Empirical Education

RESEARCH REPORT

Comparative Effectiveness of ASCD's *Understanding by Design and Differentiated Instruction* Programs

A Report of a Comparison Group Study in Griffin-Spalding County School System

Laurel Sterling
Andrew Jaciw
Boya Ma
Empirical Education Inc.

December 21, 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 Griffin-Spalding County School System for their assistance and cooperation in conducting this research. The research was sponsored by ASCD, which provided Empirical Education Inc. with independence in reporting the results.

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.

Executive Summary

Introduction

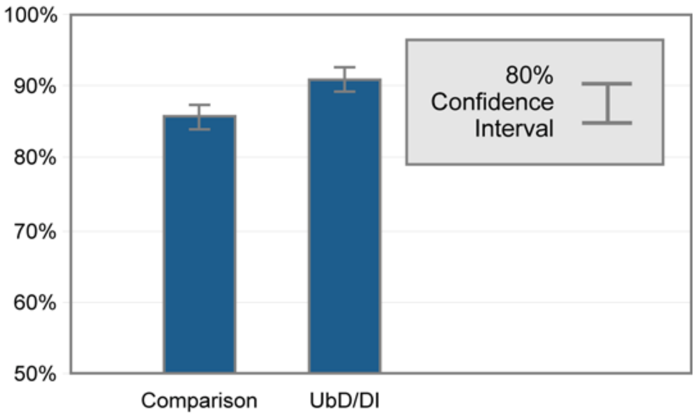
We report here on research aimed at producing evidence of the effectiveness of the professional development program that combines *Understanding by Design (UbD)* and *Differentiated Instruction (DI),* as implemented in Georgia's Griffin-Spalding County School System. The Association for Supervision and Curriculum Development (ASCD) contracted with Empirical Education for this study in order to detect differences in achievement between students at schools with *UbD/DI* and students at schools not using this professional development program.

Understanding by Design is a professional development program that trains instructors to design their own units in three stages: 1) Identify Desired Results, 2) Determine Acceptable Evidence, and 3) Plan Learning Experiences and Instruction. This is also referred to as a "backward design model." In Griffin-Spalding teachers were trained to nest *Differentiated Instruction* within the third stage of curricular design to help teachers plan lessons based on the particular needs of each student.

For this research, we used a comparison group design (also known as a quasi-experiment). Our comparison group was selected by using a matching process that first selected other districts in Georgia that share geographic proximity to Atlanta and then matched schools based on characteristics available on the state website—particularly reading scores and demographics. For each of the 11 Griffin-Spalding schools, we selected three matching schools that contained the same grade level as the one taught by the focal Teacher Leader. Since the implementation of the ASCD program began in the fall of 2005, we used test scores and other demographics from the spring of 2005 for the purposes of finding matches.

Findings

Our study yielded two main findings. We found a positive difference in reading that, even with the small sample of schools, provides some limited confidence that the program is associated with more students meeting the Georgia reading standards. In the test of English language arts, however, we found no difference between the Griffin-Spalding schools and the comparison schools in other districts. This figure shows the average percentage of students meeting or exceeding the state reading standards.



Relationship to CRCT Reading Achievement: Adjusted Means for Comparison and *UbD/DI*

The data on the extent of program implementation suggest at least a minimum level of implementation of the program among the Griffin-Spalding schools. Most Teacher Leaders participated in the training, created the requisite number of units, and provided *UbD* training in their own schools. However, the schools indicated concerns, especially regarding the amount of time required for implementation.

While our method took maximal advantage of the available data to find appropriate matches and to perform the appropriate statistical calculations, the comparison group design and the very small sample available in Griffin-Spalding put serious limitations on what we can conclude from this study. We cannot conclude that implementation of the program directly caused an improvement in reading achievement that would not have happened over that same period for other reasons. This is a basic weakness in any comparison group design.

A larger and more fine-grained sample, especially one taken from within a single district, would allow evaluations of school, teacher, or student differences that make a difference for the success of the ASCD program. The positive results for reading achievement warrants additional research using stronger controls including a richer set of student and teacher variables, a larger sample, and ideally an opportunity to randomly select schools within a district to implement the program.

