CRESST REPORT 759

Noelle Griffin
Jinok Kim
Youngsoon So
Vivian Hsu

EVALUATION OF THE WEBPLAY ARTS EDUCATION PROGRAM: FINDINGS FROM THE 2006–07 SCHOOL YEAR

JULY, 2009

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Evaluation of the WebPlay Arts Education Program: Findings from the 2006–07 School Year

CRESST Report 759

Noelle Griffin, Jinok Kim, Youngsoon So, and Vivian Hsu
CRESST/University of California, Los Angeles

July, 2009

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the WebPlay Award # U351D050036, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

EVALUATION OF THE WEBPLAY ARTS EDUCATION PROGRAM: FINDINGS FROM THE 2006–07 SCHOOL YEAR

Noelle Griffin, Jinok Kim, Youngsoon So, & Vivian Hsu
CRESST/University of California, Los Angeles

Abstract

This report presents results from the second year of CRESST's three-year evaluation of the WebPlay program. WebPlay is an online-enhanced arts education program for K–12 students. The evaluation occurred during the three-year implementation of the program in Grades 3 and 5 in California schools; this report focused on results from the second year of program implementation, 2006–07. Results show that WebPlay participation was significantly related to positive educational engagement/attitude. In terms of California Standards Test (CST) English Language Arts (ELA) scores, despite no overall WebPlay effects, a significant difference was found for limited English proficiency (LEP) students. The results support that a well-designed, theater-based education can improve student engagement; and that it may have academic benefits in language arts content, particularly for those students who are struggling with English proficiency.

Overview of Evaluation

The Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA conducted a 3-year external evaluation of the WebPlay program's implementation in a California school district, which began in the 2005–06 school year. The implementation focused on Grades 3 and 5, although across the program years a small number of classrooms at other grade levels were included. The evaluation used quantitative methodology to address the following core research questions:

1. Will WebPlay improve student performance on academic achievement, as measured by state tests, relative to the controls?
2. Will WebPlay improve student skills and development in areas outside of those covered by traditional academic achievement tests, such as arts and technology knowledge, academic engagement, academic self-esteem, and collaboration?
3. Do the effects on these outcomes persist across the different cohorts of students? To what extent do different site characteristics interact with treatment?
4. Does WebPlay show a differential effect for low-performing subgroups?
5. If the experimental condition produces higher levels of performance, what aspects of performance are most and least affected?

This report focuses on results from the second year of the evaluation, summarizing all findings from the 2006–07 school year. These results include both those focusing on

assessing the effects of WebPlay participation on student academic skills/attitudes as measured by a WebPlay student survey and program effects on state standardized test results.

WebPlay Program Background

WebPlay is an online-enhanced arts education instructional program that enables K–12 students to create and produce plays in collaboration with both a professional theater company and partner classes from different countries. WebPlay integrates arts education with the core curriculum, particularly in language arts, and is aligned to California state standards in Performing Arts, Literacy, Social Studies and Technology.

In addition to a quarterly professional development meeting and online resources for teachers and students, the participating classrooms receive support virtually from an international professional theater company in the development of "WebPlays," original theatrical productions created and performed by the students. Classrooms are partnered internationally as well, with the goal of each classroom incorporating aspects of their partner classroom's country's culture into their WebPlays.

WebPlay instructional activities consist of two weekly lessons during the regular school day: A whole-class lesson that is teacher led and a second session where the students work either individually or in small groups. Instruction each week involves both 'Research/Creation' (e.g., developing aspects of a play) and 'Communication' (i.e., targeted collaboration with partner classrooms or theatrical experts). The lessons incorporate hands-on theatrical participation in all aspects of play development and production, collaboration/teamwork, hands-on computer and multi-media instruction and experience, and activities that specifically link theater content to other skill areas (predominantly language arts, but also some math and social studies content).

Given these instructional activities that incorporate theatrical experience, cross-cultural connections, teamwork, and core content area skills, all within a creative learning environment, the program's instructional goals involve both those directly related to the theater-based experiences and other associated general outcomes. Specifically, the goals include increased theatrical knowledge and engagement, increased collaboration, increased general academic efficacy and engagement, and increased academic achievement in other content areas (such as literacy). These goals link conceptually with the 21st Century Learning Skills (Partnership for 21st Century Skills, 2008).

The program's resources, components, and goals are summarized in an implicit theory of action (see Figure 1).

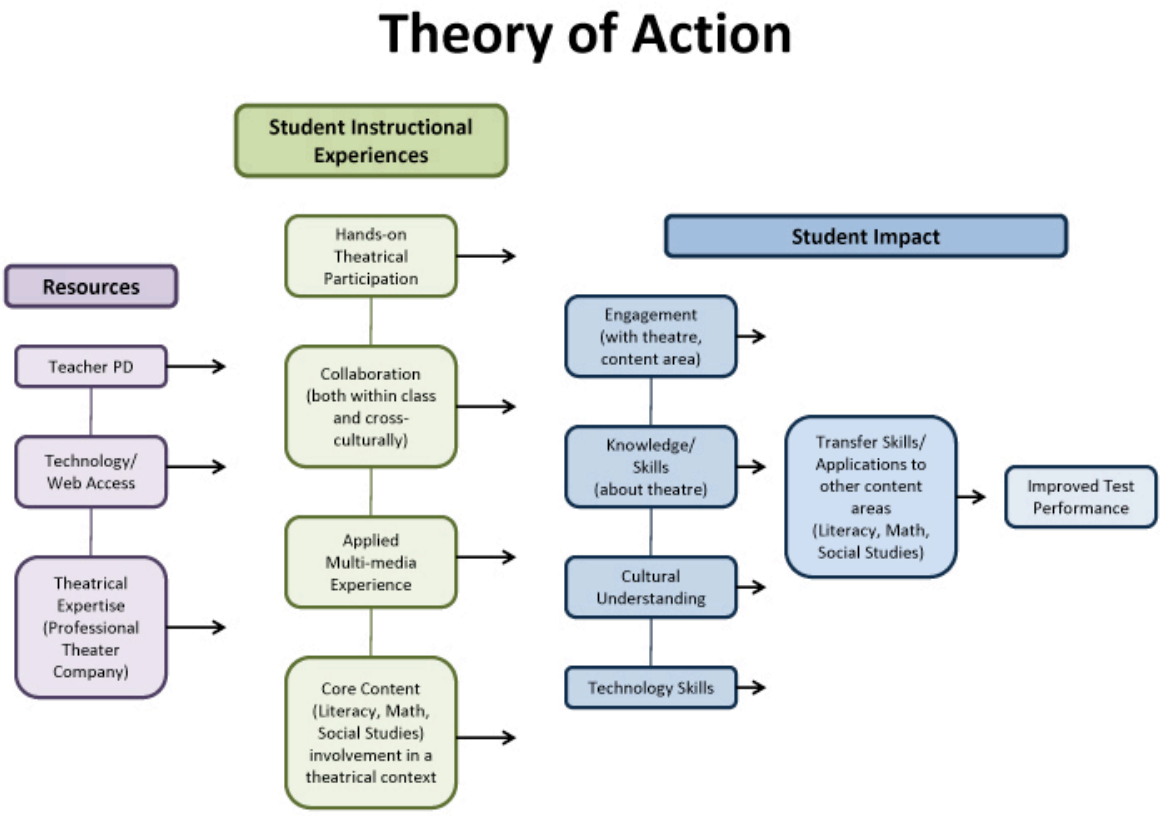

Figure 1. Theory of Action summarizing the program's resources, components, and goals.

Evaluation Methods

This report presents results from the second year of the WebPlay evaluation project, the 2006–07 school year. Below, we first present the general methodology for the WebPlay evaluation as a whole, and then information regarding sample and instrumentation specific to the 2006–07 evaluation year.

