Using program evaluation to inform and improve the education of young English language learners in US schools

Lorena Llosa and Julie Slayton

Language Teaching Research 2009 13: 35
DOI: 10.1177/1362168808095522

Language Teaching Research 13,1 (2009); pp. 35–54

Using program evaluation to inform and improve the education of young English language learners in US schools

Lorena Llosa *New York University, USA*
Julie Slayton *Los Angeles Unified School District, USA*

The purpose of this paper is to discuss how program evaluation can be conducted and communicated in ways that meaningfully affect the education of English language learners (ELLs) in US schools. First, the paper describes the Waterford Early Reading Program Evaluation, a large-scale evaluation of a reading intervention implemented in schools with substantial populations of ELLs in a large urban school district in California. Second, using the Waterford evaluation as an example, this paper discusses the conditions necessary for conducting an evaluation that yields useful information about a program's implementation and effectiveness. The paper also highlights the importance of communicating those findings in a clear way so as to be meaningful to stakeholders and decision-makers in order to facilitate the goal of improving the education of young ELLs.

In the USA, public schools are under pressure to demonstrate significant increases in student achievement under the Federal No Child Left Behind Act of 2001 (NCLB). Large urban school districts, who primarily serve low-income and language minority students, are under particularly great pressure to identify and implement educational programs that will address the needs of their students. Despite the widespread implementation of programs specifically targeting urban student populations, such as English language learners (ELLs), there is a paucity of research and evaluation documenting the effectiveness of such programs. Kiely and Rea-Dickins (2005) explain that, 'for a range of reasons, some proper, others less so, evaluation processes and findings remain either insufficiently documented or unpublished' (p. 6). Those program evaluations that do exist are often conducted by publishers and/or only infrequently provide information about the conditions under which the programs were implemented or explanations for the programs' effectiveness or lack thereof. This lack of high-quality research prompted the Department of Education's Institute for Education Sciences to create the What Works Clearinghouse (WWC, http://www.whatworks.ed.gov/), a project designed to

Address for correspondence: Lorena Llosa, Department of Teaching and Learning, Steinhardt School of Culture, Education, and Human Development, New York University, 239 Greene Street, 6th Floor, New York, NY 10003, USA; email: lorena.llosa@nyu.edu

 10.1177/1362168808095522

review studies of instructional intervention effectiveness based on strict criteria for what constitutes scientifically based research in education. Of 255 studies reviewed to assess the quality of 51 products, 75% did not meet WWC criteria for scientifically based research. In fact, 'not a single product has more than one study fully meeting WWC research standards' (Oppenheimer, 2007, pp. 28–29).

One consequence of this lack of trustworthy, readily available, and accessible research on program effectiveness is that school districts are forced to make extremely important policy decisions to adopt new programs based on limited data. Moreover, once a program is adopted, a lack of ongoing evaluation impedes a district's ability to make informed decisions about program continuation and discontinuation. Decisions then are often made by administrators based on anecdotal data from teachers in classrooms (Oppenheimer, 2007). Even when there is research available on program effectiveness, school districts are confronted with additional challenges, including the extent to which the research is presented in a way that is accessible or understandable to the stakeholders and the political environment in which the decision is being made.

The purpose of this paper is to discuss how program evaluation can be conducted and communicated in ways that meaningfully affect the education of ELLs in the complicated environment of US K-12 education. First, we present the design and the findings of a large-scale evaluation of a reading intervention implemented in schools with substantial populations of ELLs in a large urban school district in California. Second, using this evaluation as an example, we discuss the following: (a) the conditions necessary for conducting an evaluation that yields useful information about a program's implementation and effectiveness, and (b) the importance of communicating those findings in a clear way so as to be meaningful to stakeholders and decision-makers in order to facilitate the goal of improving the education of young ELLs.

I The Waterford early reading program evaluation

In this section, we describe the design and findings of a two-year evaluation of the Waterford Early Reading Program (hereafter the Waterford program) as implemented in the Los Angeles Unified School District (LAUSD) in Los Angeles, California. The description focuses on the main components of the evaluation, and in particular on those directly related to ELLs. For a complete account of the two-year evaluation, readers are referred to the published evaluation reports (Slayton & Llosa, 2002; Hansen, Llosa, & Slayton, 2004) available on the LAUSD website.

1 Background

The Waterford Program was adopted by the LAUSD to address the needs of kindergarten and first grade students who were the most at risk of experiencing reading failure. With a population of 715,541 students, 41% of whom are ELLs

(LAUSD, 2005–2006), this was the largest subgroup impacted by the district's adoption of the program. The Waterford program is a computer-based literacy program created by the non-profit Waterford Institute. It was first adopted by the district in 2001 in 2235 kindergarten and first grade classrooms in 244 low performing elementary schools with high percentages of ELLs. While the Waterford program is a complete reading/language arts curriculum with a computer-based and a teacher-directed component, only the computer component was adopted by LAUSD to supplement the district-wide primary reading program, Open Court.

Three Waterford computer stations were placed in each participating classroom, and individual students rotated from teacher-directed activities to the computer station throughout the day. Kindergarten students were expected to spend 15 minutes and first grade students were expected to spend 30 minutes a day using the Waterford courseware. The Waterford courseware is adaptive in that each student logs in at the beginning of their session and moves through the lessons at their own pace. In addition, the program has the capacity to be further tailored by the teacher to the instructional needs of each child. Level One of the Waterford program used in kindergarten classrooms focuses on print concepts, phonological awareness, and letter recognition. Level Two, used with first grade students, focuses on letter sounds, word recognition, and beginning reading comprehension. These topics covered by the Waterford courseware are only a small subset of the skills covered by the Open Court curriculum for kindergarten and first grade. (See analysis of alignment between the programs in Hansen, Llosa, & Slayton, 2004.) The rationale for the adoption of the Waterford program was that students in these low-achieving schools needed additional instructional time in reading and that they would benefit from an alternative instructional mode that was adaptable to their level of proficiency and was presumably more engaging, due to its attractive interface, than teacher-directed instruction.

