Empirical Education

RESEARCH REPORT

The Comparative Effectiveness of Professional Development and Support Tools for World Language Instruction:

A Report of a Randomized Experiment in Delaware

Jessica Villaruz Cabalo
Boya Ma
Andrew Jaciw
Empirical Education Inc.

April 5, 2007

Empirical Education Inc.
www.empiricaleducation.com
425 Sherman Avenue, Suite 210
Palo Alto, CA 94306
(650) 328-1734

Acknowledgements

We are grateful to the people in the World Languages and International Education Department of the Delaware Department of Education for their assistance and cooperation in conducting this research. Language Learning Solutions has collaborated with the Delaware Department of Education on providing all the materials. The research was funded by a grant (#R305E040031) to Empirical Education Inc. from the U.S. Department of Education. The purpose of this grant is to improve our ability to run small scale experiments to assist local decision-makers. The US ED is not responsible for the content of this report.

About Empirical Education Inc.

Empirical Education Inc. was founded to help school districts, publishers, and the R&D community assess new or proposed instructional and professional development programs through scientifically based pilot implementations. The company draws on the expertise of world-class researchers and methodologists assuring that the research is objective and takes advantage of current best practice in rigorous experimental design and statistical analysis. The company's findings let educators quantify the value of programs and help them partner with providers to implement those most effective for their students.

The Comparative Effectiveness of Professional Development and Support Tools for World Language Instruction:

A Report of a Randomized Experiment in Delaware

Table of Contents

Introduction

The Delaware Department of Education (DDOE) was interested in evaluating components of a statewide pilot program to support instruction in World Languages (particularly Spanish and French). The first component was an online student assessment, the Standards-based Measurement of Proficiency (STAMP) to measure the learning outcomes. The DDOE also wanted to evaluate their professional development program which consisted of the combined use of STAMP and ClassPak, an online teaching tool. Both STAMP and ClassPak are products and trademarks of Language Learning Solutions. Through this study, the DDOE developed the following objectives:

- Offer participating teachers and their students the online assessment in French and Spanish;
- Use the assessment data to review the degree of achievability and alignment of and between the Delaware World Language Content Standards and Delaware World Language Performance Indicators (DEWLPI);
- Compare Delaware data with those of the national obtained through the *STAMP* test administered nationally;
- Provide the participating teachers the opportunity to learn and use the additional online support tool, ClassPak;
- Conduct an additional study to obtain scientifically-based data to determine the effectiveness of ClassPak and related professional development
- Incorporate the assessment data to develop the statewide recommended WL curriculum.

This study is a replication of a previous attempt to evaluate the effectiveness of the DDOE's professional development program for World Language instruction. The question specifically addressed in this study is whether students in classes of the teachers who received the professional development and were given access to the ClassPak materials will perform better in reading and writing in Spanish or French, as measured by the STAMP, than they would if they had been in a control classroom. In this report we use the term "*PD*" to refer to the intervention consisting of training and the online tool. We were also interested in understanding the relationship between the DSTP and STAMP tests. The US Department of Education's research funds supported Empirical Education's efforts in the research. A measure of the impact of the program could provide useful evidence to support state decisions about their World Language programs.

The design of the experiments reflects the requirements of the No Child Left Behind Act, which directs schools to consult reports of rigorous research in making adoption decisions about instructional programs. The US Department of Education (2003) has been explicit in interpreting this requirement in terms of randomized experimentation for determining effectiveness. In a randomized experiment, we reduce selection bias by tossing a coin to assign teachers to use a particular product—in this case, *PD*—or to continue using their current teaching materials and methods. This design is considered appropriate for valid conclusions about effectiveness (Shadish, Cook, & Campbell, 2003). Nevertheless, given the specifics of the implementations and the particular characteristics of the districts studied here, we do not intend for the research by itself to apply generally to other districts or states.

Methods

Research Design

The study is a comparison of outcomes for groups of students taught by teachers who received professional development of the combined use of ClassPak and STAMP (*PD,* the treatment group) and students taught by teachers who received professional development on STAMP only (the control group). Thirty teachers volunteered for participation in the study. From the pool of volunteers, the DDOE randomly assigned equal numbers of teachers to the *PD* group and control group. Empirical Education advised in the procedures but did not conduct the actual randomization.

Data Sources and Collection

The data used for this study were student achievement measures obtained from the DDOE. The DDOE provided us with all the data including the 2005 Delaware State Testing Program (DSTP) scores in reading, which was used as a pretest measure and as a check on the correlation with the STAMP pretest and related analyses. The DDOE also provided STAMP scores from Learning Language Solutions, Inc (LLS). Additionally, we integrated the class roster and teacher background information and all the data from the DDOE into a standard data warehouse for the study.

STAMP Assessment

The primary outcome measure of interest are student-level proficiency scores on the STAMP test, a criterion-referenced and summative assessment developed by LLS, Inc. and the National Foreign Language Resource Center at the University of Oregon. The STAMP also follows the American Council on the Teaching of Foreign Languages (ACTFL) guidelines (Center for Applied Second Language Studies, 2006). According to LLS, STAMP is a computer- adaptive testing program that assesses student reading, writing and speaking skills. STAMP assesses student proficiency levels in World Languages, from Novice-Low to Intermediate-Mid benchmarked to the ACTFL scale. The pretest STAMP administration occurred in November-December 2005 and the posttest occurred in March 2006.

