory Structured Language Approach*. Baltimore: The International Dyslexia Association.

Woodcock, R., McGrew, K., & Mather, N. (2001). *Examiner's Manual: Woodcock-Johnson III Tests of Cognitive Ability.* Itasca, IL: Riverside Publishing.

Wrigley, H., Chisman, F., & Ewen, D. (1993). *Sparks of Excellence: Program Realities and Promising Practices in Adult ESL.* The Southport Institute for Policy Analysis. (ERIC Document Reproduction Service No. ED 373 586).

Wrigley, H., & Guth, G. (1992). *Bringing Literacy to Life: Issues and Options in Adult ESL Literacy.* San Mateo, CA: Aguirre International.

Wrigley, H., Richer, E., Martinson, K., Kubo, H., & Strawn, J. (2003). "The language of opportunity: Expanding employment prospects for adults with limited English skills." *CLASP Policy Brief*, 2.

APPENDIX A: ASSESSMENT SELECTION, ADMINISTRATION, AND SCORING

To select assessments for the final pre- and post-testing batteries, assessment staff experienced with testing the study population consulted with (1) members of our Technical Working Group (TWG) who had expertise in the content being assessed, and (2) the intervention developers, who were asked to identify the skills that the intervention was designed to improve. The following sections describe the process for selecting the final test batteries, the test administration preparation and methods, and the results of the scoring quality analysis for tests that were audio recorded and later rescored by expert scorers.

Assessment Selection

To determine which assessments to include in the final testing batteries, a pilot test was conducted on a comprehensive test battery proposed in consultation with the TWG and intervention developers:

- Study Aid and Reading Assistant (SARA) Letter Naming
- SARA Word Recognition
- SARA Spelling
- SARA Decoding
- Woodcock Johnson III (WJ) Spelling of Sounds
- WJ Reading Fluency
- WJ Oral Comprehension
- WJ Letter-Word Identification
- WJ Word Attack
- WJ Passage Comprehension
- WJ Picture Vocabulary
- Oral and Written Language Scales (OWLS)
- Receptive One-Word Picture Vocabulary Test (ROWPVT)

The pilot was conducted with volunteers from three sites. Two of the sites were among those participating in the full study later in the year. Those sites were chosen because they serve students with a range of language backgrounds, including Haitian Creole, Mandarin, and other Asian languages. An additional site was also included that serves primarily Spanish speakers.

The purpose of the pilot was to evaluate the feasibility of using the tests and the testing procedures with a low-literate ESL population (e.g., to determine whether learners could understand the test instructions and whether simplification would be needed). The pilot included testing the length of the battery components and

the battery in its entirety, as well as the instructions provided to assessors for the administration of each test. It also allowed the assessment team to evaluate how well the testing process was understood by students with limited exposure to standardized tests and whether they could understand the instructions. In addition, the pilot provided necessary data to allow study staff to check basic descriptive and psychometric properties of tests against expectations based on test technical manuals and to inform the decisions on which measures to include in the battery.

Students in literacy- or beginning-level ESL classes were invited to participate in the pilot by site staff who explained the study to students in their native languages and obtained their written consent on forms translated into their native languages. In the first pilot site, assessment staff conducted a pilot of the full assessment battery with 14 students (3 Mandarin speakers and 11 Cantonese speakers). Most students required some or all of the instructions to be translated into their native languages, and students in higher level ESL classes were asked to act as translators when necessary. In the second pilot site, 23 students participated in the pilot, including 7 Spanish speakers and 16 Haitian Creole speakers. The pilot team provided translations for the Spanish instructions, and ESL program teachers helped to translate test instructions into Haitian Creole as needed. Assessment staff also conducted additional piloting of the proposed battery with 10 Spanish speaking adults in a third site.

After piloting, data were delivered to study staff for analysis, along with recommendations from the assessment team. The recommendations included (1) simplifying the language of the test instructions, (2) providing instructions in the primary or native languages of the students in the study when needed, (3) allowing testers to use specific hand gestures along with spoken test instructions, (4) providing guidelines to help testers score students' responses given the various linguistic and pronunciation issues that were encountered, and (5) eliminating certain tests (such as the WJ Oral Comprehension and Reading Fluency tests) that most pilot students were unable to perform. Based on the pilot results and analysis, we selected a subset of the pilot battery of assessments for use in the pre-test and post-test data collection:

- SARA Letter Naming (SARALN; pre-test only)
- SARA Decoding (SARA Dec; post-test only)
- WJ Letter-Word Identification (WJID)
- WJ Word Attack (WJWA)
- WJ Passage Comprehension (WJPC)
- WJ Picture Vocabulary (WJPV; post-test only)
- Oral and Written Language Scales (OWLS) Listening
- Receptive One-Word Picture Vocabulary Test (ROWPVT)

Table A.1 presents the correlations between the assessments included in the post-test battery during the full study.

Table A.1: Correlations Between Post-Test Assessments (Full Sample)

	WJID Scale	WJWA Scale	SARA Dec	WJPC Scale	OWLS	ROWPVT	WJPV Scale
WJID Scale	1						
WJWA Scale	0.775	1					
SARA Dec	0.786	0.764	1				
WJPC Scale	0.709	0.683	0.622	1			
OWLS	0.507	0.444	0.417	0.656	1		
ROWPVT	0.524	0.456	0.473	0.620	0.682	1	
WJPV Scale	0.589	0.529	0.513	0.685	0.641	0.690	1
Sample Size: 1,137 students (587 *Sam and Pat* and 557 control).							

Note: Calculations used data for all students for whom post-test data were available.
Source: Adult ESL Literacy Impact Study student post-test assessments administered at the end of each term (fall 2008 and spring 2009).

Test Administration Preparation and Methods

During pilot testing, the standard instructions and administration protocols for many of the tests were found to be too complex for participating students. For example, in the Word Attack test, the standard instruction is "I want you to read some words that are not real words. Tell me how they sound. How does this word sound?" But testers learned that students at this level of fluency could not understand these instructions and could not discern real words from nonwords. Therefore, study staff developed a set of shorter, simplified instructions for the full study. For example, the aforementioned instruction was shortened to "Read this word to me." Similarly, phrases such as "put your finger on" were changed to "point to" so that students could recognize this simple command across all tests. The simplified instructions for each test were reviewed by assessment experts on the study team and were translated into Spanish, simplified Chinese characters, Haitian Creole, and Armenian.

For the full study, all tests were administered via computer-assisted personal interviewing technology. Testers used easels to present test items to students and recorded students' responses in the computer as correct, incorrect, or not sure. The laptop computers automatically recorded audio for each test session. To facilitate the flow of the SARA Letter Naming test, testers scored students' responses on paper and then entered scores (correct or incorrect) into the computer at the end of the testing session. Testers were instructed to use English instructions first. If the student had trouble understanding what to do, the tester was to repeat the instruction in English (using more hand gestures and speaking more slowly) and

then use the approved translation if necessary. For students who did not speak one of these languages, instructions were given in English and, when available, program staff acted as translators. Once a tester had started using the translated instructions, he or she was to continue using that translation until the start of the next assessment, at which point the tester was to try the English instructions again.

Test Administrator Recruitment and Training

Test administrators were hired locally, with preference given to those with experience administering assessments and speaking at least one language of the students at the site. Testers were required to be highly proficient in spoken and written English. In addition, all testers and field supervisors were required to pass a proficiency test in administration and scoring of the assessments before they were approved to work in the field. In total, 48 testers and 6 field supervisors were hired. Study staff conducted 4-day workshops in each site to train field interviewers and supervisors in general data collection techniques, administration of the assessments, proper scoring of students' responses (including linguistically sensitive scoring), and use of the Blaise computer-assisted interviewing (CAI) software and other computer technology.

Two key aspects emphasized in the training were (1) sensitivity to the linguistic backgrounds of the participating students and (2) the importance of following objective guidelines for scoring their spoken responses. This approach emphasized that students should not be penalized for nonstandard English pronunciations when these differing pronunciations are related to sound formations specific to their native languages. Testers were given guidelines developed by linguists on staff that detailed standard American English dialect pronunciations for each test item in the SARA Letter Naming, WJ Letter Word Identification, WJ Word Attack, and SARA Decoding test, along with correct and incorrect alternative pronunciations. Separate versions of the guidelines were developed for Spanish, Haitian Creole, Mandarin Chinese, and Armenian/Persian—taking into account the linguistic characteristics and variations of these languages and the types of pronunciation errors likely to occur among speakers of each language.

Because of differences in the languages, the names of letters and pronunciations of some words that were correct when given by speakers of one language might be incorrect for speakers of another language. Study staff reviewed the guidelines with testers in detail during the training, and testers practiced scoring the tests while listening to audio recordings of actual pilot sessions. Afterward, testers discussed the scores they gave each item and evaluated them against the guidelines to better understand the principles behind correct and incorrect responses. Three tests—WJ Letter Word Identification, WJ Word Attack, and

SARA Decoding—required an additional response category for when the tester thought the response was correct but was not entirely sure.

Although the scoring guidelines gave examples of correct and incorrect pronunciations based on the most likely errors, they could not cover all possible pronunciations. Therefore, testers were encouraged to use "not sure" as their score when they could not decide whether a student's response was acceptable. The computer program used to administer the tests, which automatically calculated test ceilings and routed testers to the next test when ceilings were reached, was programmed to count a "not sure" response similarly to a correct response. This allowed for later analysis of the audio recordings by expert scorers and ensured that tests were not ended early based on a "not sure" score that should perhaps have been scored as correct.

Audio recordings of all testing sessions were submitted to the supervisory study staff members for review during each week of data collection. Study staff noted any errors in test administration and provided feedback directly to team leaders so that they could observe testers and correct errors in a timely manner. Additionally, at the end of testing in each site, data files with audio recordings were sent to expert scorers for a more thorough review. The expert scorers reviewed and rescored a sample of 10 percent of the cases for each tester and provided the assessment team with detailed feedback on errors of administration and scoring reliability reports for each tester. The results of these reviews were conveyed to field staff as soon as possible and were also used to inform refresher trainings held throughout the study period.