The Program Evaluation and Research Branch of the LAUSD was charged with the evaluation of the program's implementation and effectiveness. It is important to point out that the Waterford evaluation was conducted in a highly politically charged environment. The district had spent $64 million to purchase and sustain the program. It had been announced as the 'Cadillac' of interventions and was seen as a significant part of the district's solution to low achievement for its at-risk populations. Additionally, the program vendor was promoting the program as a success in order to encourage the district to purchase additional software, materials, and workstations. They relied on their own research to demonstrate the program's effectiveness both inside the district and in districts around the country. Additionally, there were individuals within the district staff who were strong supporters of the program. Thus, there was a substantial bias in support of the program from the outset, which made it critical that the evaluation be carefully designed and that findings be communicated appropriately and convincingly to stakeholders, so the recommendations would be considered and not dismissed.

In order to inform the district regarding the effectiveness of the program, the evaluation needed to compare students who were exposed to the Waterford program (treatment group) with comparable students who were not (comparison group) using a quasi-experimental design. Also, given that the program was intended as an in-class supplement to the district's primary reading program *Open Court,* the evaluation had to examine the context surrounding the implementation of the courseware, in particular, the interaction between *Open Court* and the Waterford program, the extent to which students were engaged by this program, and the extent to which teachers used the program effectively. In order to gather the necessary information about the implementation and effectiveness of the Waterford program, four important considerations were included in the design: (1) an investigation of the context, (2) the use of multiple types and sources of data, (3) the use of appropriate analytic tools, and (4) the use of extensive qualitative data. Below we explain how these features were incorporated into the evaluation. Later, in the discussion section, we discuss the critical role each played in the evaluation design and in the presentation of findings.

2 Research questions

For the two-year evaluation, two overarching research questions guided investigations:

1. Does the Waterford Early Reading Program, as implemented in LAUSD, have an effect on student reading achievement?
2. To what extent is the Waterford courseware being implemented?

In each year, the study also addressed specific questions focusing on different aspects of program implementation and its effectiveness across different subgroups to investigate the program's claims that it engages students 'with an easy-to-use, fun interface that moves them from introduction to mastery of critical concepts, regardless of primary language or beginning skill level' (http://www.pearsondigital.com/waterford/). These questions included the following:

a Program effectiveness:

- To what extent was the program effective with ELLs vs. English Only (EO) students?[1]
- To what extent was the program effective with ELLs of varying levels of English proficiency?

b Program implementation:

- To what extent did use of the Waterford program provide students with additional reading instruction beyond that provided by the primary reading program?

- To what extent did teachers tailor the use of the Waterford program to the particular needs of individual students?
- To what extent did the Waterford program engage students, both ELL and EO, more than the primary reading program *Open Court*?

3 Method

The study employed a quasi-experimental design with matched samples of students in a treatment group and a comparison group. Both quantitative and qualitative methods were used to gather and analyze different types of data from multiple sources in order to address the research questions.

a Participant sample: At the outset of the study, a total of 200 classrooms were selected for participation in the Waterford Evaluation: 100 kindergarten classrooms and 100 first grade classrooms. In each grade, 50 classrooms were in the treatment group and the other 50 in the comparison group. To ensure that the comparison classrooms matched the expected characteristics of the treatment classrooms, a list of criteria was established and followed to create the matched sample. This list included the following: (a) racial/ethnic composition, (b) percentage of free/reduced lunches, (c) percentage of ELLs, (d) calendar type, and (f) track.[2]

From each of the 200 classrooms, a random sample of 10 treatment and 5–10 comparison students was selected for individual testing of their reading ability at the beginning (fall) and the end (spring) of the academic year. In Year 1 of the evaluation, 867 treatment students and 363 comparison students were tested in both fall and spring. Of those 1230 students in the Year 1 sample, 813 (66%) were classified as ELLs. In Year 2, 1019 treatment students and 849 comparison students had pre and posttest scores. Of the 1868 students in the Year 2 sample, 1229 (66%) were ELLs.

b Data collection: In order to address question 1 regarding the impact of the Waterford program on student reading achievement, individual students were tested using the Woodcock Reading Mastery Tests – Revised, Form G (WRMT-R; Woodcock, 1987) within the first four weeks and within the last four weeks of the school year. The WRMT-R is a battery of tests that measures several aspects of reading ability for kindergarteners through adults. It consists of eight subtests: Visual Auditory Learning, Letter Identification, Word Identification, Word Attack, Word Comprehension (Antonyms, Synonyms, and Analogies), and Passage Comprehension. Only the first four tests of the WRMT-R (Visual Auditory Learning, Letter Identification, Word Identification, and Word Attack) were administered to kindergarten students. In first grade, all eight of the WRMT-R tests were used.