State Achievement Test Scores

Another outcome measure of interest was the student-level scaled test scores on the DSTP in reading. The DSTP is a state-mandated end-of-year assessment, based on Delaware Performance Level Descriptions (PLDs) and Grade Level Expectations (GLEs). It was developed to be aligned with the state English Language Arts content standards. The DSTP Reading Assessment is administered in March of every academic year. DSTP in reading outcomes are reported as a scaled score from 300 to 800.

Intervention

Similar to STAMP, ClassPak was developed by Language Learning Solutions, Inc. and the National Foreign Language Resource Center at the University of Oregon. According to LLS, Inc., ClassPak is an instructional tool for teachers, a means for formative assessment and allows teachers to build their own lessons, classroom activities and exams using real world situations. It is an online, data-driven instructional support tool that helps teachers build, deliver and manage reading, writing and speaking quizzes, lessons, assignments and activities. ClassPak is pre-loaded with STAMP-like, reality-based content that assists teachers in creating a more proficiency-oriented classroom. It can also familiarize students with the STAMP test experience, since ClassPak quiz items resemble STAMP test items (Learning Language Solutions, 2006)

Classes in both the *PD* and control groups administered the STAMP Assessment. All teachers were given access to computers equipped with the STAMP Assessment. All teachers (both the *PD* and control) received a training session on STAMP Administration from LLS, Inc. Content of the STAMP training included an introduction to STAMP, coverage of state assessment and standards, use of STAMP, how to handle technology issues and LLS, Inc. resources. Practice time with STAMP was also included in the training.

In addition to STAMP training, *PD* teachers received a brief training session on the use of ClassPak. Content covered in the ClassPak training included instructions on use of the login, Quiz Builder, Lesson Builder and Activity Hunter. Practice time with ClassPak was also included in the training. *PD* teachers were equipped with the ClassPak software.

Site Descriptions

According to the US Census Bureau, the total population of Delaware in 2005 was 843,828. Among the adult population, 84% has received a high school diploma and 27.8% has received a Bachelor's Degree. The median age is 36 and the median household income is 73,597. In 2004, the DDOE

served 117,668 students across 177 schools in 19 districts. The ethnic make-up is 57.3% White, 31.9% Black, 7.9% Hispanic, 2.6% Pacific Islander and 0.3% American Indian or Alaska Native. Of the student population in the state, 33.8% are economically disadvantaged (compared to 36.7% in the nation), 3.4 % are English Language Learners (compared to 7.8% in the nation) and 14.6% have a disability (compared to 12.8% in the nation.

Sample and Randomization

Recruiting

The first meeting for the ClassPak /STAMP project occurred on November 12, 2005 and was attended by a LLS representative, the project coordinator and the researchers in the World Languages & International Education department at the DDOE, and 30 teachers across Delaware. All 30 teachers were lead teachers in their school's World Language departments and were trained in STAMP administration. Following this, the project coordinator and DDOE researchers conducted a randomization procedure among those teachers present. The study involved 11 school districts, 16 schools, and 30 teachers, equally divided into the *PD* and the control group.

Randomization

Thirty teachers were assigned (by the DDOE) using a coin toss to either the *PD* or control condition. Randomization ensures that, on average, characteristics other than treatment, which affect the outcome, are evenly distributed between treatment and control. This prevents us from confusing treatment with some other factors, technically called 'confounders', that are not evenly distributed between groups and that affects the outcome. For example, through randomization we try to achieve balance between treatment and control on the average years of teaching experience – a factor that presumably also affects the outcome.

There are various ways to randomize teachers to conditions. We used a matched pairs design whereby we first identified pairs of similar teachers, and then, within each pair, we assigned one teacher to *PD* and the other to control. Similarity was based on whether teachers were in the same grade level and whether they shared common meeting times. A pairing strategy often results in a more precise measurement of the treatment impact.

Sample Size

One concern we had was with sample size. Sample size is one of the things that determines how precisely we can measure an effect of a given size. With smaller samples we are usually only able to detect larger effects. We usually measure the size of an effect in terms of standard deviation units – which tells us how big the effect is, controlling for the spread in observed scores. Based on the available sample size, and certain assumptions about other parameters that affect the size of the effect that we can detect, we computed that we can detect an effect size as small as .35. This is computed assuming false-positive and false-negative error rates of .05 and .20 respectively. Raising the false positive rate to .20 reduces the size of the effect that we can detect to .26[1]. We emphasize that the matching design that we used potentially further lowers this value. From this we

[1] Power calculations given here are based on of reaching false positive and false negative conclusions set at the conventional values of .05 and .20, respectively; Increasing the rate of false positives to .20 dropped the minimum detectable effect size to .26. We describe effect sizes in terms of standard deviation units for a continuous normally distributed outcome. This allows us to more easily interpret how big an effect we might expect. The effect sizes that are reported in the data tables are in the log-odds metric.

see that the experiment is not designed to detect a small effect, which may be real but not discernable given the number of teachers in the study.

Because the importance of the information warranted gathering the available data even if the results ultimately proved inconclusive, the district in consultation with the researchers decided to move forward with the experiment.

Statistical Analysis and Reporting

The basic question for the statistical analysis was whether students in *PD* classrooms had higher reading and World Language scores than those in control classrooms. The mean impact is estimated using multi-level models that account for the clustering of students in classes, which provides a more accurate, and often more conservative, assessment of the confidence we should have in the findings. We use HLM SAS PROC MIXED (from SAS Institute Inc.) as the primary tool for this work. To increase the precision of our estimate, we include students' pretest scores in the analysis. In our experience, these are good predictors of achievement; including them as covariates in the impact analysis reduces the error variance, which makes it easier to discern the treatment impact.