During Cohort 1—after pre-test data collection but before post-testing—supervisory study staff members led refresher trainings with each site team. Conducted via telephone, these trainings focused on errors in administration noted during supervisors' review of the audio recordings and on changes to the computer program and protocols that had been made to prepare for the post-test phase (which involved a slightly different test battery and additional collection of background data). After the post-test data collection for Cohort 1, testers received a refresher memo with additional feedback on test administration and a short memo on scoring protocols. These materials were mailed to testers, and team leaders reviewed materials with testers before the start of data collection for Cohort 2. Finally, after pre-test data collection for Cohort 2 but before post-testing, field staff were required to participate in trainings led by field team leaders. The field staff trainings focused primarily on the scoring guidelines. Study staff sent each team a CD with audio recordings of sessions from Cohort 1 and asked the teams to review the scoring guidelines again and to practice scoring the recorded cases as a group. This exercise focused on WJ Word Attack and SARA Decoding—the two tests for which reliability scores were lowest (Table A.2). These final trainings were held before the final testing period.

Data Collection for the Full Study

Pre-test data collection was scheduled as close to the beginning of the term as possible. Testing began in the first week of classes or early in the second week, depending on the speed of the intake process and the scheduling of testers who had to work in multiple locations at the same site. Post-test data collection was scheduled to begin during the 12th week of classes. Test administrators assessed most students during class hours, by giving the teacher an ordered list of students to send out for testing. Because the administrators went into the classrooms to interact with the teacher, it is possible that they noticed differences in instructional materials and therefore were not blind to students' condition. The administrators were not, however, informed of students' assigned groups.

Pre-testing

Pre-test data collection was sequenced with sample intake. Once students were enrolled into the study and randomly assigned to a class, random assignment staff submitted the students' intake forms to the assessment team leaders at each site, who assigned each student to a tester for the pre-test assessments. Pre-test tests were conducted individually, with a tester and a single student, during class time. Students were called out of class to participate in the assessments. The tests took approximately 45 minutes to complete. Students who were absent from class during the pre-test window were contacted by phone and asked to return to the school for testing. If they could not be scheduled to test in the school, field staff went to students' homes to complete the testing. Across the two cohorts, 37 students were tested outside class.

Post-testing

At the end of the term, students were again tested primarily during class time, and students who were not present in their assigned classes during the post-test window were invited to return for testing. Aside from administering the post-test assessments, field staff collected any data that was missing from students' intake forms before instruction. The post-test battery took approximately 1 hour to complete. Students in the sample were asked to complete post-testing even if they had not completed a pre-test test or had left their assigned study class since pre-test.

For both cohorts, sites were post-tested in roughly the same order they were pre-tested, which resulted in approximately the same number of weeks between the pre-test and post-testing across sites. All students received a $40 gift card after completing the post-test assessments.

Post-testing for Students not Present in Assigned Classes During the Post-test Window

After pre-test but before post-test, team leaders visited schools 6 and 9 weeks into the term to obtain lists of students who had been absent from their assigned study class for 2 weeks or more. Team leaders made phone calls to these students to invite them to come to the school and be tested during post-test data collection. As with the testing at the beginning of the term, if students who were not currently attending their study class could not be scheduled to test in the school, field staff went to their homes to complete the testing. Across the two cohorts, 52 students were post-tested outside class. Reasons for students not attending the assigned study class during the testing window could not always be determined, but included both factors like illness and the student taking a break in enrollment from the program.

Scoring Quality

A random sample of 10 percent of the audio recordings for each tester at each site (minimum of 10 test sets per tester) was rescored by scoring experts for WJID, WJWA, SARA Letter Naming, and SARA Decoding assessments. The findings are reported in Table A.2 and show that percent agreement on items scored ranged from 78 to 92 percent for pre-tests and 73 to 88 percent for post-tests.

Table A.2: Percent Agreement on Item Scoring Between Testers and Expert Scorers, by Pre- and Post-Test Assessment

	Cohort One			Cohort Two		
Test Name	Number of Students	Number of Items	Percent Agreement	Number of Students	Number of Items	Percent Agreement
Pre-tests						
WJWA	121	1,797	78.0	89	1,708	79.0
WJID	121	4,425	88.4	89	3,785	85.6
SARALN	121	3,233	90.9	88	2,396	91.8
Post-tests						
WJWA	114	2,006	77.7	82	1,622	76.4
WJWID	114	4,777	88.4	82	3,872	85.4
SARA Dec	114	2,086	73.2	82	1,779	73.7

Note: Calculations used data for all students for whom pre- or post-test data were available.
Source: Adult ESL Literacy Impact Study student pre- and post-test assessments administered at the beginning and end of each term (fall 2008 and spring 2009).

If the tester was responsible for testing more than one language group, expert scorers rescored 10 percent of each subgroup. Of note here is that special consideration was taken to ensure the expert scoring group had its own interrater

reliability. When scores differed from the original score by more than 3 points, the tests were re-reviewed by a new tester. The final interrater reliability among expert scorers was confirmed to be greater than 90 percent across major language groups, ensuring that the expert recommendations were consistent.

Reliabilities of the Post-Tests During the Study

The post-tests were administered to two cohorts of study participants, for a total of 1,137 examinees. Internal consistency reliability estimates were calculated for the raw scores of each post-test using the Kuder-Richardson formula 20 (K-R20) in Stata.

Table A.3 summarizes the reliabilities found for the seven post-tests, which ranged from 0.809 to 0.965.

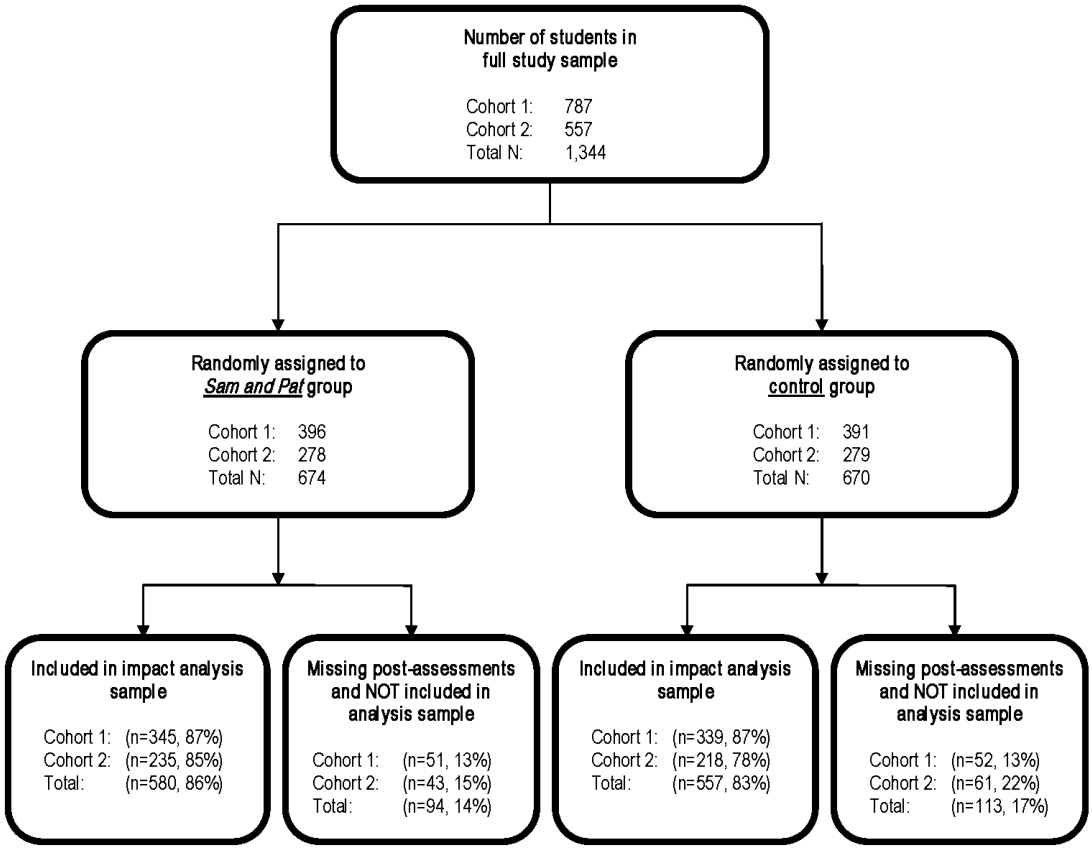
Table A.3: Post-Test Reliability Estimates

Post-Test Name	K-R20
Woodcock Johnson Letter Word Identification	0.965
Woodcock Johnson Word Attack	0.932
SARA Decoding	0.957
Woodcock Johnson Passage Comprehension	0.833
OWLS	0.937
ROWPVT	0.963
Woodcock Johnson Picture Vocabulary	0.809
Sample Size: 1,137 students (580 *Sam and Pat;* 557 control).	

Source: Adult ESL Literacy Impact Study student post-test assessments administered at the end of each term (fall 2008 and spring 2009).

APPENDIX B: SUPPLEMENTAL TABLES AND FIGURES FOR CHAPTER 2

Figure B.1: Flow of Students From Random Assignment to Analysis



APPENDIX C: CLASSROOM OBSERVATION METHODS AND INSTRUMENT

To document the implementation of *Sam and Pat* materials and related instructional practices and to describe instruction across both *Sam and Pat* and control classes during the study, members of the evaluation team conducted structured classroom observations at approximately 6 weeks into the beginning of each term.[28] The following sections describe the observation training and methods used and provide an overview of the data quality control procedures and analysis.

Observation Training and Methods

Training

The observation team consisted of six staff on the evaluation team. All observers had experience conducting structured observations of instruction, and all but one observer had experience observing adult ESL instruction. Observers received the following training:

- A day and a half of training on the content captured by the observation guide (included at the end of this appendix). This training included reviewing the glossary of instructional codes to be used, watching training videos that exemplified key practices, and practicing coding of video segments with feedback.
- Two paired practice observations in local adult ESL reading classes with group and individual trainer feedback before going into the field for the full study.
- A 2-hour retraining and individual feedback after the first study observation.

Scheduling and Preparing for the Observation

In the month before each term, the observation team prepared and sent out letters with observation schedules to the teachers. For sites that required more than one observer, we also varied which observer was assigned to each class across terms and balanced the number of *Sam and Pat* and control classes observed by each observer. It should be noted that observers were not blind to group.