In order to address question 2 on implementation, classroom observations and teacher interviews were conducted. A team of research assistants visited

classrooms for two days each during both the fall and spring semesters. In Year 1, observations and interviews were conducted in the 200 treatment and comparison classrooms to examine the quality of teacher pedagogy during *Open Court* reading instruction. Determining that the quality of pedagogy was comparable in treatment and comparison classrooms was important in order to attribute any potential reading effects to the use of the Waterford program. In Year 2, observations and interviews were only conducted in the 100 treatment classrooms. Revised protocols were designed for this phase to capture more detailed information regarding the use of the Waterford courseware in relation to *Open Court* instruction. Waterford usage data (i.e., a record of each student's use of the program over the course of the school year) was also obtained from the Waterford Institute.

All teachers were interviewed in Year 1 about their experiences with *Open Court*, specifically the training and support they had received to implement it and their perceptions of how effective it was for teaching reading. In addition to questions about *Open Court* implementation, treatment teachers were interviewed about their experiences with the Waterford Program. Questions focused on the training and support they had received, challenges they experienced in its implementation, their daily use of the courseware including number of students who used it each day and number of days a week it was used, their knowledge of the various courseware features, and their perceptions of its effectiveness with all students and ELLs in particular. In Year 2 only treatment teachers were interviewed.

c Data analysis: Qualitative data reduction and analysis (Merriam, 1998) was conducted with each type of data collected. In Year 1, classroom observation data was primarily used to examine the quality of teacher pedagogy and implementation of *Open Court*, to ensure comparability between the treatment and comparison groups. Classroom observation data was examined using a rubric created to assess the quality of pedagogy in each classroom. The rubric identified five levels of quality: high, medium-high, medium, medium-low, and low. For example, in a high quality pedagogy classroom, teachers would be observed teaching the skills presented within the *Open Court* teacher's manual with a high degree of fidelity on both days of the observation. High fidelity required both that the teacher cover a high proportion of the activities set forth in the teacher's manual and that the teacher did so using the same techniques and with the same quality as presented in the teacher's manual. (For the complete rubric, see Slayton & Llosa, 2002.) In Year 2, the analyses focused on data collected in treatment classrooms on the extent of overlap between the use of the Waterford courseware and *Open Court* instruction. Classroom observation data was also used to determine the number of students in each class who used the computer each day, the amount of time they spent, and the individual student level of engagement while using the courseware and during Open Court instruction. Level of engagement was determined using a

rubric designed to identify four levels of engagement: fully engaged, experienced only minor distractions, distracted, and off-task. (For engagement rubric and examples, see Slayton & Llosa, 2002.) In addition to classroom observation data, usage data provided by the Waterford Institute was used to determine individual student usage of the courseware. Finally, teacher interview transcripts were coded with respect to their use and perceptions of *Open Court* and the Waterford courseware.

In order to determine program effectiveness, various statistical analyses were conducted, including correlation, analysis of covariance, and hierarchical linear modeling. Hierarchical linear modeling or HLM (Raudenbush & Bryk, 2002) was used to examine differences between treatment and comparison students, and, within treatment classrooms, to examine the relationship between the time spent using the courseware and student reading achievement. HLM is particularly appropriate for analyzing K-12 data because it takes into account the nested or multilevel nature of the data (students are nested in classrooms and classrooms are nested in schools) and simultaneously considers the effect of student variables and classroom and school characteristics on outcomes.

4 Findings

In this section we present a summary of the major findings related to the effectiveness and implementation of the Waterford program. We also report on the relative effectiveness of the program with ELL vs. EO students as well as students with varying levels of English proficiency. Finally, we address the extent to which exposure to the Waterford program provided students with the presumed benefits of this particular intervention, namely, additional instructional time, individualized attention, and greater engagement. (For complete details on Year 1 analyses and findings see Slayton and Llosa, 2002, and for Year 2 see Hansen, Llosa, & Slayton, 2004.)

a Program effectiveness: *Year 1*. The results of the ANCOVA and HLM analyses indicated that exposure to the Waterford program had no or minimal impact on reading achievement. In kindergarten, no differences were found between the treatment and the comparison group on any of the WRMT-R tests. In first grade, treatment students had larger gains than comparison students in Letter Identification only. Treatment students also had larger gains than comparison students in the Synonyms test, but only in classrooms with higher quality *Open Court* pedagogy. Also, for students in the treatment group, the amount of time spent on the Waterford program had no impact or had a negative impact on gains. For example, in kindergarten, on average, time spent while using the Waterford courseware by any individual student had no effect on gains on the Word Identification test. However, for students in classrooms with higher quality *Open Court* pedagogy, time spent using the courseware had a negative effect.

Year 2. There were no statistically significant differences or any sizeable effects on the WRMT-R between treatment and comparison students in either kindergarten or first grade. Furthermore, among students in the treatment group, neither the amount of time spent on the Waterford program, nor the students' level of engagement while using the program, had any perceptible impact on achievement.

b Program effectiveness by language classification: Year 1. The analyses revealed no differences between the kindergarten ELLs in the treatment group and ELLs in the comparison group. However, within the treatment group, ELL students had larger gains than EO students in the reading readiness tests (Visual Auditory Learning and Letter Identification), and EO students had larger gains than ELL students in the basic skills tests (Word Identification and Word Attack). Overall, in kindergarten, exposure to Waterford benefited EO students in developing the basic skills, but not ELLs. The program only helped ELL students develop reading readiness skills.

In first grade, ELL treatment students had larger gains than ELL students in the comparison group on the Visual Auditory Learning test. Within the treatment group, EO students had larger gains than ELLs in the basic skills and comprehension tests (Word Identification, Antonyms, Analogies, and Passage Comprehension). With the exception of Visual Auditory Learning, Waterford did not make a difference for first grade students.