In addition to the basic analysis of the mean impact, the plan for the study identifies the teacher- and student-level covariates that we expect (through theory or prior research) to make a difference in the effectiveness of the program being tested. The analysis tests for the interactions between those covariates and the experimental condition.

In addition to examining impacts and interactions where we anticipate effects, to better understand unexpected results, we use other demographics and teacher characteristics in exploratory analyses to generate additional hypotheses about which factors potentially moderate or mediate the treatment impact.

Our analyses produce several results: among them are the estimates for fixed effects, effect sizes, and *p* values. These are found in all the tables where we report the results of the statistical models.

Estimates. The estimate can be thought of as a prediction of the size of an effect. Specifically, it is how much we would predict the outcome to change for a one-unit increase in the corresponding variable. We are often most interested in the estimate associated with the experimental conditions, which is the expected change in outcome in going from control to treatment, holding other variables constant.

Effect sizes. We also translate the difference between treatment and control into a standardized effect size by dividing the difference by the amount of variability in the outcome (also called the standard deviation). This allows us to compare the results with results from other studies that use different measurement scales. In studies involving student achievement, effect sizes as small as 0.1 (one tenth of a standard deviation) are sometimes found to be important educationally. The unadjusted effect size is the difference between treatment and control, controlling for dependencies of observations within randomized units. (This has implications for *p* values, but it also affects the estimate of the difference: it weights some cluster averages more than others – therefore we can expect inconsistency between the estimated difference and the raw difference.) The adjusted effect size adjusts for the pretest as well as other fixed and random effects used in the models with interactions that follow.)

For this study, due to the metric on which outcomes were reported, we express the effect sizes in terms of odds ratios.

***p* values.** The *p* value is very important because it gives us a gauge of how confident we can be that the result we are seeing is not due simply to chance. Specifically, it tells us what the probability is that we would get a result with a value as large as — or larger than — the absolute value of the one observed when in fact there is no effect.[2] Roughly speaking, it tells us the risk of concluding that the treatment has had an effect when in fact it hasn't. Thus a *p* value of .1 gives us a 10% probability of that happening. We can also think of it as the level of confidence, or the level of belief we have that the outcome we observe is not simply due to chance. While ultimately depending on the risk tolerance of the user of the research, we suggest the following guidelines for interpreting *p* values:

1. We have a high level of confidence when $p \leq .05$. (This is the level of confidence conventionally referred to as "statistical significance.")
2. We have some confidence when $.05 < p \leq .15$.
3. We have limited confidence when $.15 < p \leq .20$.
4. We have no confidence when $p > .20$.

Results

Formation of the Experimental Groups

Groups as Initially Randomized

The randomizing process does not guarantee that the groups will be perfectly matched. It simply guarantees that there is no intentional or unintentional bias in the selection of teachers into the treatment or the control condition. It is important to inspect the two groups to determine whether, in spite of randomization, there are any significant differences on factors that affect the outcome. (Randomization ensures lack of bias, but we are interested in knowing whether the particular estimate resulting from this randomization may be far from the true value as a result of chance imbalances on factors that affect the outcome.) The following tables address the nature of the groups in each of the sites. Table 1 and Table 2 show the distribution of teachers, classes, grades, and students between *PD* and control conditions. This is the complete number of students in the experiment at the time that the experiment began in November 2005.

[2] *p* values for estimated odds ratios have a slightly different interpretation because the distribution of the estimator is not symmetric and odds ratios don't take on negative values.

Table 1. Distribution of the *PD* Group by Schools, Teachers, Grades, and Counts of Students

School ID#	Teacher ID#	Class ID#	Grade 7	Grade 8	Grade 9	Grade 10	Grade 11	Grade 12	Total per class
531	611	1	0	0	2	43	11	6	62
533	596	3	0	0	0	15	24	6	45
	597	4	0	0	0	16	10	4	30
537	593	6	0	0	1	14	21	15	51
539	594	9	1	56	0	0	0	0	57
546	633	17	0	0	1	13	9	7	30
550	632	22	0	0	0	20	22	5	47
	642	23	0	0	6	32	29	18	85
551	646	21	0	0	18	2	23	6	49
	647	19	0	0	9	5	4	0	18
555	651	14	0	0	1	1	36	1	39
	652	12	0	0	0	0	16	0	16
559	673	24	0	0	0	5	19	26	50
Number of units			**Total students per grade**						**Students in *PD* group**
9	13	13	1	56	38	166	224	94	579

Table 2. Distribution of the Control Group by Schools, Teachers, Grades, and Counts of Students

School ID#	Teacher ID#	Class ID#	Grade 9	Grade 10	Grade 11	Grade 12	Total per class
527	645	13	2	18	7	3	30
531	635	2	0	8	30	4	42
537	636	8	0	20	7	2	29
	637	7	2	8	6	3	19
	665	5	0	0	3	28	31
542	610	10	1	1	27	34	63
	616	11	0	36	10	11	57
545	624	16	3	1	4	13	21
551	648	18	4	19	8	17	48
	649	20	12	15	7	5	39
555	653	15	0	0	12	0	12

Number of units			Total students per grade				Students in control group
7	11	11	24	126	121	120	391

Post Randomization Composition of the Experimental Groups

With randomization, we expect certain student and teacher characteristics to be equally distributed between treatment and control groups, but in any single randomization there may be discrepancies between the distributions due to chance. In checking for balance in the composition of the experimental groups, we examine student characteristics (SES and ethnicity), teacher experience, as well as student pretest outcomes.