Griffin-Spalding schools adopted the ASCD program with the intention of improving standards-based instruction. Regardless of whether the outcomes in reading are caused by the implementation of *UbD/DI*, the schools are delivering positive results in reading. The implementation of *UbD/DI* so far has not been as extensive as originally envisioned. Our recommendation to Griffin-Spalding, if the district decides to continue this program, is to ensure that all teachers receive the full ASCD training, and that they receive sufficient time and support to fully implement *UbD/DI.*

We do not recommend broader generalization beyond this particular district especially if the population differs in demographics or other standards from the limited sample in this study. The difference in results for reading and ELA suggests that further studies in a variety of jurisdictions with different standards will be important if we are to understand the areas of strength of this program and how it can be implemented to best effect.

Design and Analysis

The data for this study consist of student test outcomes and demographics provided by the district and similar school-level data obtained from the state website. To measure implementation, teachers were interviewed, and five web-based surveys were deployed. All comparison schools were called to determine that they were not implementing the ASCD program and the extent to which they implemented a backward design model for creating curricular units.

Statistical tests that took the school pretest scores into account were conducted using SAS PROC MIXED. The results must be qualified by the limitations in not having student-level data and working with percentages of students meeting each proficiency level rather than mean scale scores. The limitation of any comparison study of this kind is the inability to determine that the measured differences were directly a result of the program being implemented.

Comparative Effectiveness of ASCD's *Understanding by Design and Differentiated Instruction* Programs:

A Report of a Comparison Group Study in Griffin-Spalding County School System

Table of Contents

Introduction

We report here on research aimed at producing evidence of the effectiveness of the professional development program that combines *Understanding by Design (UbD)* and *Differentiated Instruction (DI),* as implemented in Georgia's Griffin-Spalding County School System. The Association for Supervision and Curriculum Development (ASCD) has contracted with Empirical Education for this study in order to detect differences in achievement between students at schools with *UbD/DI* and students at schools not using this professional development program.

During the 2005-2006 and 2006-2007 academic years the Griffin-Spalding County School System participated in a professional development program that combines *Understanding by Design and Differentiated Instruction*. The district adopted this program as a means of improving standards-based instruction. According to the ASCD Website,[1] *UbD* is "a framework for designing curriculum units, performance assessments, and instruction that lead your students to deep understanding of the content you teach." Specifically, training aims to teach instructors to design curricular units in three stages: 1) Identify Desired Results, 2) Determine Acceptable Evidence, and 3) Plan Learning Experiences and Instruction. This is also referred to as a "backward design model."

The following description of *Differentiated Instruction* was also taken from ASCD's Website:[2] *DI* is "an approach to teaching essential content in ways that address the varied learning needs of students with the goal of maximizing the possibilities of each learner." Instructors are expected to differentiate instruction in each of three areas: 1) student readiness,[3] 2) student interest,[4] and 3) student learning profile.[5] When integrated with *UbD, DI* is to be nested within the third stage of curricular design to help teachers plan the lessons based on the particular needs of each student.

Specifically, this study addresses the following question:

> Do students whose teachers have been trained in *Understanding by Design and Differentiated Instruction* professional development achieve higher Criterion Referenced Competency Test (CRCT) scores in reading and/or language arts than comparable students taught by similar teachers who have not received the training?

The study design is a comparison of two groups (also called a quasi-experiment) in which we started with 11 schools that had experienced the *Understanding by Design and Differentiated Instruction* program and then found matching schools that had not participated in this ASCD program. Identifying the appropriate matching units is an essential part of a comparison study such as this. For each of the focal schools, we used a matching process to select three schools in the same region that had similar demographics and similar reading scores at the focal teacher's grade level during the 2004-2005 academic year.

[1] Accessed 2/12/2007

[2] Accessed 2/12/2007

[3] The current knowledge, understanding, and skill level a student has related to a particular sequence of learning. (ASCD *Understanding by Design* On-Line Course, Lesson 2)

[4] What a student enjoys learning about, thinking about, and doing (ASCD *Understanding by Design* On-Line Course, Lesson 2)

[5] A student's preferred mode of learning (ASCD *Understanding by Design* On-Line Course, Lesson 2)

We designed our study to provide useful information to support local decisions that take into account the specifics of district characteristics and their implementation of the program. The results should not be considered to apply to school districts with practices and populations different from those in this study. The report provides a rich description of the conditions of implementation in order to assist the district in strengthening its program and to provide the reader with an understanding of the context for our findings.