[28] There was one missing observation per term.

Conducting the Observation

Observers were instructed to remind teachers when they arrived that they were conducting "naturalistic observations," meaning that the teacher should carry on as usual. If the teacher wished to explain the observer's presence to the students, he or she was asked to describe the observer as a visitor.

Observers were trained to take up a position in the class where they could see and hear the instruction clearly. If it was necessary to move around the classroom in order to code accurately, they were instructed to do so as discretely as possible.

The key section of the observation guide was Section A, where instructional practices and materials were coded in 5-minute intervals. During each interval, all instruction that occurred for at least 30 seconds was accounted for by circling any of the instructional codes that described the activity or activities occurring in that interval. (Note that the "Other" code was an exception, and was only used when no codable instruction occurred for the full interval.) If an activity rolled into the next interval, observers continued coding it for as many intervals as it lasted. Multiple codes could be used as appropriate within a coding category (e.g., L2 Phonics) or across multiple coding categories (e.g., L2 Phonics and L3 Writing and Spelling for Phonics Reinforcement). Observers were instructed to select all codes for which the activity(ies) met the definitions for those codes. Similarly, all materials used for the activities coded were also documented in Section A of the guide.

Data Quality Control Procedures and Analysis

Each term, approximately 10 percent of the observations were conducted by two staff independently so that interrater reliability (IRR) could be determined. To determine IRR, we excluded from the calculations for each interval all sections of the observation protocol in which *both observers agreed that instruction did not take place during the interval*. For example, if both observers agreed that phonics instruction did not take place during the observed interval, then all cells in the observation protocol related to phonics were excluded from the IRR calculation in that interval. In other words, empty cells were included only for areas of instruction that at least one observer had coded as having taken place. This approach reduces the number of empty cells used in the IRR calculation, but the interpretation of the results is similar to that of a less conservative approach: percentage of agreement on observed and unobserved instruction in instructional areas that took place according to at least one observer. Once we determined which cells to include in the IRR calculations, we used percent agreement as the

measure of reliability.[29] As shown in Table C.1, percent agreement between pairs of observers ranged from 0.86 to 0.95 in fall and 0.90 to 0.98 in spring.

Table C.1: Average Percent Agreement Among Observers for Fall and Spring Terms

	Fall	Spring
Instructional Area	0.953	0.950
Specific Instructional Practices within Instructional Areas	0.861	0.896
Materials	0.932	0.936
Grouping	0.938	0.975

Source: Adult ESL Literacy Impact Study classroom observation protocol.

[29] Percent agreement provides an easily interpretable measure of reliability for dichotomous items of the kind included in the observation protocol. For a discussion of other types of measures and the rationale for choosing among them, see Stemler, 2004; and Hopkins, 1998.

Adult ESL Literacy Impact Study

Classroom Observation and Coding Guide

Program:
Site:
Teacher:
Name of Class:
Class Meeting Time:
Name of Observer:
Observational Date, and Start and End Time:

Adult ESL Literacy Impact Study
Classroom Observation and Coding Guide

Summary and Overview of Class

1. Number of students: At beginning of class _____ At end of class _____

2. Number of instructional aides present: _____

3. Check any reading resources available and technology used in the classroom:

   a. Reading

   - ☐ Alphabet is displayed
   - ☐ Sound-symbol correspondence cards are displayed
   - ☐ Sight words/word wall is displayed
   - ☐ Other reading materials (describe:)

   b. Technology Used

   - ☐ Overhead projector
   - ☐ LCD Projector
   - ☐ Screen
   - ☐ Laptop
   - ☐ Desktop computer
   - ☐ Television
   - ☐ VHS/DVD Player
   - ☐ CD Player/listening station
   - ☐ Other

4. Number of hours per week students visit a computer lab as part of instruction (if zero, write "0"): _____

5. Name of software used with class (list all):

6. _____

Notes:

PART A.1 – Literacy Development

Interval _____ Time: ____ : ____ AM PM

L.1. Pre-literacy

The Teacher and Students…(circle all that apply)

Engage in activities that include

T/S1	Learning about print directionality, shapes and symbols, word boundaries, how to use a writing utensil, or how to form symbols, letters or numbers
T/S2	Developing phonemic awareness (no print)
T/S3	Recognizing individual letters and working with the names of letters in English (not phonics instruction)*+
T/S4	Working with upper vs. lower case letters of the alphabet*
T/S5	Working with the alphabet in sequence+
T/S6	Recognizing numbers in print+

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. Other: _____

L.2. Phonics

The Teacher…(circle all that apply)

T1	Explains, describes, or demonstrates sound-symbol pattern or decoding rule (e.g., leading students in using letter or key word cards to pair letter or letter combinations with the sounds they make)*
T2	Uses multi-sensory approaches (hand movements, finger tapping, blank cards, checkers, etc.) to emphasize phonemic correspondences (i.e., segmenting phonemes)*

The Students…(circle all that apply)

S1	Practice sound-symbol correspondence either independently or guided by teacher (e.g., "listen and repeat"; commonly includes the use of letter cards, key word cards, or magnetic/adhesive letters to form words)*

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. Wilson letter cards
5. *Sam and Pat* key word (sound/symbol) cards*
6. Other: _____

L.3. Writing and Spelling for Phonics Reinforcement

The Teacher and Students…(circle all that apply)

Engage in activities that include

T/S1	Matching/labeling pictures with phonetically regular words (e.g., fill in phonetic word grids)*
T/S2	Write letter(s) that represent a phoneme (e.g., "sh"; fill in the blank, copy, etc.)*
T/S3	Circling the phonetically regular word*
T/S4	Taking dictation of phonetically regular words: students write down key words (e.g., short /a/ sound words like hat, cat, mat) called out by student or teacher*
T/S5	Oral spelling of phonetically regular words*
T/S6	Copying/writing phonetically regular words*

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. Wilson letter cards
5. *Sam and Pat* key word (sound/symbol) cards*
6. *Sam and Pat* phonetic word grids*
7. Other: _____

Notes: Practices marked with '*' represent practices that may be seen during *Sam and Pat* instruction. Practices marked with '+' represent practices found to be common of adult ESL literacy classes in earlier studies (Condelli et al. 2003).

L.4. Learning Vocabulary to Reinforce Reading Instruction

The Teacher…(circle all that apply)

T1	Introduces a small number (8 or fewer) of vocabulary words or reviews old vocabulary words related to the class readings*
T2	Introduces a large number (9 or more) of vocabulary words or reviews old vocabulary words related to the class readings
T3	Engages in interactive process with students to figure out the meaning of words
T4	Associates new words with other words whose meanings students already know
T5	Writes words on board, reads aloud, students repeat*
T6	Dictates vocabulary words to students*

The Students…(circle all that apply)

S1	Air-write or trace words with their finger while spelling out loud*
S2	Match vocabulary words (orally or physically) to pictures or realia*+
S3	Label pictures (in writing) with vocabulary words*
S4	Sort cards with vocabulary words or pictures into topics*
S5	Write vocabulary words on flash cards or in notebooks (may or may not be dictation)*
S6	Do a cloze exercise to fill in new vocabulary*
S7	Students give the meaning of words (orally, pictorially, etc.)

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. Wilson letter cards
5. *Sam and Pat* key word (sound/symbol) cards*
6. Other: _____

L.5. Fluency and Accuracy in Reading (Note: Applies to reading text, practicing word lists related to phonics lessons should be coded under L.3.)

The Teacher and Students…(circle all that apply)

T1	Reads text aloud to students before having them read*
T2	Explicitly models expressive reading

The Students…(circle all that apply)

S1	Read text aloud, listen to others and read along, or take turns reading*+
S2	Repeatedly read same/familiar text (from board, text, or own writing; can be used with other fluency codes)+
S3	Listen to readings in recorded form and/or read aloud with a tape (not conversation practice; related to reading instruction)
S4	Practice reading parts of sentences (e.g., read first part of sentence, then last part, then blend together)*
S5	Read silently/quietly+
S6	Practice reading for intonation/expression in response to explicit instruction or teacher demonstration+
S7	Scan text to identify familiar words in print (in response to explicit instruction or teacher demonstration)
S8	Follow along during reading by tracing under the words with an eraser or finger (use with other fluency codes)*

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. *Sam and Pat* key word (sound/symbol) cards*
5. Other: _____

L.6. Reading Comprehension

The Teacher…(circle all that apply)

T1	Previews the text and/or pictures BEFORE reading*
T2	Interacts with students to elicit storyline and/or understanding of new words in readings BEFORE reading (e.g., Q&A; do not code this as E2)*
T3	Activates or builds students background knowledge related to the reading (e.g., relates story to students' experiences or provides additional information regarding the text)*
T4	Asks questions relevant to the text DURING reading*
T5	Asks students direct recall questions (i.e., the answer can be found in the text) AFTER reading*+
T6	Asks students inferential questions AFTER reading*+

The Students…(circle all that apply)

S1	Preview the text and/or pictures BEFORE reading guided by or independent of teacher*
S2	Make predictions about aspects of the story (based on title, pictures, etc.) or predict the ending of sentences or readings DURING reading*
S3	Sequence pictures, words, or sentence strips to tell a story
S4	Match sentences from the reading to pictures*
S5	Act out a story* (can be used in combination with ESL T/S1 if it includes dialogue practice)
S6	Retell a narrative or sequence of events
S7	Write a summary of events
S8	Respond to questions about the story DURING or AFTER reading (orally, nonverbally (e.g., yes/no cards), or in writing)+
S9	Skim text to find information DURING or AFTER reading (in response to explicit instruction or teacher demonstration)
S10	Identify/discuss the key concepts or general meaning (topic or function) of a text DURING or AFTER reading

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. *Sam and Pat* key word (sound/symbol) cards*
5. Other: _____

L.7. Writing that is Unrelated to Reading Activities (Note: Writing activities related to reading activities are coded using L.2. through L.6.)

The Teacher and Students…(circle all that apply)

Engage in activities that include

T/S1	Writing subskills (practicing punctuation, capitalization, standard spelling, etc.)
T/S2	Writing practice (copying own writing or other text)
T/S3	Guided composition (filling in blanks, sequencing, editing, responding to writing prompts)
T/S4	Free writing (journal, poems, etc.)