970 students were included in the analysis of the LSS outcome. 39 cases in the control group from Teacher ID 649 were removed because they are the only students who study German. There are no comparison cases in the *PD* group. Then, 27 cases were removed because they are disabled students. This results in a total of 904 cases.

Student Variables

Socio-Economic Status

Table 3 shows the distribution of the socio-economic status (SES) of the students in each group, as determined by participation in the Free/Reduced-price Lunch program.

Randomization resulted in SES being evenly balanced between *PD* and control. We tested this formally, and the high *p* value of .43 indicates that we should not reject the hypothesis that there is balance.

Table 3. Comparison of Free Lunch Status between *PD* and Control Group

Condition	Free Lunch Status		
	Free/reduced Lunch	Not Free/Reduced Lunch	Totals
PD	102	460	**562**
Control	55	287	**379**
Totals	**157**	**747**	**904**
Statistics	**DF**	**Value**	***p* value**
Chi-square	1	0.63	.43

Ethnicity

Table 4 summarizes the distribution of student ethnicity.

We see that ethnicity was not distributed evenly between the conditions in spite of randomization. There are proportionally more White students in the *PD* group than in the control group. Chi-square tests indicate that despite randomization, ethnicity was not balanced between conditions. The imbalance may lead the estimate of the impact to depart from its true value.

Table 4. Comparison of Ethnicity Between *PD* and Control Group

	Ethnicity				
	Asian	Hispanic	Black	White	Totals
PD	21	20	122	399	562
Control	12	33	93	204	342
Totals	33	53	215	603	904
Statistics		**DF**	**value**		***p* value**
Chi-square test		3	20.28		<.01

English Language Learner Status

A majority of the students are native English speakers. This implies that this sample is a good representation of the community. As a result of random assignment, the English proficiency level of the students is evenly distributed across the *PD* and control groups. The result of the statistical test is consistent with this assertion. Table 5 summarizes the distribution of English proficiency.

Table 5. Comparison of English Proficiency Between *PD* and Control Group

Condition	English Proficiency		
	English proficient	Not proficient	Totals
PD	555	7	**562**
Control	335	7	**342**
Totals	**890**	**14**	**904**
Statistics		**Value**	***p* value**
Fisher's exact test		0.14	.41

Note. Some categories have a small expected number of cases; hence, Fisher's exact test is reported.

Distribution by Grade

Table 6 summarizes the distribution of students by grade. We see that grade was not distributed evenly between the conditions in spite of randomization. There are proportionally more 11th grade students in the *PD* group than in the control group; and less 12th grade students in the *PD* group than in the control group. Chi-square tests indicate that despite randomization, grade was not balanced between conditions. This is a factor that was hard to randomize due to the World Language class structure, Students across grades can take the same World Language class. The imbalance may lead the estimate of the impact to depart from its true value.

Table 6. Comparison of Grade Level between *PD* and Control Group

	Grade					
	8	9	10	11	12	Totals
PD	56	32	165	216	93	562
Control	0	12	107	111	112	342
Totals	56	44	272	327	205	904

Statistics	DF	Value	*p* value
Chi-square test	4	63.13	<.01

Teaching Experience

During the randomization process, teachers identified themselves according to years of teaching experience and the initial teacher pair was formed correspondingly so that the bias due to teaching experience would be distributed among the groups evenly. As part of our data collection we were provided with information regarding teaching experience. The following table summarizes the background characteristics of the teachers in the study.

Table 7. Total Number of Years of Teaching Experience

Condition	0 to 3 years	4 and more years	Number of teachers
PD	1	12	13
Control	2	9	11
Totals	**3**	**21**	**24**

Statistics	Value	*p* value
Fisher's exact test	0.35	.58

Note. Some categories have a small expected number of cases; hence, Fisher's exact test is reported.

Characteristics of the Experimental Groups Defined by Pretest

We also checked whether randomization resulted in balance on pretest scores, a variable that we include in most of our analyses to increase the precision of our estimates.

There are three pretest scores that were involved in the analysis—the STAMP Reading test, the STAMP Writing test and the DSTP Reading test. We will report them separately in the following text.

STAMP Reading Test

A total of 904 students were left after taking out students with a disability and students taking a German class. Out of those, 62 (or 6%) students who received the STAMP Reading pretest score higher than 3 (on the STAMP scale). We are not considering them in the following analysis because after preliminary review, this small number of data greatly influenced our analysis. Table 8 shows the results of 832 students who have pretest scores of 3 or lower.

Table 8. Comparison of STAMP Reading Pretest Score between Students in *PD* and Control Group

	Pretest score			
	1	**2**	**3**	**Total**
PD	236	223	62	521
Control	126	141	44	311
Totals	362	364	106	832

Statistics	DF	Value	*p* value
Chi-square test	2	2.08	.35

Randomization resulted in STAMP Reading pretest scores being evenly balanced between *PD* and control groups. We tested this formally, and the high *p* value of .35 indicates that we should not reject the hypothesis that there is balance.

A similar chi-square test has also been done on 805 students who have the STAMP Writing pretest score 4 or lower. A low *p* value of .13 indicates that we should not reject the hypothesis that there is balance between *PD* and control groups.

DSTP Reading Test

Table 9 shows the results on 549 out of 832 students who have STAMP Reading pretest scores and DSTP Reading pretest scores.