General Evaluation Framework

The original plan for the evaluation was to use an experimental, random assignment design, with teachers/classrooms within the participating district randomly selected to either implement WebPlay or serve as a comparison group. However, logistical issues, and the need within the district for WebPlay participation to be voluntary, made such a design impossible. The decision was then made to use a quasi-experimental design, matching the teachers who elected to participate in WebPlay to groups of teachers at schools with similar characteristics, who were then recruited to serve as a control/comparison group (see the Sample section for

additional details about this process in the 2006–07 school year). Although attempts were made to match the WebPlay and recruited classrooms in terms of teacher and student characteristics, and to control for classroom factors in the analysis process, there are inherent limitations in this quasi-experimental design (Green, Camilli, & Elmore, 2006). Particularly of concern is the issue of self-selection bias; that is, the teachers who chose to participate in WebPlay may share some common characteristics—such as a positive attitude towards arts education—that would make them systematically different in some way from those teachers in the comparison group.

Two broad types of outcomes were included in the evaluation to address the core research questions. First, student academic achievement was assessed using the California standards test (CST). Given the nature of the WebPlay curriculum and its emphasis on literacy skill such as reading, writing, and communication, the decision was made to focus on the Language Arts portion of the CST as the outcome for the present study. As with all California students, the 2006–07 WebPlay evaluation sample completed their CSTs in Spring 2007. Second, an assessment was developed to capture theatrical knowledge, engagement, and motivational outcomes not typically measured as part of regular classroom assessment. This survey was completed twice by both WebPlay and comparison students (at the beginning and end of the 2006–07 school year). More detailed information about each of these sets of outcomes is provided below.

Sample

For the 2006–07 school year, 18 schools in a California school district participated in the WebPlay program. As noted above, although the overall project focus was on Grades 3 and 5, exceptions were made by the WebPlay implementation team to include a small number of fourth and sixth grade classes. Participating classrooms from these 18 schools include 13 third grade classrooms, 2 fourth grade classrooms, 14 fifth grade classrooms, and 1 sixth grade classroom. A pool of 22 matched comparison schools were selected as well for possible inclusion in the 2006–07 comparison group, based on school characteristics. Teachers at all of the comparison schools were then invited to participate in the student survey portion of the evaluation (with the assumption that not all of the schools would agree to participate). A total of 12 comparison schools eventually agreed to participate in the survey, including 7 fifth grade classes, 8 third grade classes, 1 fourth grade class, and 1 mixed third/fourth/fifth grade class.

Ultimately, 590 students (424 WebPlay and 166 control) completed the student survey in the administration prior to the WebPlay implementation (pre-measure), whereas 354

students (255 WebPlay and 99 control) completed the student survey in the administration after the WebPlay implementation with data that could be linked to the pre-implementation surveys for analysis purposes. For the analysis of CST data, the decision was made to include all data available from all of the WebPlay and comparison classes, not just those students for whom we also had a linked set of pre–post surveys. There are some limitations inherent in this strategy, in that the survey and CST results presented below are not for identical sets of students. However, it was decided that the statistical power benefits of having a larger sample for the CST analyses, as opposed to only analyzing CST scores of the students who completed the survey, which was a smaller group of students (particularly for the control group), outweighed some of these limitations.

Student Survey

The WebPlay student survey was designed to assess several key outcomes drawn from the goals of the WebPlay program that are not addressed in the state achievement tests. The items were selected to reflect both content specifically covered in the WebPlay curriculum and more general attitudinal outcomes. Items were developed and selected around the following general areas:

1. Theatrical knowledge/awareness (e.g., understanding of basic theatrical concepts)
2. Internet knowledge/use (e.g., comfort and safety in using the Internet)
3. Theatrical engagement (e.g., interest in future participation in theater-based activities)
4. Academic engagement (e.g., general engagement in school)
5. Academic self-confidence (e.g., efficacy in school academics)
6. Collaborations/external connectedness (e.g., forming instructional connections both inside and outside of the classroom)

The survey consists of two parts: one with 11 selected response items reflecting the knowledge/awareness/use that is related to the WebPlay curriculum (i.e., numbers 1 and 2 just previously mentioned), and the other with 26 Likert-type items reflecting educational engagement, self-confidence and connectedness. Based on statistical factor analysis of the Likert-style items on the survey, three underlying factors were identified: (a) *Theater Engagement/Interest*, (b) *General Academic Confidence/Engagement*, and (c) *Use of External Connections*. Detailed factor analysis information about the survey is presented in the Results section.

State Standardized Test

All of the 2006–07 WebPlay and comparison schools in the evaluation participated in California's Standardized Testing and Reporting (STAR) Program in Spring 2007. As part of STAR, students complete the CST, a series of content-based tests focused on the California academic content standards, on a yearly basis.[1] At Grades 3 and 5, the primary grades of focus for the WebPlay program implementation, the CST completed included ELA, Math, and Science (for fifth grade only). Given that the content and goals of the WebPlay curriculum are most closely tied to language arts (reading, writing) the decision was made to focus on the CST Language Arts test as an outcome measure. In addition to CST results, several other background variables were included in the 2006–07 WebPlay state test data set for analysis purposes. These variables included student gender, ethnicity, limited English proficiency (LEP) status, socioeconomic status (i.e., free/reduced lunch status), Special Education status, gifted and talented program (GATE) status, and prior CST language arts score.

Results: Student Survey Data 2006–07

Three types of results based on the student assessment/survey are presented below. First, we provide factor analysis results regarding the empirical properties of the instrument. Second, we examine WebPlay effects for the selected response and Likert items as overall outcome sets using inferential statistical techniques. Third, we present an exploratory investigation of the contribution of individual factors to these outcomes.

Factor Analysis

In constructing a measure (i.e., surveys, tests, etc.), an important property to consider is whether it measures what it is intended to measure. This issue is also known as construct validity.

Researchers typically have hypothetical constructs based on underlying theory, that they want to measure in a survey or a test, and they develop one or more items to measure the hypothetical constructs. For example, targeted outcomes of the WebPlay program included both those relating to knowledge (e.g., about the theater) and general attitude (e.g., engagement, efficacy). In such settings, confirmatory factor analysis (CFA) provides a useful way of empirically testing the hypothesized item-construct relationships; that is, are items related to constructs as expected? In settings where CFA yields the results indicating

[1] For additional information about the CST technical quality and components, please see *California Standards Test Technical Report, Spring 2007 Administration* (Educational Testing Service, 2008), available at http://www.cde.ca.gov/ta/tg/sr/resources.asp.

significant expected relationships, it serves as positive empirical evidence of construct validity of the measure. For the present study, we conducted a CFA of the student survey instrument with the pre-test data collected from the WebPlay and comparison classrooms in Fall, 2006.

Among the 11 selected response items and 26 Likert-scale items on the student survey, we used 25 Likert items in CFA. The selected response items were set aside from the analysis because CFA is not the most suited method of analyzing such items. (Thompson, 2004). We also excluded one Likert-scale item (NA14), as descriptive statistics indicated that it was negatively correlated with the test total score and the magnitude of the negative correlation is not negligible (r = -.23), suggesting the item was not a good fit with the rest of the survey.

With 25 Likert-scale items, an initial exploratory analysis suggested that the various content areas covered in the survey were best explained by three higher-order factors. The three factors include:

- Theater Engagement/Interest
- General Academic Confidence/Engagement
- Use of External Connections

The items for each of the constructions are presented in Figure 2.

Factor 1: Theater Engagement/Interest

- I think about traveling to other countries.
- Movies are more interesting than plays.
- I would like to act in a play next school year.
- I would like to write a play next school year.
- I would like to direct a play next school year.
- I could write a play about a topic that interests me.

Factor 2: General Academic Confidence/Engagement

- I like school because I am learning a lot.
- I think the activities I do in school are boring.
- I work hard at school and as a result I learn more things.
- I share my experiences and ideas with other students in my class.
- I learn new things from students who live in other countries.
- I like it when I am assigned to write stories for school.
- I like to write about myself and my own experiences.
- I would be able to do a good job if I had to write a story for school.
- I would be able to do a good job if I had to write about myself for school.
- I can do well in school if I work at it.
- Most of the things they teach at school are very hard for me to learn.