For a more comprehensive understanding of the relationship between the use of Waterford and English language proficiency, the ELL student group was examined in terms of ELD Level. In kindergarten, there were no statistically significant differences between ELD 1–2 treatment students and ELD 1–2 comparison students, nor between ELD 3–4 treatment students and ELD 3–4 comparison students. Not surprisingly, within the treatment group, students in ELD 3–4 had larger gains than ELD 1–2 students in Word Identification and Word Attack. In first grade, the only statistically significant difference between ELD 1–2 students in the treatment group and ELD 1–2 students in the comparison group was in the Visual Auditory Learning Test. No differences were found between ELD 3–4 students in the treatment group and ELD 3–4 students in the comparison group. Within the treatment group, ELD 3–4 students had larger gains than ELD 1–2 students in Word Identification, Synonyms, Analogies, and Passage Comprehension.

In general, as would be expected, students with higher English language proficiency performed better than students with lower English language proficiency on most tests regardless of their membership in the treatment or the comparison group.

Year 2. No statistically significant differences were found in kindergarten between ELL students in the treatment group and ELL students in the comparison group; nor were differences found in first grade when comparing the ELL students in the two groups.

As in Year 1, students with higher levels of English proficiency outperformed those with lower levels of proficiency: Kindergarten EO students outperformed ELL students regardless of condition in Word Identification and Word Attack. First grade EO students outperformed ELL students in Letter Identification, Antonyms, Synonyms, and Analogies.

Overall, exposure to the Waterford program had no perceptible effect for students as a whole, or English language learners in particular. Among students in the treatment group, greater time using the courseware did not result in increased student achievement. The next section reports the findings related to implementation and discusses why the presumed benefits provided by the Waterford program – additional instructional time, personalized instruction, and greater engagement – were not reflected in the scores.

c Implementation: In order to determine the extent to which the Waterford courseware was used, two sets of data were examined: usage data generated by the Waterford Institute and classroom observation data. In both years of the evaluation, use of the courseware (in terms of minutes spent over the entire school year) was very low. As mentioned previously, kindergarten students were expected to spend 15 minutes and first grade students were expected to spend 30 minutes a day, five days a week, using the Waterford courseware. Yet in Year 2, for example, on average kindergarten students used the program less than half of the recommended amount of time (47%), and first grade students used it less than one third (30%) the recommended amount of time.

The usage data also revealed that no individual student had used the courseware for the recommended amount of time. In fact, only 39% of the kindergarten students and 14% of first graders used the program at least half of the recommended time. Perhaps most striking was the fact that 25% of kindergarten and 40% of first grade students used the courseware less than one quarter of the recommended amount of time. The classroom observation data revealed additional information to explain this finding.

First, 78% of kindergarten teachers and 58% of first grade teachers used the program to some extent on all four days of observation. Computer malfunction was the most common reason why teachers did not use the computer on all the days of observation, but other reasons included, for example, lack of time due to other lessons or tests, earthquake drills, and dance practice.

Even though the Waterford program was being used in the majority of kindergarten classrooms during the four days of observation, the number of individual students who used the courseware over the four days of observation was low: Only 25% of kindergarten students and 11% of first grade students. Furthermore, 10% of kindergarten students and 18% of first grade students in treatment classrooms did not use the courseware at all during the four days of observation.

Thus, there were two different explanations for the low usage of the Waterford program. Not only did teachers not use the computers every day,

not every student had access to the computers on the days when they were being used.

d Other aspects of implementation: *Additional instructional time*. Given that the Waterford program was adopted in an attempt to supplement primary reading instruction provided by the *Open Court* program, another aspect of implementation examined was the relationship between the Waterford program and the primary reading program, *Open Court*. Data from Year 1 of the evaluation suggested that there might be some overlap in instructional time between the two programs, so in Year 2 the evaluation was designed to systematically examine the extent to which the Waterford program *supplemented* versus *supplanted* the primary reading program.

Classroom observation data in Year 2 confirmed that the use of the courseware overlapped with primary reading instruction in the majority of the classrooms. Approximately half of the kindergarten classrooms and two thirds of the first grade classrooms were observed using the courseware during *Open Court* reading instruction. In other words, on any given day, from 20% to 31% of students missed all or part of their primary reading instruction and instead were exposed to the Waterford courseware. Thus, the courseware was not consistently used to supplement the primary reading program or to provide students with additional instructional time as originally intended.

Individualized instruction. Another often mentioned advantage of a computer-adaptive instructional program is that it can be used to address the specific needs of individual students. The Waterford program includes a School Manager feature that allows teachers to set the amount of time each student uses the computer, mark students absent, and move students to the next level of the program, among other functions. However, the classroom observations and teacher interviews revealed that teachers were not using the courseware in such a way as to ensure that students were being exposed to the content most appropriate to them. For example in Year 2, only 30% of kindergarten teachers and 25% of first grade teachers used the School Manager to select lessons and activities based on their assessments of individual students' needs. Even more problematic, some teacher did not use the School Manager to monitor their students' usage of the courseware. While teachers were supposed to 'mark a student absent' so that the student would not be called to the computer when he or she was not present for the day, some of the teachers did not do this and sent the wrong student to replace a student who was absent. In other words, although the program indicated that it was Juan's turn, Oscar was sent to the computer instead to complete Juan's session. On average, on each of the 4 days of observation we documented seven instances of the wrong student using the computer (approximately 1% of turns each day). One teacher in our sample said, 'I don't go by student names that come up on the screen. I just send them when I think it's appropriate. That way it's irrelevant who is here and who is not here.' By doing

this, teachers were undermining the program's capability to adapt instruction to individual students' pace and level.