Table 9. Difference in DSTP Reading Pretest Scores Between Students in the *PD* and Control Groups

Descriptive statistics: Pretest outcomes	Raw group means	Standard deviation	Number of students	Standard error	Effect Size
PD	537.38	30.60	369	1.59	-0.18
Control	542.91	31.21	180	2.33	
***t* test for difference between independent means**	**Difference**	**DF**	***t* value**	***p* value**	
Condition (*PD* – control)	-5.53	547	1.98	.05	

The *PD* and control groups had slightly different average pretest scores on the DSTP Reading test, as shown in Table 9. However, when we accounted for the fact that outcomes for students of the same teacher tend to be dependent by modeling these dependencies, the discrepancy became less discernable. In the analyses that follow, we add the pretest covariate in order to increase the precision of the impact estimate. But we recognize that, with or without this covariate, the impact estimate is unbiased as a result of the randomization.

Attrition

Based on the cases of non-disability students from grades 8-12, a high percentage of students did not take the posttests (Spring 2006 STAMP Reading and Writing Tests).

STAMP Reading Test

Out of 842 students who have neither pre nor posttest score higher than score 4, 736 students have posttest scores. Posttest scores are missing for 106 or 12.6%. Table 10 shows the breakdown on students who have pretest only contrasted to students who have both pre and posttests by the *PD* and control groups. An exact test of differences in proportion indicates that there is no relationship between attrition and experimental condition (or, there is no differential attrition). This is important because it means that the attrition does not bias the comparison between the two groups. Data used in the tables also reflect that 10 students who have posttest scores did not have a pretest score.

Table 10. Students Missing STAMP Reading Test Score Data

	Categories of missing data		
Condition	**Having both pre- and posttest scores**	**Having only pretest score**	**Totals**
PD	463	58	521
Control	267	44	311
Totals	730	102	832
Statistics	**DF**	**Value**	***p* value**
Chi-square test	1	1.65	.20

Note. The *PD* group attrition rate is so large because 1 teacher dropped out due to the fact that his/her whole class is missing posttest scores.

Considering the large number of students having pretest scores, but missing posttest scores, 102 students could not be used in the analysis. This may be because students could have been absent during testing, some students could not finish the test and one teacher was missing posttests scores for an entire class. Despite the large number of missing students, we were able to get a high *p* value of .20, which indicates that we should not reject the hypothesis that there is balance between *PD* and control groups.

STAMP Writing Test

Out of 889 students who have neither pre nor post test score higher than score 5, 751 students have posttest scores. Posttest scores are missing for 138 or 15.5%. Table 11 shows the breakdown of students who have pre test only in contrast to students who have both pre and post test by the *PD* and control groups. An exact test of differences in proportion indicates that there is no relationship between attrition and experimental condition (i.e., there is no differential attrition). This is important because it means that the attrition does not bias the comparison between the two groups. Data used in the tables also reflect that 84 students who have posttest scores did not have a pretest score.

Table 11. Students Missing Test Score Data

	Categories of missing data		
Condition	**Having both pre- and posttest scores**	**Having only pretest score**	**Totals**
PD	443	58	501
Control	262	42	304
Totals	705	100	805
Statistics	**DF**	**Value**	***p* value**
Chi-square test	1	0.87	.35

Considering the large number of students having pretest scores, but missing posttest scores, 100 students could not be used in the analysis. This may be because students were absent during

testing or could not finish the test and one teacher was missing his/her posttest for his/her class. The high *p* value of .35 indicates that we should not reject the hypothesis that there is balance between *PD* and control groups.

Implementation Results

The project did not provide an opportunity for conducting surveys, interviews or observations. We did receive descriptive information from the project coordinator. Though the *PD* program was intended for *PD* teachers to use ClassPak in combination with STAMP, the project coordinator reported that only 4 or 5 *PD* teachers used ClassPak. The coordinator also reported was that there was very little evidence of actual use of ClassPak by these teachers. Two reasons were suggestions. First, they did not receive sufficient training on the tool. They only received a basic introductory training during the initial meeting and were not provided with on-going professional development or support. Secondly, teachers did not find ClassPak to be useful in their planning and instruction, but instead considered the tool to be an additional burden to their already heavy workload.

On the other hand, the project coordinator believed that the STAMP Project experience provided opportunities for professional development and among the positive outcomes were the following:

1. Teachers became accustomed to the idea that effective assessment does not have to be conducted with a pencil and paper.
2. After the initial anxiety that teachers may be putting their reputation as teachers on the line, they felt secure because of the confidential aspect of the STAMP.
3. Teachers appreciated the opportunity to compare themselves and their students to others on the national level.

Overall, the project coordinator felt that the STAMP test is an effective tool to measure proficiency in World Languages. The low level of pre to posttest attrition indicates that teachers made the effort to use the STAMP test.

Quantitative Results

The primary topic of our experiment was the impact of professional development. We first look at the impact of *PD* on the outcome measure (the scores on the STAMP Reading test and the score on the STAMP Writing test). Second, we ask whether student's learning ability in reading English is related to their ability in learning World Languages. We will examine reading ability using DSTP tests. Finally, we raise some concerns about the STAMP test and its usability for this kind of research.

We had also planned to model the impact of teacher experience on student outcomes, but due to the small number of teachers having fewer than 4 years of experience, we did not have a large enough sample to adequately perform this analysis.

Nature of Analysis

We present below the results for the Reading and Writing STAMP tests. STAMP Reading is measured in terms of five ordinal categories and Writing is measured in terms of six ordinal categories.

We performed a categorical outcome analysis using the program HLM, which allows us to model schools and teachers as random factors. Among other things this control for dependencies among observations within teachers and schools and gives us a more accurate and conservative estimate of the effects of interest.