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. Other: _____

PART A.2 – ESL Acquisition

Interval _____ **Time:** ____ : ____ **AM PM**

E.1. Oral Communication Skills—Listening (Note: Only code E.1. when students are explicitly instructed or motioned to "listen".)

The Teacher and Students…(circle all that apply)	
Engage in activities that include	
T/S1	Listening and repeating words, sentences, phrases, or dialogues+
T/S2	Listening to how English words are pronounced (must include teacher explicitly instructing students to listen for pronunciation)
T/S3	Listening and responding nonverbally (e.g., TPR, Bingo games, point to pictures or items)

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. Other: _____

E.2. Oral Communication Skills—Speaking

The Teacher and Students…(circle all that apply)	
Engage in activities that include	
T/S1	Practicing communication skills with structured language (e.g., repetition of phrases, dialogue practice)+
T/S2	Practicing communication with guided structure (some open-ended phrases)+
T/S3	Practicing open-ended or spontaneous communication (conversation, discussion)
T/S4	Practicing the pronunciation of English words (not general speaking practice, which is coded above; must include teacher explicitly instructing students to focus on pronunciation)+
T/S5	Practicing stress, tone and rhythm in response to explicit instruction or teacher demonstration

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. Other: _____

E.3. Grammar: Understanding How English Works

The Teacher and Students…(circle all that apply)	
Engage in activities that include	
T/S1	Doing oral practice with grammar or oral spelling
T/S2	Hearing explanations of grammar or verbalizing grammar rules+
T/S3	Writing, matching, or identifying sentences based on specific grammar patterns (includes filling in blanks in grammar worksheets; NOT oral-see T/S1 for oral code)+
T/S4	Editing/correcting sentences focusing on grammar
T/S5	Studying word parts (prefixes, suffices, endings, etc.)
T/S6	Studying parts of speech (verbs, nouns, adjectives)+
T/S7	Using problem solving to discover rules and patterns (e.g., "task-based" grammar)+

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. Other: _____

E.4. English Vocabulary and Idioms

The Teacher and Students…(circle all that apply)	
Engage in activities that include	
T/S1	Words unrelated in meaning or context (e.g., names of random objects)
T/S2	Words that arise out of a particular context+
T/S3	Words that are related (decide; decision; decisive)
T/S4	Idioms

Grouping
W Whole class
S Small Group
P Pair
I Individual

Materials
1. *Sam and Pat* workbook or worksheets*
2. Other commercial text or worksheets: Specify text name, if applicable: _____
3. Blackboard/whiteboard prompts
4. Other: _____

E.5. Socio-Cultural Knowledge

		Grouping	Materials
T/S1	The teacher and students engage in activities that include a focus on socio-cultural knowledge, such as:	**W** Whole class **S** Small Group **P** Pair **I** Individual	1. *Sam and Pat* workbook or worksheets* 2. Other commercial text or worksheets: Specify text name, if applicable: _____ 3. Blackboard/whiteboard prompts 4. Other: _____
	• Cultural facts		
	• Life skills (e.g., how to find the post office, how to navigate the welfare system, identifying community resources like libraries or police)		
	• Rights and responsibilities as a citizen (civics)		
	• Social appropriateness in language and communication		
	• Cross-cultural comparisons		

PART A.3 – Functional Reading, Writing, and Math

F. Functional Reading, Writing, and Math

		Grouping	Materials
The Teacher and Students…(circle all that apply)		**W** Whole class **S** Small Group **P** Pair **I** Individual	5. *Sam and Pat* workbook or worksheets* 6. Other commercial text or worksheets: Specify text name, if applicable: _____ 7. Blackboard/whiteboard prompts 8. Other: _____
Engage in activities that include			
T/S1	Text based functional literacy (working with forms, labels, flyers, lists, messages, etc.)		
T/S2	Alphabet based functional literacy (working with phone book, dictionary)		
T/S3	Graphic literacy (working with maps, graphs, signs, etc.)		
T/S4	Numbers and math (working with money, quantities, dates, time, types of numbers)		

PART A.4 – Other Instruction and Breaks

O.1. Other

T/S1	The teacher and students have a break in codable ESL/Literacy instruction lasting the full interval, such as:
	• A transition between activities that lasts for the full interval
	• A break or class disturbance that lasts the full interval
	• Participation in activities not coded under other codes and that last the full interval (specify):

PART B – Instructional Strategies

I.1. Links What is Learned to the Outside World (No 30 Second Rule)

T/S1	The teacher links what is learned to life outside of the classroom (e.g., points out that they will fill out similar forms when looking for a job), or brings "outside" into the classroom through the use of real life items.

I.2. Use of Students' Native Language (No 30 Second Rule)

The Teacher…(circle all that apply)		**The Students…(circle all that apply)**	
T1	Gives/clarifies instructions in students' native language	**S1**	Ask questions of teacher or other students in native language
T2	Translates individual words, idioms or phrases in native language to English (e.g., manzana = apple; Necesito ayuda = I need help)	**S2**	Translate individual words or phrases in native language to English (e.g., manzana = apple; Necesito ayuda = I need help)
T3	Translates connected text into students' native language (individual sentences to complete passages)	**S3**	Provide translation (to other students) of connected text (individual sentences to complete passages)
T4	Explicitly instructs students to practice dialogues or hold discussions in their native language	**S4**	Practice dialogues or hold group discussions in native language

APPENDIX D:
POWER CALCULATIONS AND IMPACT ESTIMATION METHODS

Power Analyses

We examined the statistical power for the main impact analyses. When we calculated the minimum detectable effect size (MDES) using the two-level model described in Chapter 4 and study sample sizes, the resulting MDES was 0.158, indicating that the study was adequately powered to detect the magnitude of impacts that the study was originally designed to detect and that was considered by the study staff to be meaningful (MDES = 0.16).

Missing Data Approach

Covariate and pre-test data

Missing data on the covariates and pre-tests were accounted for in the analyses by using a dummy variable correction: missing variables were coded as zero and were identified as missing using a dummy variable that was included as a covariate in the impact equation.

In subgroup analyses, however, we eliminated any observations with missing data on the variable defining the subgroup breakdown.

Post-test data

Missing post-test data is potentially more damaging than missing pre-test data because the measured outcomes of students may not properly represent the outcomes of their non-responding counterparts in the same research group. If that is the case, and the nonresponse patterns differ between research groups then the impacts estimated on students may be biased. However, in the current study, we achieved a response rate of 85 percent on the post-tests, and response rates were statistically equivalent between the *Sam and Pat* and control groups (see Table 2.3, Chapter 2). In addition, although some background characteristics were predictive of missing data on all post-tests—being female, Asian, Hispanic, or African American—and an additional characteristic (years in the U.S.) predicted missing WJ Word Attack post-test data, these cases met the criteria for being missing at random. We conducted probit analyses for each outcome, and found that after controlling for student characteristics (e.g., gender, ethnicity, years in the U.S., etc.), the probability of having missing post-test data was the same regardless of group assignment (Table D.1). Students with missing post-test data were dropped from the impact analysis sample.

Table D.1: Predictors of Missing Post-Test (Probit Analysis), by Post-Test

Post-test and Missing Predictors	Coef.	Std. Err.	z	P>|z|
Woodcock Johnson Letter Word ID				
Treatment or control status	-0.114	0.091	-1.260	0.208
Student Age	-0.004	0.004	-1.083	0.279
Female Indicator (Student)	-0.363	0.097	-3.728	0.000*
Asian Indicator (Student)	0.705	0.297	2.373	0.018*
African American Indicator (Student)	0.759	0.243	3.124	0.002*
Hispanic Indicator (Student)	0.837	0.237	3.527	0.000*
Years in School (Student)	0.008	0.011	0.702	0.483
Years in US (Student)	-0.017	0.010	-1.795	0.073
Woodcock Johnson Word Attack				
Treatment or control status	-0.111	0.091	-1.224	0.221
Student Age	-0.004	0.004	-0.997	0.319
Female Indicator (Student)	-0.369	0.098	-3.784	0.000*
Asian Indicator (Student)	0.701	0.297	2.360	0.018*
African American Indicator (Student)	0.737	0.243	3.033	0.002*
Hispanic Indicator (Student)	0.857	0.236	3.634	0.000*
Years in School (Student)	0.005	0.011	0.462	0.644
Years in US (Student)	-0.020	0.010	-2.066	0.039*
SARA Decoding				
Treatment or control status	-0.114	0.091	-1.260	0.208
Student Age	-0.004	0.004	-1.083	0.279
Female Indicator (Student)	-0.363	0.097	-3.728	0.000*
Asian Indicator (Student)	0.705	0.297	2.373	0.018*
African American Indicator (Student)	0.759	0.243	3.124	0.002*
Hispanic Indicator (Student)	0.837	0.237	3.527	0.000*
Years in School (Student)	0.008	0.011	0.702	0.483
Years in US (Student)	-0.017	0.010	-1.795	0.073
Woodcock Johnson Picture Vocabulary				
Treatment or control status	-0.116	0.091	-1.283	0.199
Student Age	-0.004	0.004	-1.030	0.303
Female Indicator (Student)	-0.363	0.097	-3.731	0.000*
Asian Indicator (Student)	0.764	0.295	2.587	0.010*
African American Indicator (Student)	0.755	0.242	3.116	0.002*
Hispanic Indicator (Student)	0.869	0.237	3.659	0.000*
Years in School (Student)	0.004	0.011	0.407	0.684
Years in US (Student)	-0.015	0.010	-1.566	0.117
OWLS				
Treatment or control status	-0.116	0.091	-1.280	0.200
Student Age	-0.004	0.004	-1.086	0.278
Female Indicator (Student)	-0.365	0.097	-3.748	0.000*
Asian Indicator (Student)	0.765	0.295	2.589	0.010*
African American Indicator (Student)	0.743	0.243	3.062	0.002*
Hispanic Indicator (Student)	0.843	0.237	3.564	0.000*
Years in School (Student)	0.003	0.011	0.296	0.767
Years in US (Student)	-0.016	0.010	-1.647	0.100

Table continued, next page.