Increased engagement. One of the most frequently mentioned advantages of the Waterford courseware is that it engages students with its attractive and fun interface. The majority of teachers in the Year 2 sample concurred: 80% of the kindergarten teachers and 69% of the first grade teachers interviewed said that the Waterford courseware did engage students for the duration of their turn on the computer. The majority of teachers – 82% of kindergarten teachers and 76% of first grade teachers – also reported that the program engaged students who normally had difficulty staying engaged. One teacher gave the following example: 'Claudia and Melvin, they are very hyperactive. They have a hard time staying focused. When they're on that computer they are focused. They're paying attention, they want more, they aren't distracted. They love it.' Classroom observation data confirmed teachers' perceptions: the majority of students (between 54% and 70%) were either fully engaged or only experienced minor distractions on any given day of observation. In Year 1, an even higher proportion of students (between 73% and 80%) were fully engaged or experienced minor distractions on any given day.

The high engagement provided by this technology-based program was hypothesized to be particularly important for ELLs, but this did not turn out to be true. In Year 1 for example, classroom observations revealed examples of ELLs who were using the Waterford program but who clearly did not understand the instructions on the screen. In some cases, these students sought out assistance from fellow classmates, the teacher, or the teacher's aide. In other cases, the students appeared to become disengaged when they could not understand what to do. On average, the proportion of kindergarten and first grade ELL students distracted or off-task (23% and 22%) was greater than the proportion of EO students who were distracted or off-task (14% and 14%). Similarly, within the ELL population, the proportion of ELD 1–2 students who were distracted or off-task (24% and 24%) was greater than the proportion of ELD 3–4 students who were distracted or off-task (15% and 16%).

Interestingly, the same relationship between English language proficiency and level of engagement was evident during *Open Court* instructional time. Of those students who were identified as disengaged, most were ELL students (72% in kindergarten and 63% in first grade) and, within the ELL students who were disengaged, most were ELD 1–2 students (93% in both grades).

These percentages suggest that students with lower English language proficiency have a more difficult time staying engaged during classroom instruction and use of the courseware. Therefore, the assumption that a technology-based intervention would be particularly engaging to ELLs and the program's claim that it engages students 'regardless of primary language' did not hold true in this case.

Overall, the two-year evaluation of the Waterford Program revealed that the use of the courseware as a supplement to the primary reading program did

not help ELL or EO students make improvements in reading achievement. The implementation of the courseware was low, but even among students in the treatment group, greater time spent on the courseware did not result in higher reading achievement. The evaluation also revealed that the presumed benefits of this computer-adaptive program were not realized. The courseware was often being used in class at the same time as the primary reading program, thus supplanting it instead of supplementing it. Also, teachers were not using the program so as to expose students to content appropriate for their level, and the courseware turned out not to be as engaging as anticipated for ELLs. Based on these findings, a number of recommendations were proposed to the Board of Education in the final report. These recommendations are presented in the next section.

II Necessary conditions for a useful evaluation

The evaluation of the Waterford program implementation serves as a useful example for discussing critical issues related to large-scale program evaluation in general, and to program evaluation in the US K-12 context in particular. This discussion focuses on two elements we argue are critical for carrying out a successful evaluation: the evaluation design and the reporting of findings.

1 Design elements essential for conducting a useful evaluation

The Waterford program was adopted in all schools in the district that were below a certain achievement level, therefore conducting a randomized experiment was not possible. Instead, a quasi-experimental design was adopted that included a treatment group and a comparison group composed of students with comparable characteristics. This design allowed for the direct comparison of reading achievement of students exposed to the Waterford program and those who were not in order to determine whether exposure to the program resulted in increased achievement. We argue, however, that adopting an experimental or quasi-experimental design alone is not sufficient. Next, we discuss the four additional design considerations that we believe are essential for conducting a useful evaluation.

a Investigating and understanding the context: The Waterford program was one of many interventions being implemented in the district, along with *Open Court* and other programs for math, science, and other subject areas. Thus, it was important to approach the evaluation understanding the reality that the courseware was being implemented alongside other instructional activities. Conducting an evaluation that focused solely on the variables directly related to the Waterford program, such as the amount of use of the program and whether it was effective, would have yielded information of limited utility.

The quantitative data would have revealed, as it did, that the Waterford program was not helping students improve their reading in English. However, by investigating the extent to which it was actually used, we were able to determine that one possible reason why the program did not have the intended effect was that implementation was low. Once again, though, if the evaluation had stopped here, the recommendation would have been that teachers simply needed to implement the program with greater fidelity for the amount of time recommended. This limited focus on Waterford variables alone would have resulted in a recommendation that would not have translated into improved implementation (or student achievement) because it would have failed to consider the actual context within which the program was being implemented. Our more comprehensive data collection and analysis had demonstrated, for example, that teachers could not use the program for the amount of time required, even if they wanted to, due to a number of scheduling constraints (as evidenced by the overlap with primary reading instruction).