The rationale for and methods used to perform analyses involving ordinal outcomes are detailed in Raudenbush and Bryk (2002). Basically, the method consists of modeling cumulative odds or cumulative probabilities of scoring at each category given specific levels of the covariates. This is an extension of logistic regression except the log odds refer to the probabilities of being at or below each level. The cumulative probabilities can then be decomposed into probabilities of scoring at each level given the covariates (including the performance rating a student received before the start of the intervention as well as the treatment indicator.) The HLM extension of this approach involves adding

random effects to the model which can be regarded as perturbations from the estimated fixed effects in the model.

A disadvantage of modeling categorical outcomes is that it makes the results hard to interpret. The log odds metric, which has the convenience of allowing us to express the outcome variable in terms of a linear combination of effects, does not easily reveal the relationship between the covariates and the outcome. One option is to convert the results to probabilities, but here too it is difficult to link covariate estimates in the model to graphs of probabilities, especially when interaction effects are modeled. The fact that the conversion to probabilities first gives us cumulative probabilities further complicates the situation.

To deal with the complexity of the model and provide a result that is more easily interpreted we ran the analyses two ways. First we performed the categorical outcome analyses described above. Next, we fitted linear regression models, the results of which are much easier to interpret. After confirming that the results from both kinds of analyses are similar we present the results of the analysis that is easier to interpret. Specifically, we compare the results of both approaches to see if the statistical significance of the effects of primary interest and / or the direction of these effects changes in moving to the simpler model (the effects of interest are the treatment effects as well as interactions of treatment with specific covariates.) If there is a change in the level of confidence that we have (i.e., the levels described on pgs. 4 and 5) that there is a non-zero effect then we retain the complex model. If we have at least limited confidence that there is a true effect and the direction of the effect changes then we retain the complex model Otherwise we move to the linear model. We emphasize that in the linear model we continue to model the dependencies among observations for teachers and schools. Analyses involving linear models are done using PROC MIXED in SAS.

We performed three main analyses. The first involved the STAMP Reading outcome with the STAMP Reading pretest as a covariate. The second involved the STAMP Reading outcome with the state Reading pretest as a covariate. The third involved the STAMP Writing outcome with the STAMP Writing pretest as a covariate. By the criteria described in the previous paragraph, we are able to present the results of the analysis based on the linear model for each of these three analyses.

In all cases we examined the interaction of the pretest with the treatment condition to see if treatment was differentially effective depending on a student's incoming achievement level. The interaction did not cross the threshold of statistical significance for Reading, so we present the results of models without the interaction. For STAMP Writing the interaction crossed the threshold of statistical significance (i.e., it achieved a level that would give us limited confidence that the interaction is different from zero.) Because the estimated interaction effect is small, we present the results from the model based on a linear fit and without the interaction. However, we describe how to interpret the weak interaction in the categorical outcome model.

When reporting effect sizes we provide means and standard deviations for scores in reading and writing. However, due to the categorical nature of the outcome variable we do not calculate the effect size that is normally computed when outcomes are continuous. Instead we report the odds ratio associated with the treatment effect. This is the odds of scoring at or below a given level on the outcome for the treatment group divided by the odds of scoring at or below the same level on the outcome for the control group. Odds ratios range between zero and infinity. An odds ratio of 1 means no difference between the groups.

Influential Point and Category Exclusions

The STAMP writing outcome ranges between 1 and 6. We eliminated the top two levels because a relatively small number of cases fell into these levels and their inclusion would have the potential to influence the outcome a lot. We were concerned that the fit of the linear model would depend greatly on these few cases. Likewise, we eliminated the top two levels of the STAMP reading scale. The scale ranged from 1 through 5. After reducing levels it ranged from 1 to 3. We did not remove additional cases. We believe that approaches to identifying outliers that are used with normally and continuously distributed outcomes are not suited to the analyses we are doing. We also centered the STAMP pretest on level 2. In other words, we recoded the levels of the Writing outcome from 1, 2, 3, and 4 to -

1, 0, 1 and 2, respectively. We recoded the level of the Reading outcome from 1, 2, and 3 to -1, 0 and 1. This leads to an easier interpretation of the some of the effects that are estimated in the models.

Impact of *ClassPak*™ on *STAMP* Reading Outcomes

Our first analysis addressed Reading outcomes using the *STAMP* Reading test. Table 12 provides a summary of the sample we used in the analysis and the results for the comparison of *PD* and control. The "Unadjusted" row gives information about all the students in the original sample for whom we have a posttest. This shows the means and standard deviations as well as a count of the number of students, classes and teachers in that group. The *p* value indicates the probability of arriving at a difference as large as or larger than the absolute value of the one observed when there truly is no difference. [3]The "Adjusted" row is based on the students who have both pretests and posttests. This is the sample that we use in the analyses on which we base our results reported in Table 13. The means, and therefore the effect size, are adjusted to take into account the student pretest scores.

Table 12. Effect Sizes for Students with STAMP Reading Posttest and STAMP Reading Pretest

	Condition	Means	Standard deviations	No. of students	No. of classes	No. of teachers	Effect size (Odds ratio)	*p* value[a]
Un-adjusted effect size	***PD***	1.82	0.70	467	12	12	1.21	.67
	Control	1.91	0.73	269	10	10		
Adjusted effect size	***PD***	1.82	0.70	463	12	12	1.10	.78
	Control	1.91	0.72	267	10	10		

[a] The *p* value for the unadjusted effect size is computed using a model that figures in clustering of students in teachers, but does not adjust for any other covariate. The *p* value for the adjusted effect size is computed using a model that figures in clustering and includes the pretest as a covariate.