Methods

Our study is a comparison of outcomes for schools where *Understanding by Design and Differentiated Instruction* were in place (the *UbD/DI* group) and schools using their current methods (*comparison* group). This section details the methods used to assess, with some level of confidence, the size of the difference between the groups. With this kind of study we have to keep in mind that, even where we detect a difference, factors other than *UbD/DI* may have been the actual cause of the difference. We begin with a description and rationale for the experimental design and next describe the intervention, the research sites, the sources of data, the composition of the experimental groups, and finally, the statistical methods used to generate our conclusions about the impact of *Understanding by Design and Differentiated Instruction*.

Experimental Design

Experiments are used to estimate the impact of an intervention on the basis of a sample. By this, we do not mean the impact of the intervention on the teachers and students in our sample. Instead, we mean the impact that the intervention would have on a larger population from which the sample was drawn (e.g., all the students or all the teachers in a state). The design of the experiment attempts to reduce bias and uncertainty and to make our impact estimates (based on the sample) as precise as possible. There is always a level of uncertainty and an associated level of imprecision. We think of the uncertainty as related to the likelihood that we would get a different result if we took a new sample of students or of teachers from the same larger population. Our design attempts to efficiently deploy the available resources to reduce uncertainty and improve precision, in other words, to reduce the likelihood that we would get a different result if we tried the experiment again.

An up-front effort to fully specify a design or plan for the experiment has two advantages:

- First, we identify, before seeing the outcomes, where we expect to see differences. In this way, we avoid fishing for results in the data, a process that can lead to mistaking chance differences for differences that are probably important as a basis for decisions. Because some effects will be large simply by chance, "mining" the data in this way can capitalize on chance; that is, we would conclude that there is an effect, when really we're just picking the outcomes that happen to appear large as a result of chance variation.
- Second, a study design will include a determination of how large the study should be in terms of students, teachers, and schools in order to get to the desired level of confidence in the results. In the planning stage of the experiment we calculate either how many cases we need to detect a specific sized difference between the *UbD/DI* and comparison groups, or how big a difference we can detect, given the sample size that is available. Technically this is called a power analysis. In this report, we explain how many aspects of design determine the size of the experiment.

Design Features to Address the Research Questions

How the *UbD/DI* Group was Identified

How the participants for a study are chosen can often generate a bias in the study. For example, where schools are chosen because of exceptional characteristics, it may be difficult to find comparable schools. In this case, ASCD identified a district that had already begun implementing the program and was interested in participating in the research. Because all the eligible schools in the district had already begun implementation, it was necessary to identify a comparison group outside the district.

Identifying the Comparison Sample

Since we want to know the impact of *Understanding by Design and Differentiated Instruction*, we attempt to isolate its effects from the other factors that might make a difference for how or what teachers and students do. This is always a challenge when the program group is selected before the comparison group is identified. The possibility always exists that factors related to the reason that the program group was selected, called “confounders,” rather than the program itself, account for the difference in outcome between the program and comparison group. For this study, because we had to find comparison schools outside the district, we knew from the beginning that district characteristics would be confounded with the program. We attempted to reduce these influences by considering potential comparison schools only in districts that were in close geographic proximity (in this case, the counties surrounding Atlanta) and that shared other cultural and economic characteristics. Aside from these characteristics that were not necessarily quantifiable, we also had a fair amount of quantitative information on demographics and test scores. These were used to identify (within the geographic region) schools that were as similar as possible to the *UbD/DI* schools in Griffin-Spalding.

The challenge to matching is to find cases that are similar to the treatment on factors that matter. As noted, these factors are technically known as confounders. Such factors are correlated with treatment (i.e., they are not balanced between conditions) and have a causal influence on the outcome. If we fail to identify all the critical confounders and control for their influence, our estimate of the treatment impact will be biased. The confounders can be either observed, in which case we can model them directly in order to control for imbalance on them between conditions, or unobserved. Though bias due to confounding of treatment with unobserved variables cannot be reduced through a statistical adjustment, other strategies can be used to minimize it.