Table D.1: Predictors of Missing Post-Test (Probit Analysis), by Post-Test (Continued)

Post-test and Missing Predictors	Coef.	Std. Err.	z	P>|z|
ROWPVT				
Treatment or control status	-0.117	0.091	-1.291	0.197
Student Age	-0.004	0.004	-1.047	0.295
Female Indicator (Student)	-0.364	0.097	-3.743	0.000*
Asian Indicator (Student)	0.761	0.295	2.577	0.010*
African American Indicator (Student)	0.755	0.242	3.115	0.002*
Hispanic Indicator (Student)	0.868	0.237	3.657	0.000*
Years in School (Student)	0.004	0.011	0.404	0.686
Years in US (Student)	-0.015	0.010	-1.558	0.060
Sample Size: 1,344 (675 *Sam and Pat*; 670 Control)				

*Indicates variable is a significant predictor of missing post-test data after controlling for other student characteristics, based on probit ($\chi 2$) analysis.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the end of each term (fall 2008 and spring 2009).

Estimation Model

Reading and English language impacts were calculated by comparing mean *Sam and Pat* and control group scores for each assessment. The model used to estimate those impacts is described below:

$$Y_I = \overline{Y_T} - \overline{Y_C}$$

where
Y_I = impact for outcome Y,
$\overline{Y}_T$ = mean outcome Y for the treatment group, and
$\overline{Y}_C$ = mean outcome Y for the control group

The model can also be expressed as a regression model:

$$Y_i = b_0 + b_1 E_i + \varepsilon_i,$$

where
Y_i = outcome Y for student i
E_i = 1 if student i is part of the *Sam and Pat* group, and
= 0 if student i is part of the control group.
b_0 = intercept
b_1 = coefficient associated with being in the *Sam and Pat* group
ε_I = random error term for student i

To increase the precision of the impact estimates, the student- and teacher-level covariates listed below were included in the model.

I. Student covariates
   a. Assessment pre-test scores
   b. Age
   c. Female indicator
   d. Asian indicator
   e. African American indicator
   f. Hispanic indicator
   g. Years in school
   h. Years in the United States

II. Teacher covariates
   a. Female indicator
   b. Asian indicator
   c. African American indicator
   d. Hispanic indicator
   e. Certified to teach ESL indicator

III. Site indicator variables

This transformed the model to a two-level regression model:

$$Y_{ij} = b_0 + b_1 E_i + \sum_{k=1}^{K} c_k X_{ijk} + \sum_{n=1}^{N} f_n Q_{jn} + \varepsilon_{ij} + \gamma_j \qquad (1)$$

where
Y_{ij} = outcome Y for student i, taught by teacher j
E_i = 1 if student i is part of the *Sam and Pat* group, and
= 0 if student i is part of the control group.
b_0 = intercept
b_1 = coefficient associated with being in the *Sam and Pat* group
c_k = vector of coefficients associated with K baseline student covariates
X_{ijk} = vector of K baseline covariates for student i
f_n = vector of coefficients associated with N baseline teacher covariates
Q_{jn} = vector of N baseline covariates for teacher j
ε_{ij} = random error term for student i
γ_j = random error term for teacher j.

In summary, we estimated a two-level regression model, where the first level was the student and the second level was the teacher.[30] In each regression equation, the dependent variable was the post-test result for each assessment and the independent variables included a *Sam and Pat*-control group dummy variable, the pre-test result for that assessment, student-level covariates and teacher-level covariates.

The statistical significance of the coefficients in this model were assessed using a two-tailed t-test. If this test determined that an impact (as expressed by coefficient b_1) has a less than 5 percent chance of being zero or having a different sign (being negative when the point estimate is positive or vice versa), the impact was considered statistically significant (subject to multiple comparison adjustments).

Effect sizes for impacts were calculated by dividing the unadjusted impact by the pooled standard deviation at each site. The pooled standard deviation is a weighted average of the control and *Sam and Pat* group standard deviations. Effect sizes for pre- to post-test gains were calculated by dividing the unadjusted overall gain by the pooled standard deviation (0.5*pre-test s.d. + 0.5*post-test s.d.).

In addition to calculating overall treatment impacts, we performed subgroup analyses by running the impact analysis described by equation (1) on the following subgroups:

- Non-Roman-based Alphabet Background
- Spanish Native Language Speakers
- Students with Lower/Higher Literacy Levels at the beginning of the term
- Students from Cohort 1/2

An attendance service contrast was also estimated as described in equation (1), with total hours of attendance treated as the outcome.

Instructional service contrasts were estimated using a one-level model at the teacher level that included site indicators as covariates:

$$Y_j = d_0 + d_1 E_j + \sum_{m=1}^{M} f_m Z_{jm} + v_j, \qquad (2)$$

[30] Clustering due to the randomly assigned pods was accounted for by applying the Huber-White correction to the standard errors.

where
Y_j = instructional outcome for teacher *j*
E_j = 1 if teacher *j* is part of the *Sam and Pat* group, and

= 0 if teacher *j* is part of the control group.
d_0 = intercept
d_1 = coefficient associated with being in the *Sam and Pat* group
d_m = vector of coefficients associated with M site indicator covariates
Z_{jm} = vector of M site indicator covariates for teacher *j*
v_j = random error term for teacher *j*

The statistical significance of the coefficients in this model was assessed as described above.

Adjusting for Multiple Comparisons

There is a risk of spurious findings in large-scale evaluation studies because these studies often include a large number of independent hypothesis tests, each of which has a small chance of producing a statistically significant result when there is no real impact. Following recommendations by Schochet (2007), we minimized this risk and used appropriate statistical corrections to mitigate it.

To minimize the risk posed by the multiple comparisons problem, we included a relatively narrow set of outcome measures and subgroup breakdowns in our impact analysis. As indicated earlier in this report, our outcome analysis is focused on a set of seven key outcomes, including assessments of reading and English language outcomes. Our subgroup breakdowns were limited to the study sites, cohort variable, language background of the students, and their baseline reading level. This focus served to limit the number of impact estimates, which in turn, reduced the number of independent statistical tests.

In addition to this preventive effort to minimize the number of independent tests, we used a statistical correction to account for the multiple-comparison problem in our analyses. These statistical corrections unfortunately reduce overall statistical power for a study. In other words, they reduce the study's ability to detect impacts when true differences exist between the treatment and control group. Benjamini and Hochberg (B-H) (1995) developed a correction that minimizes this negative effect on the study's statistical power. Statistical power increases with the percentage of true impacts among the total number of outcome comparisons. As indicated by Schochet (2007), applying the B-H correction to a set of impact estimates reduces the power of any given test from 80 to 74 percent when the program has a non-zero impact on 80 percent of the outcomes measured. However, in the case where only 20 percent of the outcome comparisons have

non-zero impacts, the power to detect a single impact is reduced to 55 percent with the B-H correction. Therefore, as long as there is preponderance of evidence of the effectiveness of the intervention in increasing reading and English language test scores, this method largely preserves the study's ability to detect impacts. Only when impacts are rare does the B-H adjustment substantially reduce statistical power.

To implement the B-H adjustment, we used the procedures described by Thissen, Steinberg, and Kuang (2002), which rely on simple Excel spreadsheets to make the necessary adjustments. They do so by calculating a specific B-H critical value for each statistical test, which is based on the number of tests accompanying it, the degrees of freedom in the analysis, and the statistical significance of each of the tests. Specifically, the procedure developed and validated by Thissen and colleagues (2002) uses a version of the following:

Comparison	P-Value	Index	B-H Critical
A vs. B B vs. C C vs. D Etc.	From SAS or STATA output	Rank of p-values from large to small	[=((X-Index+1)*0.05/2*X)] (where X is the number of tests)

Adapted from Thissen, D., L. Steinberg, and D. Kuang. (2002, Spring) "Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons." *Journal of Educational and Behavioral Statistics,* 27(1), 77–83.

We followed the same approach in the current study. An advantage of this procedure is that it is completely transparent and can be implemented regardless of whether the analyst uses SAS, STATA, or other statistical software. Although some of these software programs include built-in procedures to make B-H adjustments, they do not demonstrate the impact of multiple comparisons on statistical power as clearly.

To implement multiple comparison adjustments, Schochet (2007) recommends dividing the outcomes into distinct domains and then conducting the adjustments for multiple comparisons within these outcome domains. Therefore, we divided our outcomes into the following domains:

English Reading Skills

- SARA Word Identification
- Woodcock Johnson Word Attack
- SARA Word Attack
- Woodcock Johnson Passage Comprehension

English Language Skills

- OWLS
- ROWPVT
- Woodcock Johnson Picture Vocabulary

The B-H adjustment was applied within the reading domain for the low literacy subgroup impact analysis, resulting in four comparisons to adjust.[31]

[31] Adjustments were only made for the analysis in which we found a statistically significant impact.

APPENDIX E:
SUPPLEMENTAL TABLES FOR CHAPTER 3

Table E.1: Percentage Distribution of Final Sam and Pat Lesson Number Covered in Class, as Reported by Sam and Pat Teachers

	Overall
Final Lesson Number	
3–6	13.6
7–10	18.2
11–13	18.2
14–16	27.3
17–22	22.7
Sample Size: 22 classes (10 missing cases).	

Note: Means are unadjusted, and based on all *Sam and Pat* classes for whom data were available. Details may not sum to 100 due to rounding.
Source: Adult ESL Literacy Impact Study spring 2009 *Sam and Pat* teacher data form.

Table E.2: Percentage Distribution of Students Attending Varying Numbers of Class Hours, Overall and by Group

	Total	*Sam and Pat*	Control
Number of Hours Attended			
0 hours	4.2	3.6	4.9
1–50 hours	32.7	28.6	36.9
51–100 hours	29.1	29.8	28.4
101–150 hours	26.9	31.8	21.9
Over 150 hours	7.1	6.2	7.9
Sample Size	**1,344**	**674**	**670**

Note: Percentages are unadjusted, and based on all students for whom attendance data were available. Details may not sum to 100 due to rounding.
Source: Adult ESL Literacy Impact Study attendance database.