Table 13 shows the estimated impact of *PD* on students' performance on the STAMP Reading test. The row in the table labeled "Impact of *PD*" gives us information about whether *PD* made a difference in performance on the STAMP test. The coefficient associated with the treatment is -0.06, which shows a small negative difference associated with *PD* for a student with an average score on the pretest. The *p* value of .65 indicates that we can expect to see a difference, as large as or larger than the absolute value of the estimate, 65% of the time when in fact there is zero impact. Using the criteria outlined earlier in the report, we conclude that we have no confidence that the true impact is different from zero.

[3] The *p* value indicates the probability of attaining an odds ratio as large or larger than the one observed (or as small or smaller than the point below which the reference distribution has the same area as the area to the right of the point estimate (since we're performing a two-tailed test)) when the odds ratio is actually 1.

Table 13. The Impact of *PD* on Student Performance in STAMP Reading

Fixed effects	Estimate	Standard error	DF	*t* value	*p* value
Predicted value for a control student with a score of a 2 on the pretest	2.33	0.25	9	9.37	< 0.01
Impact of *PD*	-0.06	0.14	9	-0.46	0.65
Predicted change in the control outcome for each unit increase on the pretest	0.43	0.04	707	12.05	< 0.01
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	0.03	0.02		1.65	.05
Within-teacher variation	0.33	0.02		18.82	<.01

[a] Schools are also modeled as a fixed factor but not included in this table

[b] Teachers were modeled as a random factor.

Impact of *Professional Development* on Writing Outcomes

Our second analysis addressed Writing outcomes using the STAMP Writing test. Table 14 provides a summary of the sample we used in the analysis and the results for the comparison of *PD* and control. The "Unadjusted" row gives information about all the students in the original sample for whom we have a posttest. This shows the means and standard deviations as well as a count of the number of students, classes and teachers in that group. The *p* value is interpreted the same way as explained in Table 12. The "Adjusted" row is based on the students who have both pretests and posttests. This is the sample that we use in the analyses on which we base our results reported in Table 15. The means, and therefore the effect size, are adjusted to take into account the student pretest scores.

Table 14. Effect Sizes for Students with STAMP Writing Posttest and the STAMP Writing Pretest

	Condition	Means	Standard deviations	No. of students	No. of classes	No. of teachers	Effect Size (Odds ratio)	*p* value [a]
Un-adjusted effect size	***PD***	1.79	0.81	472	12	12	1.09	.87
	Control	1.89	0.86	279	10	10		
Adjusted effect size	***PD***	1.81	0.81	443	12	12	0.99	.99
	Control	1.89	0.85	262	10	10		

[a] The *p* value for the unadjusted effect size is computed using a model that figures in clustering of students in teachers, but does not adjust for any other covariate. The *p* value for the adjusted effect size is computed using a model that figures in clustering and includes the pretest as a covariate.

Table 15 shows the estimated impact of *PD* on students' performance on the STAMP Writing test. The row in the table labeled "Impact of *PD*" gives us information about whether *PD* made a difference in performance on the STAMP test. The coefficient associated with the treatment is -0.06, which shows a small negative difference associated with *PD* for a student with an average score on the pretest. Using the criteria outlined earlier in the report, the *p* value of .74 leads us to conclude that we have no confidence that the true impact is different from zero.

Table 15. The Impact of *PD* on Student Performance in STAMP Writing

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Predicted value for a control student with a score of a 2 on the pretest	2.38	0.30	9	8.05	< 0.01
Impact of *PD*	-0.06	0.16	9	-0.34	0.74
Predicted change in the control outcome for each unit increase on the pretest	0.54	0.03	682	16.54	< 0.01
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	0.05	0.03		1.70	.05
Within-teacher variation	0.38	0.02		18.48	< .01

[a] Schools are also modeled as a fixed factor but not included in this table

[b] Teachers were modeled as a random factor.

We present here the results of a model that does not include the interaction of treatment with incoming achievement. In the model with the categorical outcome the interaction crossed the threshold of statistical significance (i.e., it achieved a level that would give us limited confidence

that the interaction is different from zero.) The estimated effect of the interaction was positive with a *p* value of .17. Therefore, in what follows, we describe how to interpret this interaction.

The positive interaction means that the odds of being at or below a given level on the posttest for a one-level increase on the pretest, divided by the odds of being at or below that level on the posttest holding constant the level of the pretest, is .16 for the control group and .20 for the treatment group.

This suggests that a gain in pretest is less advantageous for treatment since it's associated with greater odds of being at or below the given level. For example, there is no level lower than 1, so this translates into greater odds of being at 1 for treatment than control. This result should be considered tentative; that is, it is suggestive but not conclusive about possible differential effects of *PD* depending on how the student scores on the pretest. The *p* value for the interaction suggests caution in drawing conclusions from this result for which we have limited confidence.

Relationship Between DSTP Reading and *STAMP* Reading Outcomes

Our third analysis addressed the relationship between DSTP Reading and STAMP Reading outcomes. Initially, we were interested in whether the condition's effect varies across DSTP Reading as well as whether *PD* was differentially effective for low- and high-performing students. As a question independent of our experiment on *PD*, we also wanted to know whether achievement in world languages (as measured by STAMP) was related to level of English reading ability.