There are various approaches to matching. Determining which is appropriate depends on several factors, including the sample size of treatment cases, the number of comparison cases, and the number of variables in terms of which we can perform the match.

In this study, the small number of treatment cases limits the matching strategy that we can use. The total number of treatment cases is small to begin with (N=11), and we subdivided this group further to perform matches within grades. We matched within grades because cross-grade matches would have been inappropriate: for one, the outcome scales were not vertically equated; moreover, grade level itself may be a confounder. In 7th grade we had only one treatment case to match control cases to. Fifth grade had the largest number of treatment cases, with five.

As described above, we matched grade-level teams to grade-level teams. This allowed us to control for potential selection at the school and grade levels. (*Controlling for selection* means that we control for unevenness on factors other than treatment that affect the outcome; that is, we aim to control for intrinsic factors that lead teacher teams to choose to participate in, or select into, treatment.) Specifically, we judged that it was more suitable to identify similar grade levels first, and then to consider a comparison of the performance of the treatment teachers’ students to the students of the comparison grade levels. We did this instead of matching each treatment teacher to a whole grade level, which would have hindered our efforts to partial out selection effects. Also, adopting the latter of these strategies would have made finding suitable matches more difficult. (Matching grade-level teams to other grade-level teams, but analyzing the difference between teacher-level scores in the treatment condition and team-level scores in the control condition, introduces complexities to the analysis which will be discussed later in this report.)

Our matching procedure can be divided into two stages. The first strategy was used to limit potential bias arising from unobserved confounders. The second strategy was used to limit potential bias due to confounding on variables that we observe.

In the first stage, we limited the pool of potential comparison cases. Following the results of Bloom et al. (2004), we focused on comparison cases that could be considered “local”[6]. In our study, although the outcomes are average achievement levels rather than earnings levels (as was the case in the Bloom et al. study), we believe that localness matters for a wide range of social outcomes when the goal is to reduce selection bias due to unobserved confounders. Localness can be defined variously. For the comparison group, we had available grade-level teams across the whole state of Georgia to choose from. We limited the pool of possible comparison cases to grade-level teams located in the south beltway region of Atlanta (bordered by I-20 to the north, Barnesville to the South, I-85 to the West, and I-75 to the East). There are more than 2000 grade-level teams in Georgia. We limited the comparison pool to 467 teams.

In the second stage of the matching procedure, from among the 467 potential comparison cases, we identified grade-level teams that were similar to the teams participating in the intervention. We did this by comparing each treatment case to all the comparison cases simultaneously on a series of variables that were available to match on. Information on these variables was obtained from publicly available datasets. The variables we matched on include the percentage of students scoring proficient in reading on the CRCT, the percentage of students participating in the National School Lunch Program, the percentage of African American students, the percentage of Hispanic students, and the student/teacher ratio.

The preferred approach to matching when a small number of individuals participate in the intervention is to find cases that are similar to the treatment cases simultaneously on all variables being modeled. In other words, we match treatment cases to control cases that are nearby in the covariate space that is defined by the variables in terms of which we are matching. One such type of “distance” is called the “Mahalanobis distance””. It has certain desirable properties; for instance, it adjusts for the fact that covariates are usually scaled differently.

We used the program *Match It,* written using *R* software, to identify comparison groups. Specifically, we used the “Optimal Matching” option in *Match It*. This form of matching is shown to be the best at minimizing the total Mahalanobis distances between treatment and control cases (Rosenbaum, 1989). We then matched each treatment case to three controls, based on the general finding that using more comparison cases than this does not improve precision significantly[7].

Organizational Levels Considered in the Experiment

This research works within the existing organizational hierarchy of schools, in which students are grouped under teachers who belong to schools. One level in the hierarchy, identified as the level or unit of analysis, is generally determined on the basis of the kind of intervention being tested. School-wide reforms call for a school-level analysis, while a professional development

[6] Bloom et al. showed that by restricting the comparison cases to ones that could be considered geographically local, variation across certain economic outcomes could be reduced in the absence of treatment even when modeling potential differences due to observable factors. This is equivalent to limiting selection bias, and suggests that keeping matches local reduces potential mismatch on unobservable factors.