In order to get at these underlying issues, it was critical that we collect not only information about student performance and time spent on the courseware, but also information about the context surrounding the program implementation, such as the quality of the teacher pedagogy during reading instruction and the activities in which the class was engaged when Waterford was being used. We agree with Maxwell (2004) that 'to develop adequate explanations of educational phenomena, and to understand the operation of educational interventions, we need to use methods that can investigate the involvement of particular contexts in the processes that generate these phenomena and outcomes' (p. 7). The careful examination of the context made it possible for us to uncover the reasons why the program was not being implemented as intended and provide recommendations that 'made sense' in the actual context of these schools and classrooms.

b Using multiple types and sources of data: In order to fully understand the Waterford program and the context in which it was being implemented, we relied on multiple types and sources of data. For example, when looking at implementation, we sought the usage data generated by the Waterford computers. We also used classroom observation data to corroborate the usage data as well as to understand usage at the classroom level. As explained in the findings, the classroom observation data allowed us to examine factors like the number of classrooms that used the Waterford program on any of the four days of observation, the number of students within each classroom who used the courseware on any given day of observation, and, at the individual student level, the extent to which any given student used the courseware on one, two, three, or all four days of observation. Interview data from teachers allowed us to examine their perceptions about their own use of the courseware. The various types of quantitative and qualitative data allowed us to put together a picture of the use of the program and the context in which it was being used. This

multi-methodological approach allowed us to gain a deeper understanding of the factors affecting the program implementation and avoid the incorrect interpretations that would have resulted from a limited focus on achievement data and Waterford usage data alone.

c Using appropriate analytic tools: The use of appropriate analytic tools is essential for maximizing the information that can be gathered from an evaluation. The use of HLM in the Waterford evaluation was critical both for meaningfully analyzing the data and helping shape the evaluation from Year 1 to Year 2. HLM allowed us to look not only at differences between the treatment and comparison groups but also allowed us to look within the treatment group to examine the relationship between use of the courseware and achievement, while controlling for other classroom characteristics such as quality of teacher pedagogy. As a result, we were able to detect nuanced yet critical relationships in the data (Newton & Llosa, 2008). For example, we found that time spent on Waterford had no effect on kindergarten students on the Word Identification test, but in classrooms with higher quality pedagogy, time spent on Waterford had a negative effect. An explanation was that the ability to recognize words is not a focus of the Waterford program in kindergarten, but it is a focus of *Open Court* instruction. Thus, it is possible that time taken away from high quality *Open Court* pedagogy impacted the students' level of achievement on this skill. These and similar findings from the HLM analysis led us to focus on the overlap between Waterford and primary reading instruction during the extensive classroom observations in Year 2.

d Incorporating qualitative data: One of the most important decisions made in the design of the evaluation was to incorporate qualitative classroom observation data. Classroom observations are often used in program evaluation (e.g., Van den Branden, 2006), but rarely in large-scale evaluations of the size of Waterford. We observed 200 classrooms in Year 1 (treatment and comparison) and 100 classrooms in Year 2 (treatment only), each for four days; as a result, we collected invaluably rich data about the teachers' actual instructional practices. In treatment classrooms we also collected detailed information about individual students' use of the Waterford program through observations. This incorporation of extensive qualitative data allowed us to investigate not only the use of the program itself but the context in which it was implemented. In Slayton and Llosa (2005), we demonstrated how the use of qualitative methods allowed us not only to generate findings that were meaningful and useful to stakeholders, but also to improve the evaluation design, from one year to the next. For example, in Year 1 we observed treatment and comparison classrooms to determine the quality of teacher pedagogy. At the outset of the evaluation we were primarily concerned about our ability to isolate the effects from the Waterford program from those of the primary reading program. The data collected in Year 1 confirmed that in fact,

Open Court quality of pedagogy was comparable in the two groups and as a result, in Year 2, we were able to focus our resources on the observation of the treatment classrooms and the way the Waterford program was being implemented.

In summary, and critically, the Waterford evaluation design took into account the environment in which the program was being implemented, looking beyond outcomes in order to find explanations for those outcomes. The importance of getting at these explanations has also been stressed by Norris (2006) in relation to outcomes-based assessment in foreign language programs. He argues for a system:

> through which the findings about learning outcomes can be understood vis-à-vis the elements of FL programs that bring them about. Without such a system for processing the evidence and turning it into well-articulated recommendations and actions, SLO assessments will invariably end up as unused reports collecting dust on the department chair's shelf, if not in the recycle bin. (p. 580)

In the case of the Waterford evaluation, we sought explanations for the findings by considering multiple sources of data, using qualitative data to understand the classroom context, and using appropriate techniques to thoroughly analyze the data. The next section explains how we then turned those findings into 'well-articulated' recommendations.

2 Reporting the findings: The importance of presentation

A critical final step in conducting a useful evaluation is to present the findings in a clear and accessible manner so that stakeholders can take action. In the case of the Waterford evaluation, making the evaluation findings useful required that we take into consideration the political context within which we were presenting the findings and that we frame them accordingly. There was a substantial bias in support of the program from the outset, which made it critical that we find the appropriate ways to communicate findings we expected to be received unfavorably (and possibly dismissed). Beyond the political overtones, we had to consider the level of sophistication of our audience as consumers of research, and the nature of information that would allow the Board of Education and district staff to take appropriate action to meet students' needs. Bearing all of these factors in mind, we decided to: (1) provide explanations for the findings, (2) contextualize the findings in relation to the existing literature, and (3) make realistic recommendations.

a Providing explanations for the findings: In the Waterford evaluation, we attempted to avoid presenting the findings in a vacuum but rather to include detailed explanations for what we had observed. As in many school districts, LAUSD presented an environment in which some stakeholders had preconceived ideas about the effectiveness of a particular program. Thus, it was

important to demonstrate that the full case had been considered and a variety of factors had been taken into account, in order for the findings to be credible to members of the district community. The critical features of design discussed above, the investigation of the context, and the use of multiple types and sources of data, including extensive qualitative data and appropriate analytic techniques, made it possible to provide such detailed and careful explanations for the findings.

b Contextualizing the findings in relation to the existing literature: In addition to providing careful explanations for the findings, it was also important to situate the findings within the existing literature. First, including a literature review in the report may help the reader understand the larger context of program implementation. It was important for us to discuss the quality of other evaluations conducted on the Waterford program and explain how our evaluation was different, especially since our findings were contrary to these other evaluations and the inclination of the district's program staff (who were committed to continuing to implement the program for reasons that were not grounded in evidence of program effectiveness). Second, we also included references to the literature as part of the explanation of the findings that we did get.