Factor 3: Use of External Connections factor

- I use the Internet to get information for school projects or homework.
- I use the Internet to learn more about things on my own.
- I use the Internet to work with other students on school projects or homework.
- I work together with other students on assignments.
- I learn something from other students in my class.
- I use the internet to learn more about other countries/cultures.
- I share my experiences and ideas with other students in my class.
- I learn new things from students who live in other countries.
- It is easy for me to use the Internet to get information for a school project.

Figure 2: Items per factor.

After establishing these three factors based on both the exploratory analysis and theoretical considerations, we conducted a CFA with three factors measured by multiple indicators (i.e., items) as specified above. This initial CFA model was modified several times based on both theoretical and empirical considerations. Specifically, from a theoretical standpoint we checked to ensure that the modified model does not substantively contradict the underlying theory and focus of the WebPlay Program. At the same time, empirical considerations addressed include modification indices (Jorskog & Sorbom, 1989; Muthen, 1998) and how to better modify the model to fit most closely to the data.

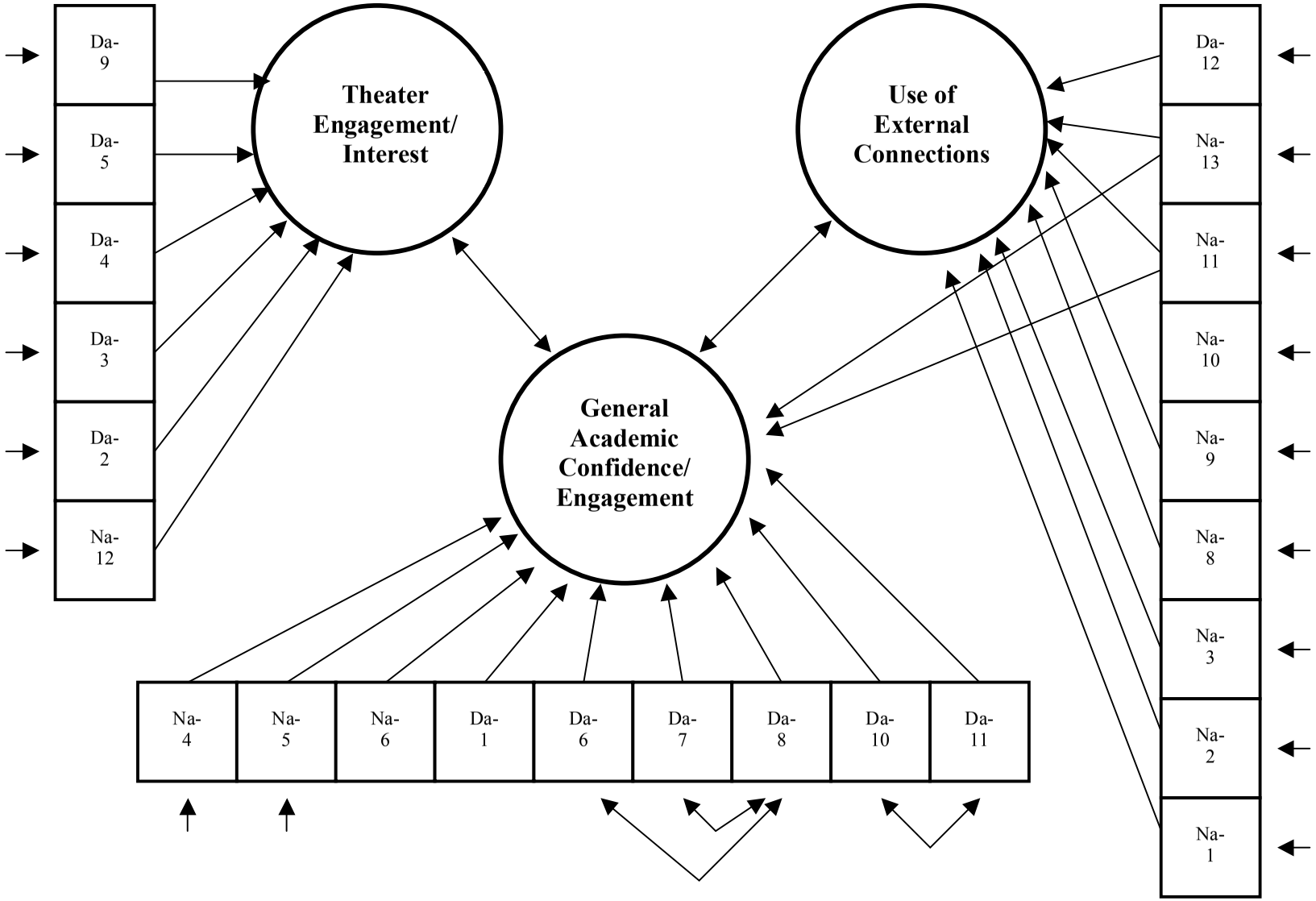

Figure 3. Final CFA model.

The final CFA model is represented graphically in Figure 3. Three circles represent the factors; squares represent the items or indicators; long one-way arrows indicate item-construct relationships; short one-way arrows indicate residual errors of indicators; and two-way arrows indicate covariances among the factors or among the items.

Table 1

Tests of Model Fit for WebPlay Student Survey

Method	Fit
CFI	0.858
TLI	0.888
RMSEA (Root Mean Square Error Of Approximation)	0.065
SRMR (Standardized Root Mean Square Residual)	0.068
WRMR (Weighted Root Mean Square Residual)	1.389

Note. CFI = Comparative fit index, TLI = Tucker Lewis index.

We used software Mplus (Muthen & Muthen, 2004a; 2004b) for model estimation which provides several indices of model fit to the data. According to Mplus technical appendices (Muthen, 1998), TLI > .95, CFI >.95, RMSEA <.06, SRMR <.08, WRMR <.90 may indicate well-fitting models. Table 1 shows the indices from the CFA model shown in Figure 3. As can be seen, the conclusions we can draw based on the suggested cut-offs are mixed. RMSEA is about at the borderline from the cut-off, SRMR indicates a good fit, while CFA, TLI, and WRMR are below or above their respective cut-offs. Given these mixed results, and the desire to find a model that is the best fit both empirically and theoretically, further investigation was warranted. Along these lines, Table 2 represents the estimates of the final model.

Table 2

CFA Results for WebPlay Student Survey

Items by factor	Estimates	*S.E.*	Est./*S.E.*	Std. Est.
Factor Loadings				
Theater engagement/interest (Factor 1)				
NA12	1.00	0.00	0.00	0.25
DA2	-1.76	0.45	-3.93	-0.44
DA3	2.84	0.64	4.43	0.71
DA4	2.90	0.66	4.40	0.72
DA5	2.50	0.55	4.59	0.62
DA9	1.95	0.43	4.58	0.49
General academic confidence/engagement (Factor 2)				
NA4	1.00	0.00	0.00	0.63
NA5	-0.70	0.09	-7.65	-0.44
NA6	0.98	0.09	10.54	0.62
DA1	1.12	0.10	11.07	0.71
DA6	0.66	0.09	7.09	0.42
DA7	0.77	0.09	8.62	0.48
DA8	0.53	0.09	6.06	0.33
DA10	0.58	0.10	6.10	0.37
DA11	-0.11	0.09	-1.24	-0.07
NA11*	0.55	0.09	6.35	0.35
NA13*	0.39	0.10	4.12	0.25
Use of external connections (Factor 3)				
NA1	1.00	0.00	0.00	0.59
NA2	0.80	0.10	7.74	0.47
NA3	1.12	0.11	10.11	0.66
NA8	0.89	0.10	8.94	0.52
NA9	0.69	0.10	6.96	0.41
NA10	0.91	0.10	9.49	0.54
NA11	0.43	0.09	4.56	0.25
NA13	0.56	0.10	5.81	0.33
DA12	0.68	0.11	6.42	0.40

(*table continues*)

Table 2 (*continued*)

Factor variances					
	THEATER	0.06	0.03	2.25	1.00
	GENERAL	0.40	0.06	6.89	1.00
	EXTERNAL	0.35	0.05	6.83	1.00
Factor covariances					
	Factor 1 and 2	0.09	0.02	3.92	0.60
	Factor 1 and 3	0.07	0.02	3.59	0.49
	Factor 2 and 3	0.13	0.03	4.74	0.34
Residual covariances					
	DA10 and DA11	-0.33	0.06	-6.03	-0.33
	DA6 and DA8	0.27	0.05	5.66	0.27
	DA7 and DA8	0.25	0.04	6.00	0.25

Note. *Item loads on more than one factor.