[7] A limitation of the method just described is that we need more than one treatment case in order to carry out the matching algorithm. As stated earlier, we conducted separate matches within each grade level. For 7th grade we had only one treatment case. We therefore took a different approach: We considered the Euclidean distance of the treatment case to each of the comparison cases in the multivariate space of all matching variables, where these variables were all scaled the same way. We chose the three closest of these as the comparison set.

program can use a teacher-level analysis. For this study, our units for comparison were the grade levels within a school. In the *UbD/DI* schools, there is a Teacher Leader for each pair of grade levels: K-1, 2-3, and 4-5, except in the middle schools, where there is a Teacher Leader for each middle school department: English/language arts, history, math, and science. In addition, we were able to obtain state testing data separately for each grade level of the participating Teacher Leaders. So in effect, the grade level was our unit for comparing the *UbD/DI* schools to the comparison schools. While we did not have information more fine-grained than grade levels for the comparison schools, we were able to obtain class roster information for the *UbD/DI* schools. This allowed us to isolate and examine separately the outcomes for the Teacher Leader as well as the outcomes for the whole grade. In most cases, however, the Teacher Leader was the instructor for all the students at the particular grade level through formal or informal departmentalization. Consequently our comparison of program and comparison schools used data from the whole grade level. In terms of a hierarchy, we did not have to address the complexity of obtaining individual scores while examining differences at the school level. In all cases, we used the combined result for the whole grade (or alternatively, the combined result for the Teacher Leader's class).

How Small a Difference Can We Detect?

A process called power analysis is normally used to plan the number of schools that the experiment will need in order to say with any confidence whether the program has an impact of a certain size. For this experiment, however, we were limited to the number of schools available in the participating school district. For the comparison sample, we had a much larger pool of schools available in the nearby school districts, and we identified three comparisons for each *UbD/DI* school. The question then is not how many schools we need but, given the schools in the sample, how small a difference we can detect with confidence. This is important to clarify at the outset because there may be a real difference between schools using *UbD/DI* and similar schools that is smaller than can be detected. Some additional factors go into calculating what is called the Minimal Detectable Effect Size (MDES). The calculation is somewhat simpler because, in this case, we are not using student-level data directly, but rather grade-level aggregates. Therefore we do not have to ask how the variation is divided up among various levels. We are also not attempting to determine differential effects of the program on different subgroups of students; with the small numbers available, this would not be feasible. We do get considerable value from the pretest score or from the aggregate school score (for the program and comparison schools) for the year prior to the implementation of *UbD/DI* in Griffin-Spalding. The pretest is almost always the variable most closely associated with the outcome. In this case, the pretest is a "covariate." By including the covariate we can increase precision by "removing" this source of variation in the results. Technically, a covariate-adjusted analysis is called an analysis of covariance (or ANCOVA). In this study, we assumed a fairly substantial correlation between the pre- and posttests (.64). In a power analysis determining MDES, a good pretest correlation will increase precision and thereby require fewer Schools to detect the same level of difference.

How Confident Do We Want to be in the Results?

We describe uncertainty in terms of the likelihood that, if we ran the experiment again with a different sample from the same region, we would get the same result. Although the results would never be exactly the same, we can design the experiment so that the different results that we would get would likely fall within a certain range. An experiment that produces a very high level of confidence that the results would be very similar requires a larger number of units than an experiment that produces a lower level of confidence or a wider range of likely outcomes for the other hypothetical experiments. Because we can never be entirely certain of our results, the final step in the power analysis is to determine an acceptable or tolerable level of uncertainty. Conventionally, researchers have called for a high level of certainty, specifically, that getting a result like that observed would happen only 5% of the time if the program schools were not different from the comparison schools. For the purpose of the power analysis for this experiment, we used that criterion although, as we explain later, we report the results using a range of confidence levels.