Table E.3: Percentage Distribution of Students Attending Varying Percentages of Class Hours, Overall and by Group

	Total	*Sam and Pat*	Control
Percentage of Class Hours Attended			
0–25 percent	19.9	15.7	24.0
26–50 percent	13.8	11.9	15.7
51–75 percent	22.7	22.8	22.5
76–100 percent	43.7	49.6	37.8
Sample Size	**1,344**	**674**	**670**

Note: Percentages are unadjusted, and based on all students for whom attendance data were available. Details may not sum to 100 due to rounding.
Source: Adult ESL Literacy Impact Study attendance database.

APPENDIX F:
SUPPLEMENTAL ANALYSES FOR CHAPTER 4

Table F.1: Impacts Based on Reading Assessments Before Rescoring

Outcome	*Sam and Pat* Group	Control Group	Diff.	Effect Size	P-Value for Difference
Reading Assessments					
Woodcock Johnson Letter Word Identification	441.815	443.968	–2.154	–0.041	0.349
Woodcock Johnson Word Attack Scale	467.527	466.534	0.993	0.025	0.582
SARA Decoding	13.563	13.756	–0.193	–0.005	0.693
Sample Size: 1,137 students (587 *Sam and Pat* and 557 control).					

Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Table F.2: Impacts Based on Scaled English Language Assessment Scores

Outcome	*Sam and Pat* Group	Control Group	Diff.	Effect Size	P-Value for Difference
English Language Assessments					
OWLS Scale (based on age 12)	40.960	41.058	-0.098	-0.031	0.563
OWLS Scale (based on age 17)	40.492	40.562	-0.070	-0.035	0.526
ROWPVT Scale (based on age 12)	54.073	54.424	-0.351	-0.172	0.060
ROWPVT Scale (based on age 17)	53.170	53.331	-0.161	-0.041	0.144
Sample Size: 1,137 students (587 *Sam and Pat* and 557 control).					

Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Table F.3: Impacts Based on Raw Woodcock Johnson Scores

Outcome	*Sam and Pat* Group	Control Group	Diff.	Effect Size	P-Value for Difference
WJID	34.163	34.898	-0.735	-0.049	0.282
WJWA	11.964	12.069	-0.105	-0.013	0.779
WJPC	10.310	10.417	-0.107	-0.029	0.409
WJPV	7.115	6.986	0.128	0.036	0.429
Sample Size: 1,137 students (587 *Sam and Pat* and 557 control).					

Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Table F.4: Mean Pre- vs. Post-Test Scores on Reading and English Language Assessments, by Group

Outcome	*Sam and Pat* Group					Control Group					P-Value for Mean *Sam and Pat* vs. Control Group Gain Difference
	Mean Pre-Test Score	Mean Post-Test Score	Mean Gain (Diff.)	Effect Size	P-Value for *Sam and Pat* Group Gain	Mean Pre-Test Score	Mean Post-Test Score	Mean Gain (Diff.)	Effect Size	P-Value for Control Group Gain	
Reading Assessments											
WJID Scale (Rescored)	406.119	442.438	13.004	0.194	0.000*	402.224	443.978	15.841	0.232	0.000*	0.249
WJWA Scale (Rescored)	434.972	468.510	8.705	0.150	0.000*	429.488	465.871	9.454	0.159	0.000*	0.692
WJPC Scale	403.228	433.060	6.675	0.159	0.000*	401.560	434.260	6.763	0.168	0.000*	0.899
English Language Assessments											
OWLS	13.342	18.084	3.979	0.399	0.000*	13.367	17.876	3.689	0.361	0.000*	0.496
ROWPVT	21.381	28.524	6.018	0.381	0.000*	21.636	29.829	6.800	0.414	0.000*	0.224
Sample Size: 1,113 students (567 *Sam and Pat;* 546 control).											

*Indicates that difference is significant at 5 percent level, based on 2-tailed t-tests.
Notes: These figures are not regression-adjusted. Only assessments administered at both pre- and post-testing were included in this table. Calculations used data for all students for whom both pre- and post-test data were available.
Source: Adult ESL Literacy Impact Study assessments administered at the beginning and end of each term (fall 2008 and spring 2009).

Additional Sensitivity Analyses

In addition to testing whether impacts were sensitive to (1) site, (2) the use of scaled versus raw scores for the ROWPVT and WJ assessments, and (3) subgroup membership, we also tested the impacts' sensitivity to the following statistical assumptions:

- Students who are "no-shows" do not bias the estimates of impact; and
- The effect of each pre-test covariate on the corresponding post-test score does not vary across the pre-test distribution.

The remainder of this section provides a discussion of these assumptions and the results of our analyses.

Correcting for "No-Shows"

In this study, random assignment was conducted on the student's first day of class.[32] However, because the first day of class was largely taken up by intake activities and students were randomly assigned before instruction began, we counted students who did not show up again after the first day of the term as no-shows.

To test our assumption that no-shows did not bias our impact estimates, we implemented the no-show correction first proposed by Bloom (1984). This correction is based on the assumption that the overall net impact of a program, divided by the percentage of individuals who received *any* services is an unbiased estimate of the average impact per service recipient. Thus, if 10 percent of all sample members dropped out immediately after the first class (i.e., did not show up), the overall net impact could be divided by 0.9 (i.e., inflated by 11 percent) to estimate the impact on those who did participate. If students showed after the first class, they were counted as "shows" and the no-show correction does not apply to them.

[32] Random assignment typically occurred over a 2-week period at the beginning of the term. Students were randomly assigned on the first day they arrived at the site. However, because students who enrolled in the class after the first day of the term could have received some services, we can only identify no-shows among the students who show up on the first day of class each term and thereafter do not show up to class again.

In total, 63 students (5 percent) were considered no-shows under our definition.[33] Results of the no-show correction indicate that impact results were not biased by the exclusion of these students from analyses (Table F.5).

Table F.5: Impacts After No-Show Correction

Outcome	*Sam and Pat* Group	Control Group	Diff.	Effect Size	Adjusted Diff.	Adjusted Effect Size	P-Value for Difference
Reading Assessments							
WJID (Rescored)	440.808	442.086	–1.278	–0.025	–1.342	–0.026	0.573
WJWA Scale (Rescored)	466.698	465.757	0.941	0.024	0.988	0.025	0.595
SARA Decoding (Rescored)	13.251	13.368	–0.117	–0.011	–0.123	–0.012	0.809
WJPC Scale	432.825	433.519	–0.695	–0.038	–0.730	–0.040	0.295
English Language Assessments							
OWLS	17.804	17.787	0.017	0.002	0.018	0.002	0.974
ROWPVT	28.526	29.550	–1.023	–0.061	–1.074	–0.064	0.111
WJPV Scale	431.652	431.239	0.414	0.020	0.435	0.021	0.663
Sample Size: 1,137 students (587 *Sam and Pat* and 557 control).							

Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers, and adjusted using Bloom's no-show correction (1984). Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Spline Estimation

As was explained earlier in the report, we included pre-test scores as covariates in each of the impact analysis equations in order to increase the precision of the impact estimates. For the impact equations, we assume that the effect of the pre-test on the post-test score does not vary across the pre-test distribution. However, this assumption of constant pre-test effects may not be valid; for example, students with lower pre-tests might experience higher marginal gains from the intervention than those who started with higher pre-test scores, and the assumption could bias measured impacts, depending on the differences between

[33] There was no statistically significant difference in the number of no-shows in the *Sam and Pat* and control groups (26 compared to 37, respectively; p = 0.149).

the *Sam and Pat* and control groups in pre-test score distributions. To determine the extent to which the assumption of constant pre-test effects bias impacts, we estimated the impact equations with spline regression terms in place of the pre-test score. This specification allowed the slope of the pre-test score to vary according to whether or not the pre-test was below or above the group mean score.

The impacts measured with the spline regression equations are presented in Table F.6. Results show that the addition of the pre-test spline terms did not change the pattern of impacts in most cases. For example, for the Woodcock Johnson Letter-Word Identification assessment, the effect size was –0.06 using a spline function for the pre-test covariate, while the effect size for the standard impact equation was –0.03. However, for one assessment, the ROWPVT (vocabulary), the effect size was –0.13 and statistically significant using a spline function for the pre-test covariate, whereas the effect size for the standard impact equation was –0.06 and not statistically significant. This finding implies that students with below-mean pre-test scores achieved lower returns to their pre-test scores than they did in the linear impact specification; therefore, the model predicted a lower spline adjusted mean for students with below-mean pre-test scores than was predicted for the linear model. Similarly, the spline model predicted a higher adjusted mean than the linear model for students with at- or above-mean pre-test scores. Because the *Sam and Pat* group had a higher proportion of students with ROWPVT pre-test scores below the mean, the estimated impact using the spline function for the pre-test covariate (difference between *Sam and Pat* and control group) differs from the estimated linear model impact. This pattern of findings suggests that for most of the study outcomes, a linear estimation model was appropriate. For the ROWPVT impact analyses, however, we have less confidence that the results have met the required assumption of linearity, and therefore the impact estimates found for the ROWPVT should be interpreted with caution.

Table F.6: Impacts Estimated With Spline Terms

Outcome	*Sam and Pat* Group	Control Group	Diff.	Effect Size	P-Value for Difference	Slope for Pre-Test Scores Below Mean	P-Value for Slope for Pre-Test Scores Below Mean	Slope for Pre-Test Scores At or Above Mean	P-Value for At or Above Mean Pre-Test Score Slope
Reading Assessments									
Woodcock Johnson Letter Word Identification (Rescored)	440.502	443.322	–2.820	–0.055	0.242	0.530	0.000	0.276	0.000
Woodcock Johnson Word Attack Scale (Rescored)	466.573	466.167	0.406	0.010	0.815	0.080	0.286	0.436	0.000
Woodcock Johnson Passage Comprehension Scale	433.048	433.865	–0.817	–0.045	0.243	–0.220	0.000	0.788	0.000
English Language Assessments									
OWLS	18.050	18.041	0.009	0.001	0.986	0.021	0.795	0.655	0.000
ROWPVT	28.390	30.472	–2.081	–0.125	0.032*	0.030	0.689	0.704	0.000
Sample Size: 1,113 (567 *Sam and Pat* and 546 control)									

*Indicates that impact is significant at the 0.05 level, based on 2-tailed t-tests.