The estimates and relationships between the items in the fitted model agree very well with the hypothesized relationships between constructs. Factor loadings were all consistent with the hypothesized relationships of the individual items to each of the constructs.. In addition, items that are negatively scaled (i.e., items with higher scores indicate lower levels in factors) show significant loadings that are negative.

In summary, all of the findings (factor loadings) correspond well to the hypothesized relationships between items and factors. Also, the three factors show moderate to moderately high correlations with each other, which is also expected given the nature of the constructs. *Theater Engagement/Interest* and *General Academic Confidence/Engagement* are most highly correlated (r =.60); and *Use of External Connections* is moderately correlated with both factors (r = .49 and .34, respectively). Thus, the CFA results provide positive evidence that supports the construct validity of a large part of the student survey, consisting of 25 Likert-scale items. First, all the items are significantly loaded on the proposed underlying factors. Second, negative significant loadings show up in reversely scaled items. Third, some empirically-based revisions to the model from the initial hypothesized constructs make sense substantively and do not contradict the original design of the survey. These findings thus support the continued use of the survey for this project, as well as for use in similar arts-based program evaluations.

Analyses of Overall WebPlay Outcomes

We used Hierarchical Models (HMs) or multilevel models to estimate the differences in outcomes of interest between students who are in WebPlay participating schools and students who are in control schools. HMs are particularly useful in analyzing "nested" data that includes multiple levels of groups—for example, children who are in classrooms, which are in schools, which are in districts, etc. In terms of the WebPlay Program, students are nested within schools, which are either included in WebPlay or control conditions. Given this nested structure of the data, HMs are an appropriate approach to draw sound inference about the differences in outcomes between WebPlay and control schools (see Goldstein, 2003; Raudenbush & Bryk, 2002, for a more detailed description of the use of HMs).

Using HM techniques, we examined two general outcomes from the survey. One outcome is constructed from all selected response items in the survey, which intend to measure Theatrical Knowledge/Awareness and Internet Knowledge/Use. The outcome is the number of items that a student got right among the 11 items. The other outcome is constructed from Likert-type items in the survey, which intend to measure Theatrical engagement, Academic Engagement and Self-Confidence, and Collaborations/External Connectedness. The outcome is the mean of 25 Likert-type items. The coefficient alpha for these items was 0.78. In what follows, the first outcome will be referred to as Knowledge outcome, and the second outcome will be referred to as Engagement outcome. Two-level HMs, in which students are nested within schools, are applied to each of the two outcomes.

Specifically, the following two-level HM was used:

Level-1 or student-level model:

$$y_{ij} = \beta_{0j} + \beta_{1j}(\text{Pretest}_{ij} - \overline{\text{Pretest}}_{..}) + r_{ij}, \qquad r_{ij} \sim N(0, \sigma^2),$$

Level-2 or school-level model:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} WebPLay_j + u_{0j}, \qquad u_{0j} \sim N(0, \tau_{00}),.$$

$$\beta_{1j} = \gamma_{10}.$$

Figure 4. Two-level HM for WebPlay student survey data 2006–07.

In the Figure 4 equation, y_{ij} is the outcome score for student *i* in school *j* (i.e., either the Knowledge outcome score, or the Engagement outcome score); $Pretest_{ij}$ is the score that student *i* in school *j* got in the pre-implementation administration in the same scale as the corresponding outcomes; and $WebPlay_{ij}$ indicates whether school *j* is a WebPlay participating school.

Although the control group schools are carefully matched in terms of numerous school characteristics, due to its quasi-experimental nature the study lacks a true randomization procedure. Therefore, the students in WebPlay schools may be systematically different from the students in control schools on various characteristics prior to the implementation of the WebPlay curriculum. In such quasi-experimental settings, study designs with both pretest and posttest, such as the one employed in this evaluation, help strengthen the inferences concerning the differences between the two groups.

Table 3 shows the descriptive results of the pretest measures by WebPlay vs. the comparison group. In the Knowledge pretest, the control group mean is higher by 0.5 points, which is about one-third of one standard deviation of the variable. On the other hand, in the Engagement pretest, the WebPlay group mean is higher by 0.17 points, which is about a quarter of one standard deviation. Independent *t*-tests are conducted for both pretests as a preliminary analysis, and for both measures, WebPlay and control groups show significant differences (*p* values less than 0.001 and 0.02 respectively).

Table 3

Descriptive Statistics of Two Pre-Implementation Measures by Treatment Group

Measures	WebPlay			Control		
	(*N* = 424)			(*N* = 166)		
	N	Mean	Std Dev	*N*	Mean	Std Dev
Pretest: Knowledge	423	6.56	(1.54)	164	7.06	(1.41)
Pretest: Engagement	399	4.00	(0.71)	158	3.83	(0.76)

In an effort to control for such pre-existing differences, the employed HM controls for the pretest survey scores. As can be seen in the Figure 4 equation, the Pretest variable is centered on its grand mean. By virtue of this grand mean centering, β_{0j} is the mean of posttest scores for classroom *j*, adjusted for the differences in pretest scores (see Raudenbush & Bryk, 2002, Chapters 2 & 5). Table 2 presents the results from the HM analysis for the Knowledge outcome. Holding constant the pre-implementation score in the same scale in the model, participating WebPlay implementation for the first year did not significantly relate to the overall Knowledge outcome (*p value* = 0.4; statistical significance is generally denoted by *p value* = 0.05 or smaller).

Table 4 also shows that the expected posttest Knowledge score is 7.10 for a typical student in the study sample, which means that a typical student got seven items right out of eleven Knowledge items in the post-implementation survey. The pretest score is positively associated with the posttest score, with an increase of 0.2 points when a student got one more item right in the pre-implementation survey. The adjusted school means varied significantly across schools.

Table 4

HM Results for Knowledge Outcome

Fixed effects	Coefficient	*SE*	*p* Value
Intercept	7.10	0.21	<.0001
WebPlay	-0.22	0.26	0.41
Pretest	0.20	0.06	0.00
Random effects	**Variance component**	***SE***	***p* Value**
Adjusted means	0.18	0.10	0.03
Student residual	2.12	0.17	<.0001

Table 5 presents the results from the HM analysis for the Engagement/Attitude outcome. Holding constant the pre-implementation score in the same scale in the model, participating WebPlay implementation for the first year was significantly related to the Engagement/Attitude outcome (*p value* = 0.02; statistical significance is generally denoted by *p value* = 0.05 or smaller). Participating in WebPlay was associated with the 1.9 points increase in the posttest score, which is about a quarter of its standard deviation. This indicates a small to medium effect size (e.g., Cohen, 1988).

Table 5 also shows that the expected posttest Engagement/Attitude score is 3.80 for a typical student in the study sample, which means that a typical student self-report about a medium level in the 6-point scale items of the post-implementation survey. The pre-implementation score in the same scale is positively associated with the posttest score, with an increase of 0.6 points when a student reported one higher point in the 6-point scale in the pre-implementation survey. Unlike the Knowledge outcome, the adjusted school means did not vary significantly across schools.