Sample Size Calculation for This Experiment

Taking all the above factors into consideration, we estimated that 44 schools would constitute a sufficiently large sample to detect an effect size as small as 0.52.

Understanding by Design and Differentiated Instruction: Program Description

Participants were provided professional development, seven books, a binder of print materials, and access to the *UbD Exchange*. Trainers asked the Teacher Leaders to create curricular units and to submit these units to ASCD for review. Participants were expected to submit their units to the *UbD Exchange* and encouraged to access other teachers' units that are provided on the *Exchange*.

Training/Professional Development

ASCD provided 15 full days of professional development. The dates are listed in Table 1. Sessions began at 8:00 am and ended at 4:00 pm.

The first training session was designed as an overview for the school principals and central office administrators. All other sessions were primarily designed to train school teams; however, central office administrators were also expected to attend. The school teams were expected to consist of a school administrator and three or four teachers. The teachers on the teams are designated as "Teacher Leaders." The Teacher Leaders in the elementary schools were to each represent a different span of two grades (i.e., K-1, 2-3, 4-5). The middle school teams were expected to consist of one teacher from each of the four primary subject areas (i.e., mathematics, science, history, and reading/English language arts).

Table 1. Dates and Primary Participant for Professional Development

Date	Primary Participants
2005-2006 School Year	
August 30	Principals and Central Office Administrators
August 31	School Teams
September 1	School Teams
October 27	School Teams
November 10	School Teams
November 11	School Teams
January 9	School Teams
January 10	School Teams
May 16	School Teams
2006-2007 School Year	
September 29	School Teams
October 25	School Teams
December 1	School Teams
January 5	School Teams
February 2	School Teams
May 1	School Teams

The emphasis of the training was to prepare site teams to master the content and to redeliver the content (i.e. replicate the training) at their schools. All of the training sessions during 2005-2006 pertained to *Understanding by Design*. The sessions during 2006-2007 pertained to *Differentiated Instruction* and to integrating *UbD* with *DI*. Training topics included each of the three stages of curriculum design, differentiation strategies, assessment, rubrics, self-assessment, and preparation for redelivery and coaching. Each training session included time for participants to work on designing curricular units.

Understanding by Design and Differentiated Instruction Materials

In addition to a participant binder and handouts, participants receive the following books:

- John L. Brown, *Making the Most of Understanding by Design*
- Jay McTighe and Grant Wiggins, *Understanding by Design Professional Development Workbook*
- Carol Ann Tomlinson, *The Differentiated Classroom: Responding to the Needs of All Learners*
- Carol Ann Tomlinson, *Fulfilling the Promise of Differentiated Classroom: Strategies and Tools for Responsive Teacher*
- Carol Ann Tomlinson, *How to Differentiate Instruction in Mixed Ability Classrooms*
- Carol Ann Tomlinson and Jay McTighe, *Integrating Differentiated Instruction and Understanding by Design: Connecting Content and Kids*
- Grant Wiggins and Jay McTighe, *Understanding by Design, Expanded 2nd Edition*

Comparison Materials

Because comparison schools were not part of the study, we do not have information on their materials.

Implementation Schedule

Table 2 provides the major milestones of this study.

Table 2. Major Milestones of Study

Date	Major Milestone
November 2006	Initiation of one-year quasi-experiment
November – December 2006	Interviews with training facilitators
January 2007	First training observation
January – February 2007	Interviews with administrators
January – May 2007	Administration of monthly web-based teacher surveys
February 2007	Receipt of first dataset from district
March 2007	Selection of comparison schools
May 2007	Second training observation
September 2007	CRCT assessment data provided on state website

Site Descriptions

Recruiting

We invited all Teacher Leaders at each of the elementary and middle schools in the Griffin-Spalding district who teach reading and/or English language arts to students in grades 4-8 to participate in the study. This focal group was selected for several reasons, explained below.

- Although the *UbD/DI* program in Spalding is designed to impact all schools in the district, all grade levels, and each of the four major subject areas (i.e., mathematics, English language arts/ reading, science, and history), district personnel believed that the program was currently having the greatest impact in the elementary grades and in the subject area of English language arts and reading, and that it was having somewhat of an impact on the middle grades.