The following excerpt from the Year 2 report illustrates how we attempted to explain the problem of overlap between Waterford and primary reading instruction, by demonstrating a thorough understanding of the school and classroom context and framing the discussion in terms of existing literature:

> The findings regarding student courseware usage strongly suggest that the Waterford courseware was not being successfully integrated into the majority of classrooms in the district during the 2002–03 school year. According to researchers (Kosakowski, 1998; North Central Regional Educational Laboratory, 1999), successful technology implementation requires integration into the already existing educational program. The technology cannot simply be added on to the existing program. As the Waterford program has been implemented to date, this integration has not occurred. Instead of establishing a policy or procedure for integrating the program into the existing educational program, the courseware was set up in classrooms and teachers were simply directed to have students use the courseware during the existing reading/language arts block. They were given little if any guidance on how to use the Waterford courseware as a supplement to the primary Open Court instruction that was already being provided.
>
> Moreover, the addition of technology to the educational program may require that adjustments to the length or organization of the school day occur (Kosakowski, 1998; North Central Regional Educational Laboratory, 1999). Yet, the amount of time dedicated to reading/language arts instruction was not altered to accommodate the inclusion of an additional component. For kindergarten, the total time dedicated to reading/language arts remained 90 minutes and the time dedicated to first grade reading/language arts remained 150 minutes. Because the program is intended to supplement the phonics portion of Open Court instruction, teachers were told not to use the courseware during the Sounds and Letters or Preparing to Read (Green Section) portion of Open Court instruction. In kindergarten, the Sounds and Letters portion of instruction can last up to 50 minutes and

> in first grade the Preparing to Read section can last as long as 75 minutes. Thus, kindergarten teachers had approximately 40 minutes and first grade teachers had approximately 75 minutes during which they could have students use the courseware during their reading/language arts block of instructional time. Yet, for all students in a kindergarten classroom to use the courseware, 105 minutes of the day are needed and for *all* first grade students to use the program, 210 minutes are needed. By not extending the reading/language arts block or the school day to accommodate the addition of the courseware to the educational program, the policy put teachers in the position of having to either 1) use the *supplementary* reading program during the *primary* reading program instruction; 2) use the courseware during other portions of the instructional day (e.g., during math instruction); 3) or not have all of their students use the courseware on every day. These time constraints also do not take into account other events that interfere with typical daily instruction like Banked Time Tuesdays, assemblies, parent teacher conference weeks, testing schedules, and a host of other events that further limit the amount of time available during the regular instructional day. Consequently, it was not surprising to find that approximately half of the kindergarten classrooms and two-thirds of the first grade classrooms were observed using the courseware during Green Section instruction. In other words, on any given day, between 20–31% of students missed all or part of their primary phonics instruction and instead were exposed to the Waterford courseware, the supplementary reading program. Thus, the program was not used to supplement the primary reading program. (Hansen, Llosa, & Slayton, 2004, pp. 63–64)

c Making realistic recommendations: Finally, we reasoned that the best way to make our recommendations useful was to ensure that they reflected a comprehensive understanding of the realities and constraints of an elementary school classroom in our district, as well as the policies and practices in place around reading instruction. We did not present alternatives that we knew would be politically unfeasible. We also provided a continuum of options for the district. We did not limit ourselves to only the 'best' alternative but instead provided a number of recommendations, each of which we knew could potentially improve the quality of reading instruction for students in these classrooms.

In light of the evaluation findings, at the end of Year 2 we made the following recommendations to the LAUSD Board of Education to ensure that the Waterford program effectively supplemented the primary reading program:

- Teachers should be provided with ongoing, in-depth professional development that not only instructs them about the mechanics of the technology, but also teaches them how to meaningfully use it (Hasselbring & Tulbert, 1991; Kinzer & Leu, 1997; North Central Regional Education Laboratory, 1999). This training is essential so that they can individualize the courseware in order to support primary reading instructional activities conducted during the phonics instructional components of *Open Court*.
- Consider ways to integrate the Waterford courseware content into the existing educational program so as to facilitate the Waterford program's use as a supplement to the primary reading program.

- Consider targeting use of the Waterford courseware to only those students who have the greatest need for additional reading instruction and practice.
- Consider extending the school day to allow for the additional time required to provide all students access to the courseware on a daily basis.
- Consider using the courseware during the already existing after school and Saturday school intervention, Intersession, and Summer school programs for students who need the supplementary instructional time instead of during the regular instructional day. (Hansen, Llosa, & Slayton, 2004, p. 70)

As a result of the evaluation, the district decided not to expand the implementation of the Waterford program beyond its initial schools. Schools and teachers who already had the program were provided with new guidelines regarding the use of the Waterford program during instructional time. The district also moved the use of the Waterford program to instructional time dedicated to intervention instead of during the primary instructional block. Finally, the findings contributed to the district's decision to move to an all-day kindergarten schedule.