Table 5

HM Results for Engagement/Attitude Outcome

Fixed effects	Coefficient	*SE*	*p* Value
Intercept	3.80	0.07	<.0001
WebPlay	0.19	0.08	0.02
Pretest	0.57	0.05	<.0001
Random effects	**Variance component**	***SE***	***p* Value**
Adjusted mean	0.00	0.01	0.43
Student residual	0.36	0.03	<.0001

A significant and positive increase associated with the WebPlay group, as compared to the control group, is notable after only a single year of implementation.

Exploratory Investigation of Individual Scales/Constructs

Given the significant results for the Engagement/Attitude items as a whole, we examined specific areas of the outcomes that may be more or less related to the WebPlay implementation, both in terms of Knowledge and Engagement constructs. We divided the Knowledge outcome into two areas: *Theatrical Knowledge/Awareness* and *Internet Knowledge/Use*. We also divided the Engagement/Attitude outcome into the three factors identified through the CFA: *Theater Engagement/Interest*, *General Academic Confidence/Engagement*, and *Use of External Connections*. For these five sub-areas, we examine which areas may have been more affected by the WebPlay implementation than others. This probing is intended to be exploratory, and limited only to descriptive analysis. We do not present inferential statistical tests to avoid possible complications and misinterpretations regarding scaling, missing data, and multiple comparisons.

Figure 5 shows the average scores of each treatment group (WebPlay and control) in two sub-areas, Theatrical Knowledge and Internet Knowledge. It also shows the scores both in pre-implementation and post-implementation: 1 indicates pre- and 2 indicates post-implementation in the X axis. Thus the lines between the pre- and post-implementation reflect increment or decrement in average scores between the pre- and post-implementation. Note that two sub-area scores are not comparable because they are in different scales. Visual and exploratory comparison is warranted between pre- and post-implementation, and between WebPlay and control groups in the same area.

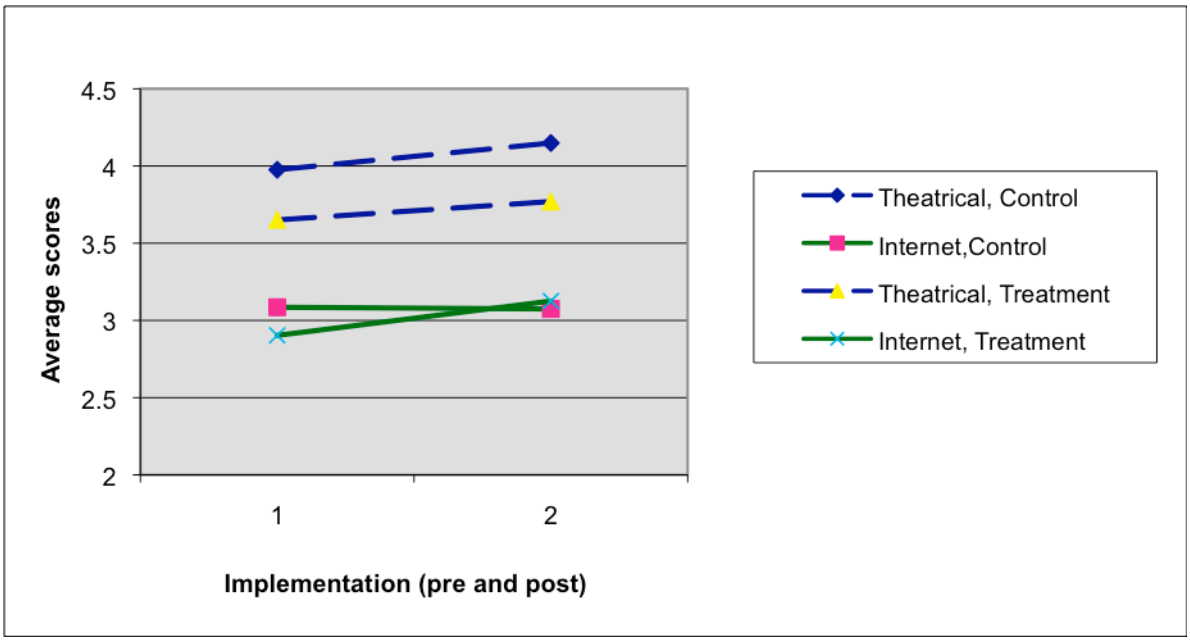

Figure 5. Knowledge outcome sub-area scores by treatment group.

The dotted lines in Figure 5 represent changes in the Theatrical Knowledge sub-area. It seems that both the WebPlay group and the control group increased slightly in their average scores. Because the WebPlay group does not differentially increase in this area, this may not be the sub-area that WebPlay affected the most.

The plain lines in Figure 5 represent changes in the Internet Knowledge sub-area. Although the control group tended to stay in the similar level, the WebPlay group tended to increase slightly in this area. However, the change is only about 0.1, which may not be meaningful both substantively and statistically. Thus, the exploratory analysis suggests that both Knowledge sub-areas are not affected by the WebPlay implementation after 1 year of participation.

Looking more specifically at each of the Knowledge items, Table 6 presents the percentage of WebPlay and control students who answered the selected response items correctly for both the pre- and post- implementation student surveys.

Table 6

WebPlay Student Survey 2006–07: Percentage of Students with Correct Answers to True/False Items at Pre and Post Implementation

Label	Control % correct		WebPlay % correct	
	Pretest	Posttest	Pretest	Posttest
1. Establishing a conflict should happen in the middle of a written play.	47.8 (*N* = 161)	63.8 (*N* = 94)	61.5 (*N* = 418)	68.6 (*N* = 236)
2. The location or place ("setting") of a play is where the play is being presented.	81.1 (*N* = 164)	72.3 (*N* = 94)	73.9 (*N* = 422)	79.5 (*N* = 239)
3. Blocking generally occurs after a play is performed for an audience.	39.1 (*N* = 161)	28.7 (*N* = 94)	46.4 (*N* = 412)	42.9 (*N* = 231)
4. The way you move your body is important when developing a character for a play.	92.5 (*N* = 160)	94.7 (*N* = 94)	81.0 (*N* = 420)	91.6 (*N* = 238)
5. The work of the design department could help show how wealthy a character is.	62.9 (*N* = 159)	65.6 (*N* = 93)	55.6 (*N* = 417)	59.5 (*N* = 237)
6. When you do improvisation, you do not read lines from a script.	47.5 (*N* = 160)	53.3 (*N* = 92)	47.7 (*N* = 417)	55.1 (*N* = 234)
7. A major difference between film and theater is that in theater there is a live audience.	72.3 (*N* = 159)	68.8 (*N* = 93)	67.9 (*N* = 420)	66.9 (*N* = 236)
8. If I am talking with someone on the Internet who says she/he goes to my school, I would talk to my parents before deciding if I would find myself with that person at school.	88.4 (*N* = 164)	84.0 (*N* = 94)	83.4 (*N* = 421)	86.6 (*N* = 238)
9. E-mails as well as the documents sent via e-mail ("attachments") can have viruses.	69.4 (*N* = 160)	67.7 (*N* = 93)	67.0 (*N* = 412)	70.2 (*N* = 235)
10. The Internet lets you send messages to other people in the United States, but not people in Europe.	25.6 (*N* = 160)	23.9 (*N* = 92)	27.6 (*N* = 421)	22.6 (*N* = 239)
11. If I am chatting on the Internet with someone at 8:30 in the morning using the school computer, it will be 8:30 in the morning where they are too no matter which part of the world they live in.	18.1 (*N* = 160)	18.1 (*N* = 94)	29.5 (*N* = 420)	20.1 (*N* = 239)

These results are purely descriptive in that they include all pre- and post-responses collected (i.e., the groups are not identical at pre- and post-). With these limitations in mind, the descriptive findings for these items do not present any consistent trends. For some items there was an increase in the percentage of correct responses for the WebPlay group only (e.g., Item 2, 8, 9), for some items there was an increase for both WebPlay and control

groups (e.g., Item 1, 5, 6) and for some there was an increase for neither (e.g., Item 3, 10, 11). These observed differences are descriptive in nature and thus cannot be assumed as evidence of WebPlay impact, and could be due to differences in group composition at pre- and post-implementation.