Notes: Spline terms were incorporated into linear regression models using dummy variables to measure the slope and intercept terms for scores below the mean and at or above the mean. In addition, pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers were included in the model to control for these characteristics. Calculations used data for all students for whom there were both pre- and post-test data. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups.

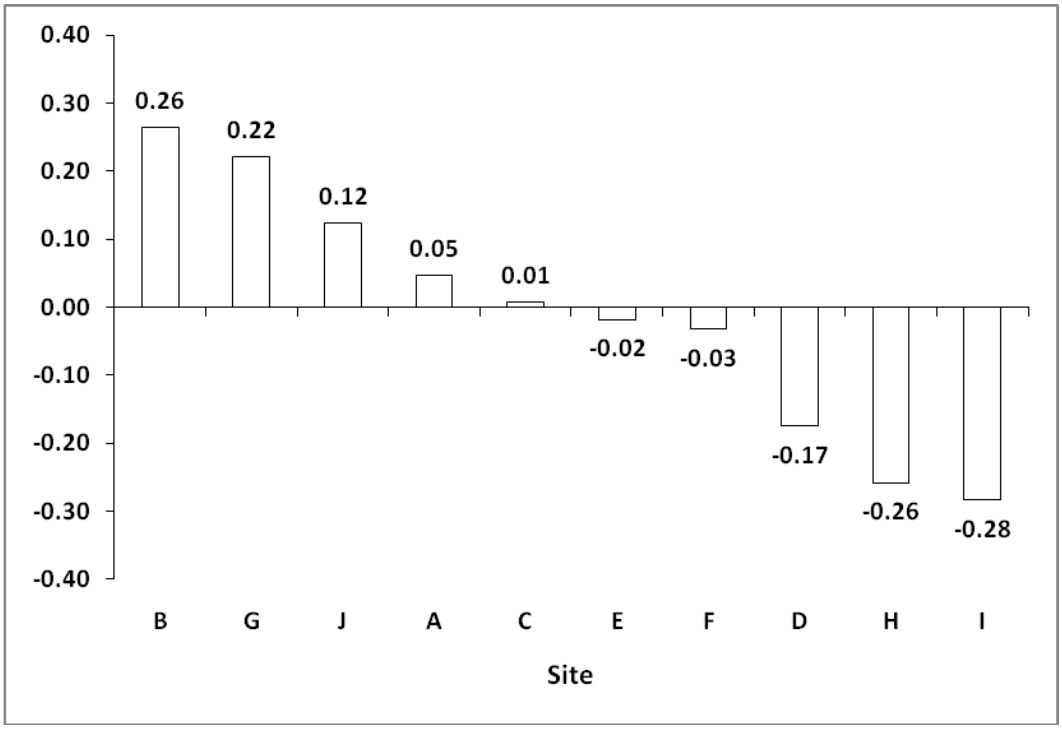
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Table F.7: Number and Percentage of Students Meeting the Study's Definition of Lower Literacy, by Site

Site	Number of Students Defined as Lower Literacy	Total Study Sample Size	Percent of Students Defined as Lower Literacy
A	158	222	71.2
B	38	54	70.4
C	75	109	68.8
D	48	86	55.8
E	35	72	48.6
F	140	349	40.1
G	17	61	27.9
H	22	98	22.4
I	34	205	16.6
J	13	88	14.8
Totals	**580**	**1,344**	**43.2**

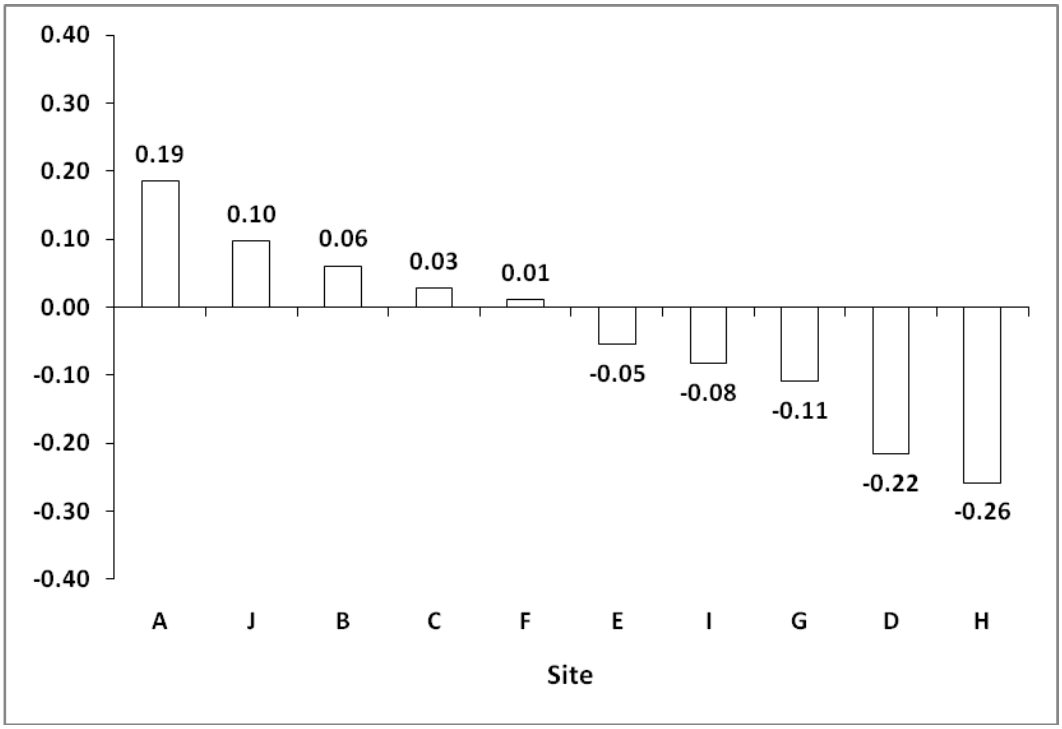
Note: Lower literacy was defined as having both (1) a raw pre-test score of 31 or below on Woodcock Johnson Letter Word Identification and (2) a raw pre-test score of 9 or below on Woodcock Johnson Word Attack.
Source: Student assessments administered at the beginning of each cohort (fall 2008 and spring 2009).

Figure F.1: Impacts on Woodcock Johnson Letter Word Identification Scale Scores, by Site



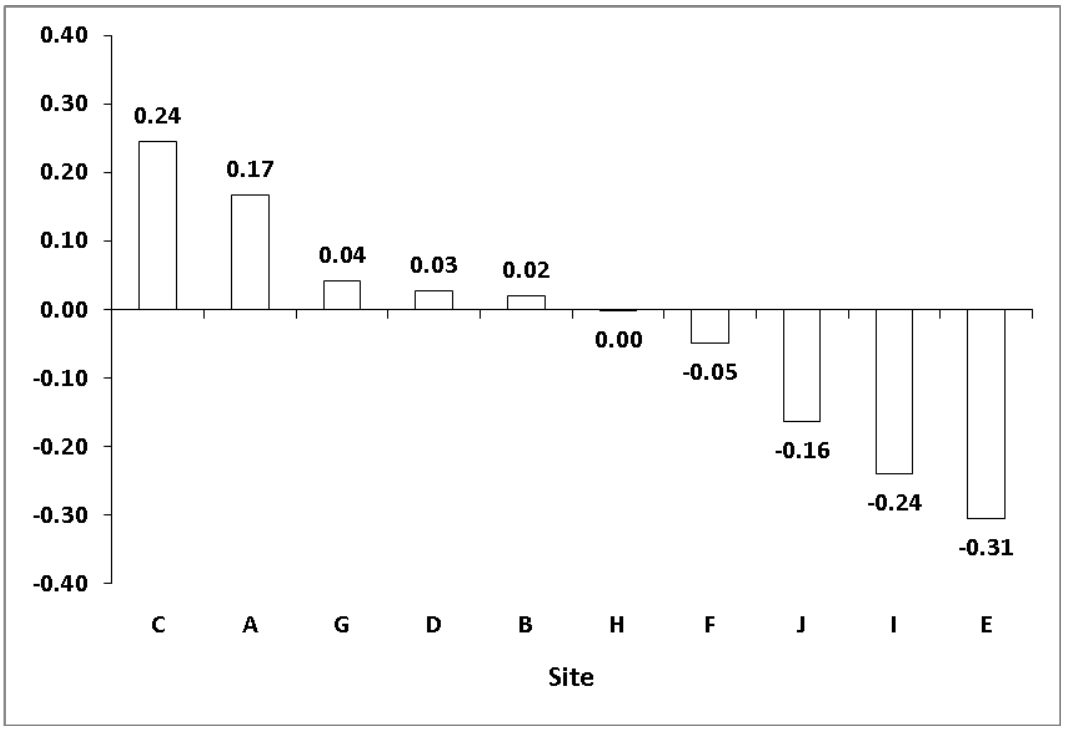
Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level. Impacts are ordered by magnitude.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Figure F.2: Impacts on Woodcock Johnson Word Attack Scale Scores, by Site