III Conclusion

Carefully designed program evaluations can reveal whether programs adopted to help ELLs in fact have the intended effect. Furthermore, when the intended effect is not achieved, a well-designed evaluation should identify the issues that impede success. In the case of the Waterford program, the district adopted a program specifically to address the needs of schools serving large population of low-achieving students and ELLs. The program was intended to be used as a supplement to the regular reading program. It was believed that it would give students additional opportunity to practice readings skills learned in class. It was also believed that the program would be engaging to students and that it would be particularly helpful to ELLs in that it is an adaptive program that moves at the students' pace. It was believed that these program features would contribute to an increase in achievement for these students.

By conducting a carefully designed evaluation, we were able first to determine that the Waterford program was not having its intended effect and second to provide explanations for why, despite the presumed advantages of the program, it was unsuccessful in the specific context of the Los Angeles Unified School District. By collecting multiple types of data from various sources, using appropriate analytic techniques, and incorporating rich qualitative data, we were able to determine, that due to the constraints of the school day and district policies, the Waterford program was supplanting rather than supplementing the primary reading program and thus students were not getting additional instructional

time. We also discovered that teachers were not using the program to adapt instruction for individual students, and that even though the majority of students were engaged by the program, ELLs were not. The appealing visuals and animation did not make up for the fact that students with low levels of proficiency were unable to understand the instructions and purpose of the lessons.

Also, by providing a report that offered explanations for the findings in the context of school realities, as well as the broader literature, and by making realistic recommendations, the evaluation resulted in actions that might contribute to the ongoing improvement of education of ELLs. By carefully documenting the conditions under which the program implementation took place, the evaluation also offered a useful point of interpretable comparison to other districts, schools, and teachers who are considering adopting or implementing a similar program in their schools, thereby contributing to a much-needed research base on instructional interventions for such populations.

Notes

[1] In LAUSD, English-Only or EO refers to students whose parents reported speaking only English in the home when first enrolling them in the school. Those students whose parents report speaking a language other than English are tested using the California English Language Development Test (CELDT), the statewide standardized test of English proficiency, and if not found proficient, are identified as ELLs and placed into one of five English Language Development (ELD) levels. ELD Level 1 means that the student has minimal or no English, and ELD Level 5 means that students are almost ready to reclassify as Fluent English Proficient.

[2] Calendar type refers to whether the school is on a nine-month (or traditional) or year-round calendar. Students on a year-round calendar are placed in different *tracks* or schedules (three or four, depending on the school). Year-round schools exist to accommodate large numbers of students attending a school intended for a smaller student population.

IV References

Hansen, E. E., Llosa, L., & Slayton, J. (2004). *Evaluation of the Waterford Early Reading Program as a supplementary program in the Los Angeles Unified School District 2002–2003*. Planning, Assessment and Research Division Publication No. 177. Program Evaluation and Research Branch, Los Angeles Unified School District.

Hasselbring, T. S., & Tulbert, B. (1991). Improving education through technology: Barriers and recommendations. *Preventing School Failure, 35*(3), 33–37.

Kiely, R., & Rea-Dickins, P. (2005). *Program evaluation in language education*. Hampshire and New York: Palgrave Macmillan.

Kinzer, C., & Leu, D. J. (1997). Focus on research – The challenge of change: Exploring literacy and learning in electronic environments. *Language Arts, 74*(2), 126–136.

Los Angeles Unified School District (2005–2006). R30 Language Census Report 2005–2006. *Planning, Assessment and Research Publication* No. 313. Retrieved on May 18, 2007 from http://search.lausd.k12.ca.us/cgi-bin/fccgi.exe

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher, 33*(2), 3–11.

Merriam, S. (1998). *Qualitative research and case study applications in education*. Revised and expanded from *Case study research in education*. San Francisco, CA: Jossey-Bass.

Newton, X. A., & Llosa, L. (2008, April). *Towards a more accurate and nuanced approach to program effectiveness assessment: Hierarchical linear models (HLM) in K-12 program evaluation*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Norris, J. M. (2006). The why (and how) of assessing student learning outcomes in college foreign language programs. *Modern Language Journal, 90*(4), 576–583.

North Central Regional Educational Laboratory (1999). Critical Issue: Using Technology to Improve Student Achievement. Retrieved on May 18, 2007 from http://www.ncrel.org/sdrs/areas/issues/methods/technlgy/te800.htm

Oppenheimer, T. (2007). Selling Software: How vendors manipulate research and cheat students. *Education Next* (2), pp. 22–29. Retrieved on May 24, 2007 from http://www.hoover.org/publications/ednext/6017486.html

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd edn)*. Thousand Oaks, CA: SAGE Publications.

Slayton, J., & Llosa, L. (2005). The use of qualitative methods in large-scale evaluation: Improving the quality of the evaluation and the meaningfulness of the findings. *Teachers College Record*, *107*(12), 2543–2565.

Slayton, J., & Llosa, L. (2002). Evaluation of the Waterford Early Reading Program 2001–2002: Implementation and student achievement. *Planning, Assessment and Research Division Publication* No. 144. Program Evaluation and Research Branch, Los Angeles Unified School District.

Van den Branden, K. (2006). Training teachers: Task-based as well? In K. Van den Branden (Ed.), *Task-based language education: From theory to practice* (pp. 217–248). Cambridge: Cambridge University Press

Woodcock, R. W. (1987). *Woodcock Reading Mastery Tests – Revised*. Circle Pines, MN: American Guidance Service.