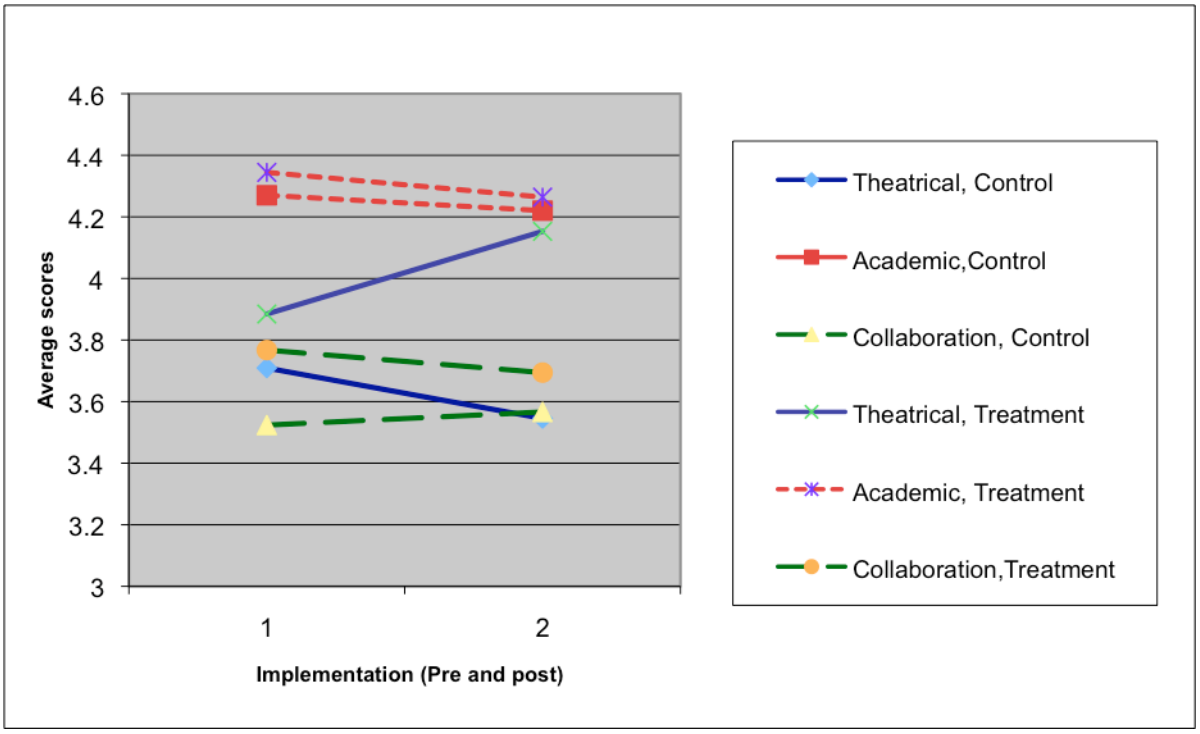

Figure 6. Attitude outcome sub-area scores by treatment group.

Figure 6 shows the average scores of each treatment group (WebPlay and control) in three sub-areas: Theatrical Engagement, Academic Engagement/Self-confidence, and Collaboration/External Connectedness. It also shows the scores both in pre-implementation and post-implementation: 1 indicates pre- and 2 indicates post-implementation in the X axis. Thus the lines between the pre- and post-implementation reflect increment or decrement in average scores between the pre- and post-implementation.

The small dotted lines represent changes in the sub-area of Academic Engagement/Self-Confidence. For both WebPlay and control groups tend to decrease slightly in the area. Because the two groups show an identical pattern of change, the area appears to be unrelated to the WebPlay curriculum. The Collaboration/External Connectedness do not show much change between pre- and post-implementation for both WebPlay and control groups, as represented by the longer dotted lines.

The solid lines represent changes in Theatrical Engagement. The change pattern is unique, because the WebPlay group tends to increase substantially, whereas the control group tends to decrease. The increase of the WebPlay group in the Theatrical Engagement area may be due to the WebPlay implementation. However, why the decrease of the control group occurred is unclear. It may be developmental change for children in upper elementary grades. It may be due to non-random missing of the responses. Or, it may be due to other reasons, such as overall decreased student motivation and interest in test-taking at the end of the school year.

Table 7 presents average pre- and post-scale scores for each of the Engagement/Attitude factors.

Table 7

WebPlay Student Survey 2006–07: Average WebPlay and Comparison Group Scores on Attitude/Engagement Factors Pre- and Post Implementation

	Control		WebPlay	
Factor	Pre mean (*SD*)	Post mean (*SD*)	Pre mean (*SD*)	Post mean (*SD*)
1. Theatrical engagement	3.7(1.2) ($N = 158$)	3.5/1.0 ($N = 94$)	3.9/1.1 ($N = 402$)	4.2/1.1 ($N = 238$)
2. General academic engagement	4.3/0.9 ($N = 161$)	4.2/0.8 ($N = 94$)	4.3/0.8 ($N = 405$)	4.3/0.8 ($N = 238$)
3. Collaboration/external connections	3.5/0.9 ($N = 164$)	3.6/0.9 ($N = 94$)	3.8/0.9 ($N = 414$)	3.7/0.9 ($N = 240$)

2006–07 Evaluation Findings: State Test Data (CST) Results

Based on the nested nature of the data, as described above, we again used HMs to estimate the differences in the state ELA achievement test outcome (CST ELAs) between students who are in WebPlay participating classrooms and students who are in control classrooms.

Although the 2006–07 sample included some classes covering Grades 4 and 6, our HM analysis focuses on only the two core grades in the study, which are Grades 3 and 5. This decision was based on the fact that this WebPlay implementation was targeted for Grades 3 and 5 and that, given the project was primarily implemented in those grades, we have sufficient data at those grades to adequately balance or statistically control for pre-implementation characteristics. The small number of students at the other grade levels would

not allow for such rigorous analysis, and the nature of the CSTs does not allow for combining grade levels for purposes of analysis. In the following sections we first describe the basic background characteristics of the WebPlay students and control students in the CST data set, and then we present the specific HM results separately for Grade 3 and Grade 5.

Student Background Characteristics

In addition to Spring 2007 CST ELA scores, we analyzed basic background variables such as student gender, socioeconomic status (i.e., free/reduced lunch status), Special Education status, GATE status, ethnicity, and LEP status in addition to prior year's (2006) CST ELA scores (note that 2007 CST ELA score will be addressed in the HM analysis sections below). Descriptive analyses and simple *t*-tests were used to compare WebPlay and control group students on these background variables to determine if the groups differed in anyway.

In the third grade, the WebPlay and the control groups are found comparable in terms of their prior CST ELA scores ($t = 0.790$, $df = 386$, $p = 0.430$). Students in the control and WebPlay groups are also similar in terms of student gender, Special Education status, and GATE status. However, more students in the WebPlay group are receiving free/reduced-priced lunch than those in the control group (95% vs. 88%, respectively).

In terms of the composition of ethnic groups, the percentages of students of each ethnicity are mostly comparable in the control and WebPlay groups, except for Asian and Hispanic representation. Specifically, there was a greater percentage of Asians in the control group than in the WebPlay group (11% vs. 1%). The pattern is reversed for Hispanics students, with 85% Hispanic students in the WebPlay group, and 77% in the control group. There are also a greater percentage of LEP students in the WebPlay group (62%) than in the control group (53%).