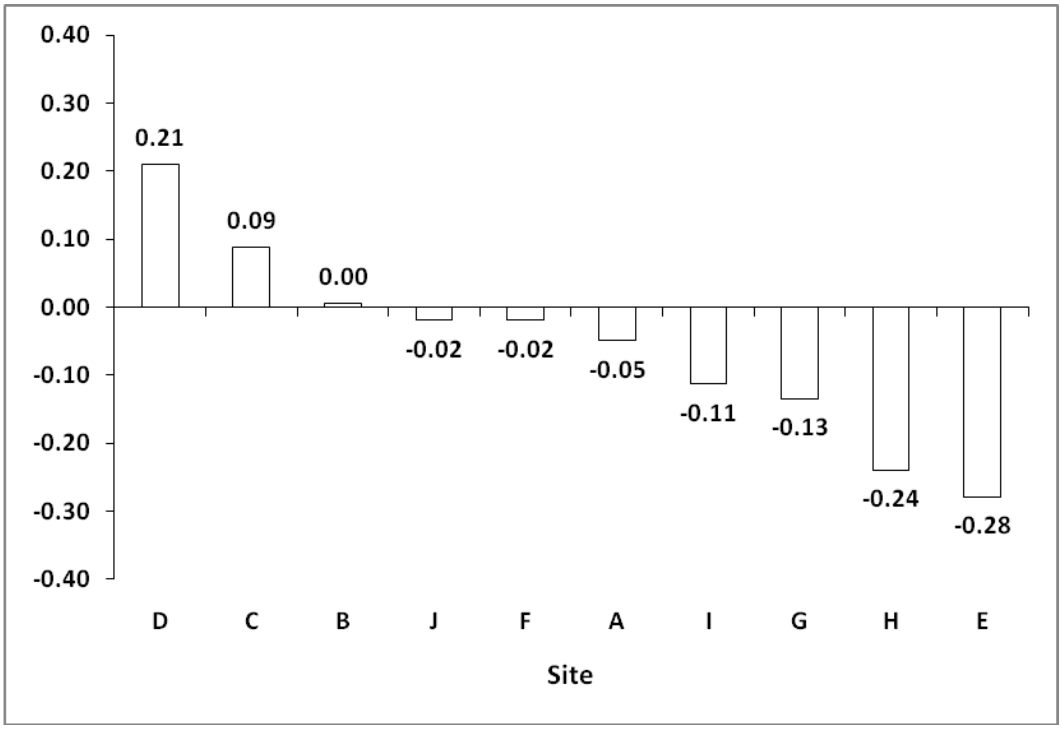
Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level. Impacts are ordered by magnitude.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Figure F.3: Impacts on SARA Decoding Scores, by Site



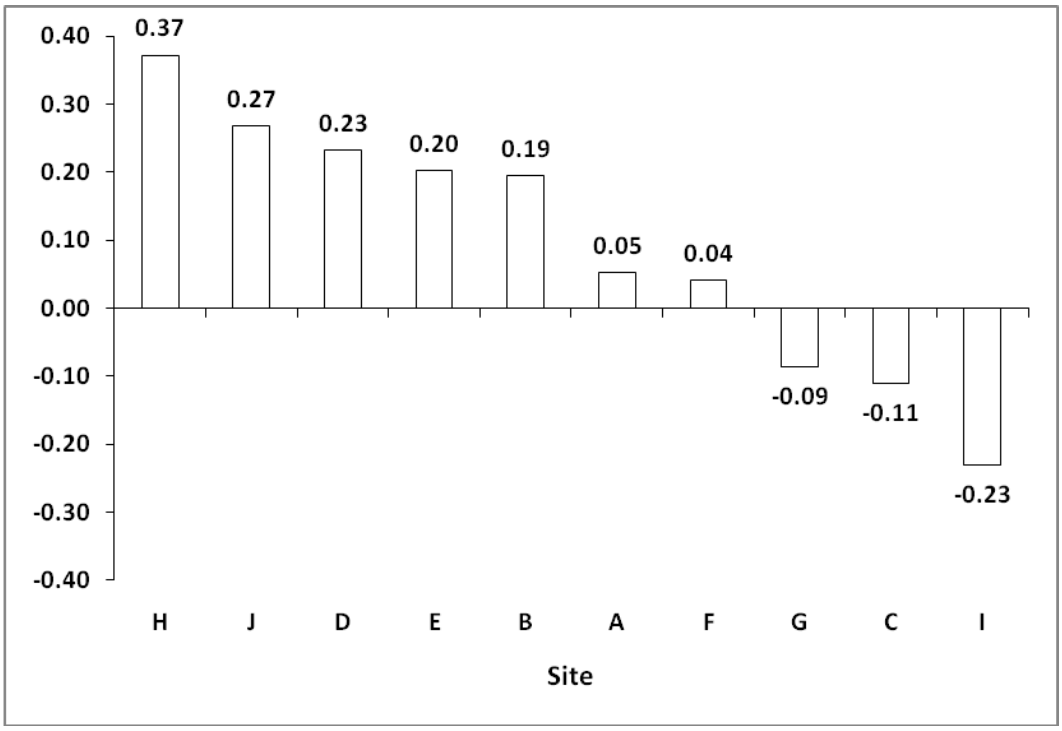
Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level. Impacts are ordered by magnitude.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Figure F.4: Impacts on Woodcock Johnson Passage Comprehension Scale Scores, by Site



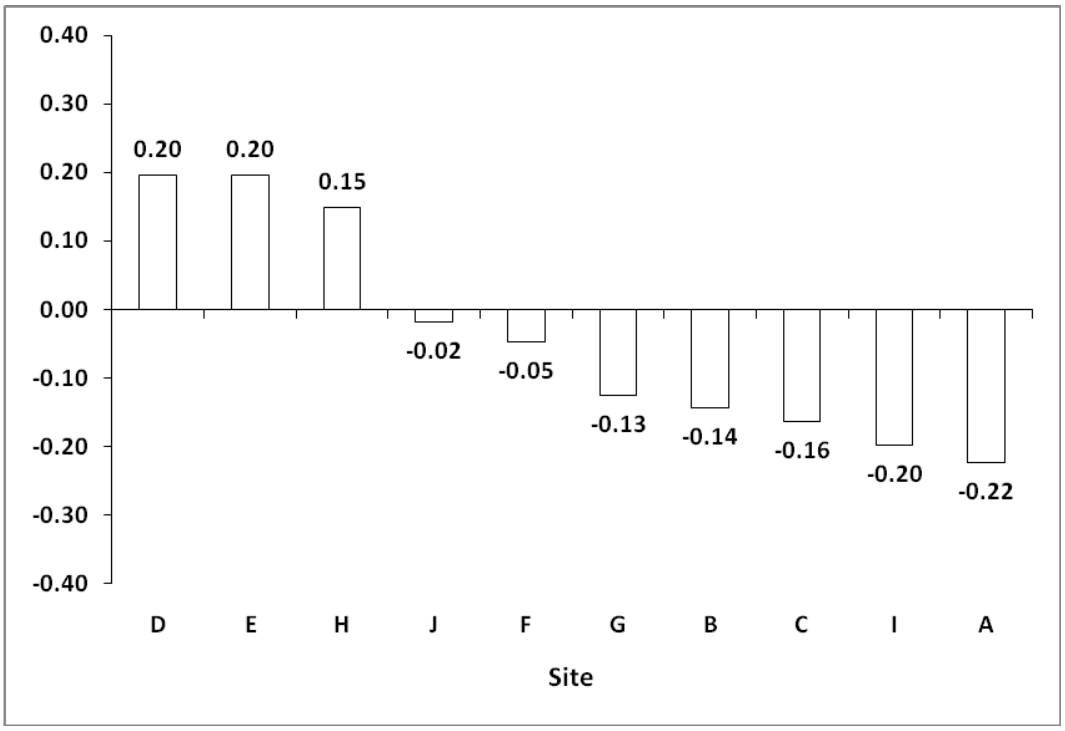
Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level. Impacts are ordered by magnitude.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Figure F.5: Impacts on OWLS Scores, by Site



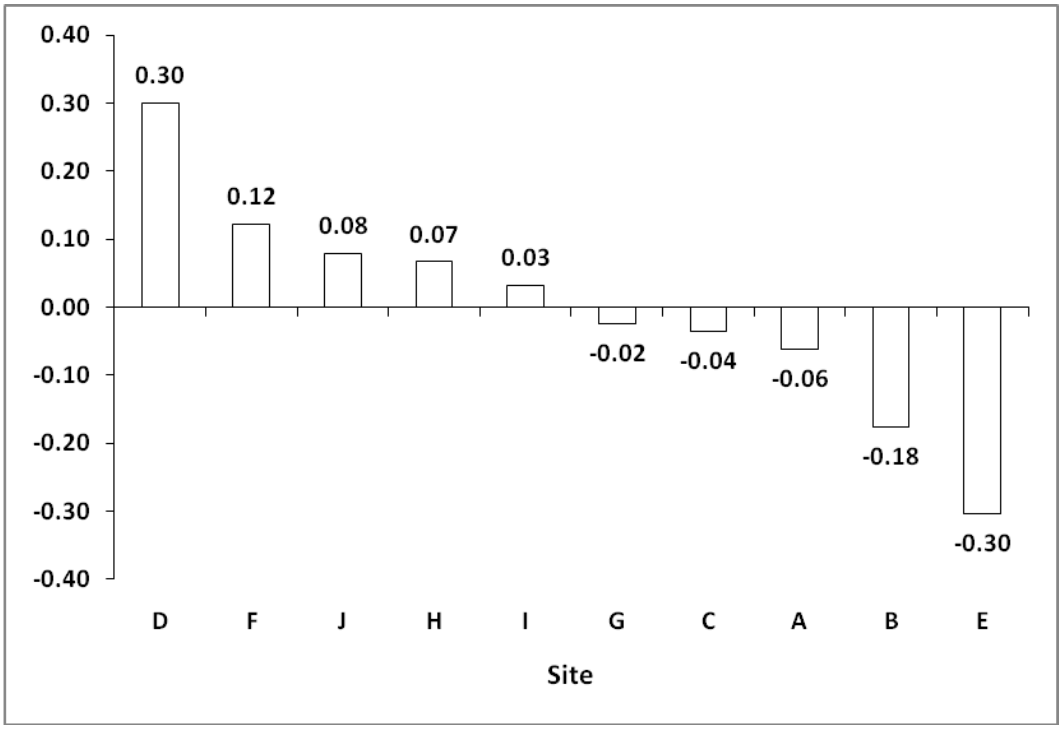
Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level. Impacts are ordered by magnitude.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Figure F.6: Impacts on ROWPVT Scores, by Site



Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level. Impacts are ordered by magnitude.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.

Figure F.7: Impacts on Woodcock Johnson Picture Vocabulary Scale Scores, by Site



Notes: Estimates were regression-adjusted using ordinary least squares, controlling for pre-random assignment characteristics of students, pre-test scores, and background characteristics of teachers. Calculations used data for all students for whom post-test data were available. A two-tailed t-test was applied to the differences between the *Sam and Pat* and control groups. The differences were not statistically significant at the 0.05 level. Impacts are ordered by magnitude.
Source: Adult ESL Literacy Impact Study student intake forms and assessments administered at the beginning and end of each term (fall 2008 and spring 2009), and fall 2008 teacher data form.