Table 16. The Relationship between STAMP Reading Posttest and DSTP Reading Pretest

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Predicted value for a control student with an average pretest	-0.85	0.56	9	-1.51	.17
Impact of *PD*	0.02	0.15	9	0.12	.91
Predicted change in control outcome for each unit increase on the STAMP reading pretest	0.33	0.04	522	7.80	<.01
Predicted change in control outcome for each unit increase on the Delaware reading pretest	0.01	0.01	522	6.42	<.01
Random effects[b]	**Estimate**	**Standard error**		***z* value**	***p* value**
Teacher mean achievement	0.04	0.02498		1.66	.05
Within-teacher variation	0.30	0.01847		16.19	<.01

[a] Schools are also modeled as a fixed factor but not included in this table

[b] Teachers were modeled as a random factor.

Table 16 shows us that for each one-point increase on the pretest measure of English Reading, a student gains .01 points on the STAMP outcome, holding constant the pretest measure of world language reading achievement.

Issues of Reliability of the STAMP Test as Used in this Research

In this final section we raise some concerns about the STAMP test as used in this research. In closely examining the likelihood of students moving up from one level to another, we noticed that many students dropped back a level. Table 17 shows for each of the initial performance levels (x axis) the percent of students that found themselves at each of the levels (y axis) of the posttest.

First we can see that a majority of students stayed at the same level. For those who did move, they were very likely to drop a level. For example, for students with a pretest level of 3, 33.4% dropped one or more levels while only 18.6 moved up one or two levels. Understanding that these levels are not equal interval and that it may be harder to move from 3 to 4 than to move from 2 to 3, we were surprised to see a large number of students loosing ground. We do not believe that this reflects the actual success of teachers in teaching World Languages. We suspect that students were not as motivated at the end of the year as they were at the beginning perhaps because of other tests being given and of the fact that the stakes were low for this particular test. We were also concerned that a large percent of students did not move from the pretest level. A more fine-grained scale may prove more useful in assessment of growth both as a state assessment and for purposes of experiments on products and programs such as ClassPak designed to improve achievement.

In commenting on these issues of reliability, the project coordinator noted that language learning is a complex process where as students progress in the language acquisition process, they take risks and tend to make more mistakes. It is as if they were going back to the drawing table to think things out and solve communication problems. In the initial levels of language proficiency, students are called upon to *react* to given information using lower levels of learning, such as comprehension and memorization. As they move forward in the learning process, however, the level of difficulty increases and calls for the use of higher levels of learning as they *create* with language. These complexities may account for some of the falling back observed in the posttest. Finally, the project coordinator noted that STAMP assesses proficiency, which is a global assessment. As such, it does not delineate discrete point information and does not discriminate between subcategories.

Table 17. STAMP Test Results Showing for Each Level of the Pretest the Percent of Those Students in Each of the Levels of the Posttest

Posttest level showing percent of students at that level for each pretest level	**5**	0.0	0.6	3.9	4.8	33.3
	4	0.6	2.3	14.7	47.6	22.2
	3	4.8	20.4	48.0	28.6	33.3
	2	31.7	62.2	26.5	19.0	11.1
	1	63.0	14.5	6.9	0.0	0.0
		1	**2**	**3**	**4**	**5**
		Pretest level				

Discussion

Our goal in this research was to provide the World Languages and International Education Department of the Delaware Department of Education information that would be useful in determining the impact of professional development and support including ClassPak and the STAMP test for World Language teachers in their state. There were secondary questions of interest including the usability of the STAMP test as a measure of reading and writing proficiency in World Languages and the relationship between gains in proficiency and incoming English reading achievement.

In our experiment we did not find that *PD* led to differences in outcomes when compared to the performance of students taught without *PD*. This held for outcomes in reading as well as writing. We got the same result analyzing the data as categorical and as continuous. The one indication of an effect was for writing where there was possible evidence that for treatment, where a student starts in terms of achievement is less of a determinant of where a student ends up, compared to control

We are cautious about these results for two reasons. First, the sample size of students was large, but the number of teachers was fairly limited. The ability of an experiment to detect small differences is highly dependent on the number of "upper level" units (i.e., teachers in this case) and with the given sample of teachers, we would expect that we could detect a moderate size impact (in standard deviation units, effect sizes as large as or larger than .35). We are unlikely to detect effects smaller than this, should they exist. The observed differences, however, were very close to zero.

The second reason for caution was the outcomes as measured by the STAMP posttest. We were concerned that a large number of students had lower scores at the end of the year than at the beginning. We do not believe there was a failure of the teachers to increase or maintain student proficiency. It is more likely that students took the test more seriously at the beginning than at the end. In any case, the STAMP test is not very fine-grained. Almost all students fell within the first three levels. The fact that most students stayed at the same level from pre- to posttest, gave us no information about growth for those students.

The STAMP test did appear to give us some information on growth. The correlation of growth in World Language proficiency and English reading level can be interpreted as an indication of the validity of the STAMP test. In the absence of other tests standardized to clear norms, the state can potentially gain useful information through a broader testing program with STAMP. This potential will be increased, we believe, by attention to the student motivation in taking the test and by further development of the test to increase the differentiation of levels and eventually to providing a continuous growth scale measure that can be used more readily in studies such as the one we undertook.

References

Center for Applied Second Language Studies (2006). White paper on ClassPak and STAMP.

Delaware Department of Education (2006). Retrieved from http://www.doe.k12.de.us/ on Retrieved on March 17, 2006.

Learning Language Solutions (2006). Retrieved from http://www.onlinells.com/classpak.php ,on March 17, 2006.

Raudenbush, S.W. & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods (2nd ed.). Newbury Park, CA: Sage.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

U.S. Census Bureau (2007). Retrieved on June 30, 2006 from http://www.census.gov/.