As with the third grade students, fifth grade students' prior year performances on the CST ELA is not significantly different between the control and the WebPlay groups ($t = 1.734$, $df = 617$, $p = 0.083$). The groups also do not differ significantly in terms of background variables such as free/reduced-priced lunch, gender, GATE status, and Special Education status. However, as with third grade, the WebPlay group is more Hispanic-dominant and the percentage of Asian students is higher in the control group than in the WebPlay group. There are also more fifth grade LEP students in the WebPlay group (46%) than in the control group (34%).

Overall, the WebPlay and control groups did not differ significantly on a majority of students' background characteristics. However, there were consistent differences in both

Grades 3 and 5 in terms of student ethnicity and LEP status, with the WebPlay group having more LEP and Hispanic students and the control group more Asian students. It should be noted that the WebPlay group had a higher percentage of groups that have been associated with achievement gaps on multiple California student academic outcome measures (Baker, Griffin, & Choi, 2008).

HM Analysis and Results for Third Graders

In considering the nesting of the data, the following two-level HM model was specified to analyze differences between WebPlay and comparison students. The Level-1 or student-level model is

$$y_{ij} = \beta_{0j} + \beta_{1j}\left(PRE_{ij} - \overline{PRE_{.j}}\right) + \beta_{2j}\left(IFEP_{ij} - \overline{IFEP_{.}}\right) + \beta_{3j}\left(RFEP_{ij} - \overline{RFEP_{.}}\right) + \beta_{4j}\left(LEP_{ij} - \overline{LEP_{.}}\right) + \beta_{5j}\left(GIFTED_{ij} - \overline{GIFTED_{.}}\right) + r_{ij}, \qquad r_{ij} \sim N(0, \sigma^2_{WebPLay})$$

,

and the Level-2 or school-level model is:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} WebPLay_j + \gamma_{02}\overline{PRE_{.j}} + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00}),$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \quad u_{1j} \sim N(0, \tau_{11}),$$

$$\beta_{4j} = \gamma_{40} + \gamma_{41} WebPLay_j,$$

$$\beta_{kj} = \gamma_{k0}, \quad k = 2,3,5.$$

In the equation above, y_{ij} is the ELA score for student i in school j; PRE_{ij}, $IFEP_{ij}$, $RFEP_{ij}$, LEP_{ij}, and $GIFTED_{ij}$ are respectively the ELA scores for the previous academic year, whether student i in school j was Initial Fluent English Proficient (IFEP), Reclassified Fluent English Proficient (RFEP), LEP in terms of English language proficiency (ELP) status, and whether classified as gifted or talented. $WebPlay_j$ is a binary variable indicating whether school j is a WebPlay participating school.

Because all student pre-implementation characteristics are centered on their respective means (this is often referred to as grand-mean centering in the HM literature) except student pretest scores, the Level-1 intercept is the average posttest score for school j, adjusted for ELP status and gifted status but unadjusted for pretest scores. At Level 2, the intercept is a function of the WebPlay indicator and school average pretest scores. Thus, the key parameter in this model is the coefficient of WebPlay, which represents the expected differences in the ELA scores, holding constant various student characteristics as well as school average pretest scores.

It is notable that the coefficient of LEP status, which captures the expected difference between LEP students and their English-only peers in ELA scores for school *j*, controlling for all other predictors and the WebPlay or control membership, is specified as a function of the WebPlay membership indicator with no random component attached. This means that the difference between LEP and English-only students does not vary across schools, but may be different for students receiving WebPlay and students in the control schools. Also, when significant, this implies interaction between the WebPlay program and LEP status. Because intact schools are assigned to either WebPlay or control conditions, this interaction becomes a cross-level interaction between treatment and student characteristics.

Table 8

Results from HMs Predicting Performance in CST in ELA for WebPlay and Control Students: Grade 3

Fixed effect	Estimate	*SE*	*Df*	*t* value	*p* value
Model for class means					
Adjusted grand mean	343.51	3.81	18	90.06	<.0001
WebPlay contrast	-6.61	4.71	18	-1.40	0.18
Class pretest average	0.63	0.09	18	6.69	<.0001
Average within-class slope					
Pretest/posttest	0.61	0.04	569	16.59	<.0001
ASIAN/posttest	15.45	4.42	569	3.50	0.00
RFEP/posttest	5.01	2.55	569	1.97	0.05
LEP/posttest in WebPlay	6.51	4.43	569	1.47	0.14
LEP/posttest in control	-9.47	3.49	569	-2.71	0.01
Gifted/posttest	12.26	2.84	569	4.32	<.0001
Random effect	Estimate	*SE*		*Z* value	*p* value
Between class					
Variance in adjusted means	83.88	33.79		2.48	0.01
Variance in pre/post slopes	0.01	0.01		2.16	0.02
Within class					
Variance in WebPlay	442.88	34.92		12.68	<.0001
Variance in control	454.87	42.33		10.75	<.0001

Note. HM = Hierarchical Model, CST = California Standards Test, IFEP = Initial Fluent English Proficient, RFEP = Reclassified Fluent English Proficient, LEP = Limited English Proficient.

Table 8 presents the results from the third grade HM analysis. Holding constant all the other variables in the model, participation in the WebPlay program did not significantly

relate to student ELA achievement for the overall population (p = 0.22; statistical significance is generally denoted by p = 0.05 or smaller). However, for students with LEP status, participating in the WebPlay program was significantly related to the ELA achievement (p = 0.05). Specifically, the expected difference in the ELA achievement between WebPlay and control students was about 13 points for typical students in terms of various pre-implementation characteristics in typical schools in terms of pretest average scores, whereas the expected difference between WebPlay LEP students and control LEP students was about 27 points, which is statistically significant. As the sample standard deviation of the outcome (i.e., ELA achievement scores) is 56 points, the difference of 27 points reaches almost ½ standard deviation, which is considerable.

All the other variables that are included in the final model are significantly related to student ELA achievement. Not surprisingly, student prior CST ELA score and school prior CST ELA average were positively related to the outcome of interest, (i.e., the 2006–07 CST ELA scores). Language status was also a significant variable in the CST ELA score predictor. Compared to English Only (EO) students, Initial Fluent English Proficient (IFEP) and Redesignated Fluent English Proficient (RFEP) students, on average, performed better, holding constant all the other variables in the model, by 9 and 15 points respectively. Also, students who were classified with gifted or talented status tended to perform better by 20 points, holding constant all other variables.

HM Analysis and Results for Fifth Graders

Similarly to the model applied to third graders, in considering the nesting of the data, the following two-level HM model was specified to analyze differences between WebPlay and comparison students. The Level-1 or student-level model is

$$y_{ij} = \beta_{0j} + \beta_{1j}(PRE_{ij} - \overline{PRE_{.j}}) + \beta_{2j}(ASIAN_{ij} - \overline{ASIAN_{.}}) + \beta_{3j}(RFEP_{ij} - \overline{RFEP_{.}}) + \beta_{4j}(LEP_{ij} - \overline{LEP_{.}}) + \beta_{5j}(GIFTED_{ij} - \overline{GIFTED_{.}}) + r_{ij}, \quad r_{ij} \sim N(0, \sigma^2_{WebPlay})$$

,

and the Level-2 or school-level model is

$$\beta_{0j} = \gamma_{00} + \gamma_{01} WebPLay_j + \gamma_{02}\overline{PRE_{.j}} + u_{0j}, \quad u_{0j} \sim N(0, \tau_{00}),$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \quad u_{1j} \sim N(0, \tau_{11}),$$

$$\beta_{4j} = \gamma_{40} + \gamma_{41} WebPLay_j,$$

$$\beta_{kj} = \gamma_{k0}, \quad k = 2,3,5.$$

As can be seen, the HM specified for fifth graders is very similar to third graders, based on the assumption that pre-implementation characteristics and the WebPlay program will relate to student ELA achievement similarly in both grades. The changes between the two models are with regard to the student pre-implementation characteristics that are included as predictors, which may not be the primary focus of this study. These changes were based on empirical results about the relationships between background variables. The IFEP status was no longer a significant predictor of the student ELA achievement for fifth graders, and, unlike third graders, being Asian turned out to be a significant predictor for fifth graders controlling for all the other student characteristics and treatment status. Other than these changes, the same explications about the model and the interpretation of parameters for the third graders apply to this model as well.

Holding constant all the other variables in the model, participating in the WebPlay program did not significantly relate to student ELA achievement neither for overall population (p = 0.18), nor for LEP students (p = 0.14) unlike third grade results. Specifically, the expected difference in the ELA achievement between WebPlay and control students was about 7 points in the direction that favors the control school students, for typical students in typical schools, which is negligible and statistically non-significant. The expected difference between WebPlay LEP students and control LEP students was about 7 points, in the opposite

direction (i.e., favoring WebPlay school students), which again is of negligible magnitude as well as being statistically non-significant.

All the other variables that are included in the final model were significantly related to student ELA achievement. Not surprisingly, student-prior-CST-ELA score and school-prior-CST-ELA-average score, were on average positively related to the outcome of interest, (i.e., the 2006–07 CST ELA scores). Language status was also a significant variable in the CST ELA score predictor. Compared to EO students, RFEP students on average perform better, holding constant all the other variables in the model, by 5 points, and LEP students in control schools on average performed worse by 9 points. Also, students who are classified with GATE status tended to perform better by 12 points, holding constant all the other variables. Lastly, Asian students tended to perform better than White students by 15 points controlling for all other variables in the model.

Table 9

Results from HMs Predicting Performance in CST in ELA for WebPlay and Control Students: Grade 5

Fixed effect	Estimate	*SE*	*df*	*t* value	*p* value
Model for class means					
Adjusted grand mean	343.51	3.81	18	90.06	<.0001
WebPlay contrast	-6.61	4.71	18	-1.40	0.18
Class pretest average	0.63	0.09	18	6.69	<.0001
Average within-class slope					
Pretest/posttest	0.61	0.04	569	16.59	<.0001
ASIAN/posttest	15.45	4.42	569	3.50	0.00
RFEP/posttest	5.01	2.55	569	1.97	0.05
LEP/posttest in WebPlay	6.51	4.43	569	1.47	0.14
LEP/posttest in control	-9.47	3.49	569	-2.71	0.01
Gifted/posttest	12.26	2.84	569	4.32	<.0001
Random effect	**Estimate**	***SE***		***Z* value**	***p* value**
Between class					
Variance in adjusted means	83.88	33.79		2.48	0.01
Variance in pre/post slopes	0.01	0.01		2.16	0.02
Within class					
Variance in WebPlay	442.88	34.92		12.68	<.0001
Variance in control	454.87	42.33		10.75	<.0001

Note. HM = Hierarchical Model, CST = California Standards Test, IFEP = Initial Fluent English Proficient, RFEP = Reclassified Fluent English Proficient, LEP = Limited English Proficient.

Conclusion and Discussion

Significant effects of WebPlay participation were found on both the student survey measure and the CST, although not consistent across all scales or for all students. Keeping in mind the relatively small sample size with both pre- and post-data available, a significant WebPlay participation effect was found on the Attitude/Engagement portion of the survey, with WebPlay students performing significantly better on posttest scores on this portion of the survey than the comparison students did (controlling for pretest scores). Further exploratory investigation suggests that this difference may primarily represent increases in the WebPlay students' Theatrical Interest/Engagement scale/factor.

There were no statistically significant differences between WebPlay and control students in the knowledge items during overall pre- to post-implementation. It should be

noted that these knowledge items were revised for the 2007–08 WebPlay implementation, with the goal of providing more targeted information about the impact of WebPlay on student theatrical knowledge.

In terms of the analysis of program effects on the ELA CSTs, results are mixed between the two main grades included in the sample. For third graders, there was no significant overall effect of WebPlay, but there was significant interaction between WebPlay program participation and student LEP status. Specifically, for LEP students, the expected difference between WebPlay and the control group was approaching ½ standard deviations, which indicates a considerable size of treatment effect. However, in Grade 5, the WebPlay program did not significantly relate to ELA achievement, overall or specific to LEP students.

The finding of the most significant program effects at third grade with LEP students suggests that program impact on literacy learning may be strongest with those at earlier stages of literacy skills development—in this case, third grade LEPs, whose English literacy development could be assumed to be relatively lower than the other non-LEP students. Using theater as an entry to literacy activities may be of particular benefit to this group. The mixed effects between two grades may also be influenced by a variety of other factors. For example, the WebPlay curriculum may have been better integrated into the district's existing ELA curriculum in the third grade than in the fifth grade. Or, the different findings between grades may be due to student social development characteristics. For example, it may be easier to engage third graders in the performance-based activities that WebPlay employs, whereas older students may be more concerned about their peer group's perceptions of them when they engage in activities. Differences in characteristics between third grade LEP students and fifth grade LEP students might also have had an impact on these findings. It can be suggested that students who are not reclassified as English fluent until fifth grade may differentially represent "longer term" English language learners (ELLs)—that is, students who are not reclassified within 2 years on entering the system (Wolf et al., 2008). These students may thus encounter many additional challenges in schools that ELLs at the lower grades do not face. Improving ELA achievement significantly for such students may take more intensive, targeted interventions than what can have impact for third grade LEP students.

As noted earlier, there are several methodological limitations to the evaluation design, including the fact that WebPlay participant teachers were self-selected rather than randomly assigned, the different sample sizes used for the survey and CST analyses, and the relatively low survey response rates for the comparison classrooms. Furthermore, both additional broader (i.e., larger, randomly sampled) and deeper (i.e., interviews, observations) data would provide a better understanding of the nature and scope of the WebPlay program

impact. For example, classroom observations or teacher logs might help identify certain instructional strategies that teachers implement as part of the program in their classrooms that better predict positive student outcomes than other strategies do.

However, even with these caveats in mind, the findings of some significant differences between WebPlay and comparison classrooms on both survey and standardized test outcomes, utilizing the most rigorous statistical techniques appropriate, is noteworthy. The findings of significant differences between WebPlay and control groups are particularly encouraging given that this was the first exposure to WebPlay for all of the students and most of the teachers, and that the program represented one relatively small piece of the overall classroom curriculum. In summary, even with a relatively small sample, WebPlay participation appeared to have a positive impact on student engagement and, for certain students, academic achievement relative to a comparison group in similar schools. Again, additional research, both quantitative and qualitative, would help to both provide additional support to these findings, and also better understand how the WebPlay program can be best implemented to maximize program impact on students.

References

Baker, E. L., Griffin, N., & Choi, K. (April, 2008). *The achievement gap in California: Context, status, and approaches for improvement* (Policy Brief). Davis, CA. University of California, Davis School of Education Center for Applied Policy in Education.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Edward Arnold.

Green, J. L., Camilli, G., & Elmore, P. E. (Eds.). (2006). *Handbook of complimentary methods in education research*. New York: Routledge.

Jorskog, K., & Sorbom, D. (1989). *LISREL 7: A guide to the program and applications.* Chicago: Statistical Package for the Social Sciences.

Muthén, B. O. (1998). Mplus technical appendices [Computer software appendices]. Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2004a). Mplus (Version 3.11) [Computer software]. Los Angeles: Author.

Muthén, L. K., & Muthén, B. O. (2004b). Mplus: User's guide [Computer software manual]. Los Angeles: Author.

Partnership for 21st Century Skills. (2008). *21st century skills, education, and competitiveness.* [Policy paper]. Washington DC: Author

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newberry Park, CA: Sage Publications.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Wolf, K., Kao, J., Herman, J. L., Bachman, L. F., Bailey, A., Bachman, P. L., et al. (2008). *Issues in assessing English language learners: English language proficiency measures and accommodation uses* (CRESST Tech. Rep. No. 713). Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing (CRESST).