- Researchers chose to exclude Kindergarten through grade 3 because of limitations in gathering assessment data in lower grades.
- Although the program is also designed to impact non-Teacher Leaders, the first two years of the program focuses far greater resources in the training of Teacher Leaders than on the rest of the teachers in the schools.

The district sent out information regarding the study to the teachers in this focal group and to their principals on December 7, 2006. Researchers e-mailed all principals on December 12, and e-mailed the Teacher Leaders on December 20, to invite the Teacher Leaders to participate in the study. At the January *UbD/DI* training, all Teacher Leaders who fit the focus condition, together with administrators, met with the Research Manager from Empirical Education to discuss the research. At the end of the meeting, all Teacher Leaders who fit the focus condition were asked to sign consent forms to participate in the study.

On December 8 the district provided 15 teacher names and e-mails, which resulted in one teacher from each school, except for one school that had two names and one school that did not have any names (one of the schools did not have any Teacher Leaders who taught ELA in grades 4-8). From the 15 names that were provided by the district, one teacher was excluded because she is not a classroom teacher and two others were excluded because they do not teach either reading or English language arts. Another Teacher Leader was excluded from the study because her/his school did not exist prior to the 2005-2006 school year when the ASCD program began. This lack of data prior to the beginning of the study prevented us from selecting comparison schools for that particular focal school. Therefore, the resulting focal group consists of 11 Teacher Leaders: one teacher each from eight of the district's 11 elementary schools and one teacher each from three of the district's four middle schools.

Griffin-Spalding County School System

County

Spalding County is part of Georgia's Atlanta Metropolitan Area. The county encompasses 197 square miles and has an estimated population of 61,289, according to the 2005 census.[8] The median household income in the county in 2003 was $35,239. About two-thirds of the county's residents are white and about one-third are African American, whereas other racial groups constitute less than 5% of the population, as displayed in Table 3.

[8] Census data were accessed on the Web at http://quickfacts.census.gov/qfd/states/13/13255.html on February 19, 2007.

Table 3. County Racial Makeup

Ethnicity	Percentage of population
American Indian and Alaska Native	0.2%
Asian and Pacific Islander	0.8%
Black or African American	32.4%
Hispanic or Latino of any race	2.3%
White	65.8%
Two or more races	0.7%

City

Griffin is located 40 miles south of Atlanta and encompasses 14.6 square miles. According to the 2000 census, Griffin has a population of 23,451.[9] The median household income in Griffin in 2000 was $30,088. About half of the city's residents are African American, about half are white, and other racial groups constitute less than 5% of the population, as displayed in Table 4.

Table 4. City Racial Makeup

Ethnicity	Percentage of population
American Indian and Alaska Native	0.2%
Asian and Pacific Islander	1.0%
Black or African American	49.9%
Hispanic or Latino	2.2%
White	47.0%
Other	1.0%
Two or more races	1.0%

District

The Griffin-Spalding County School System includes schools in Spalding County and in the cities of Griffin, Orchard Hill, and Sunny Side. It serves PreK-12 students in 17 schools. The district consists of 11 elementary schools, four middle schools, and two high schools. One of the elementary schools is a charter school. The total district population consists of 10,967 students. About 58.5% of students are classified as economically disadvantaged.[10] Seven of the elementary schools and two of the middle schools are Title 1 schools. The racial composition of

[9] Census data were accessed on the Web at http://censtats.census.gov/data/GA/1601335324.pdf on February 19, 2007

[10] District data were accessed on the Web at http://schoolmatters.com on February 2, 2007.

the district, as displayed in Table 5, is similar to the demographics of the city of Griffin. English Language Learners constitute .8% of the district student population. In 2006, 85.2% of the district's students who were assessed on the Criterion Referenced Competency Test (CRCT) scored proficient or better in reading and 76.7% scored proficient or better in mathematics.

Table 5. District Racial Makeup

Ethnicity	Percentage of population
American Indian and Alaska Native	.1%
Asian and Pacific Islander	.8%
Black or African American	44.9%
Hispanic or Latino	3.2%
White	48.6%
Two or more races	2.4%

Data Sources and Collection

The data for this study includes demographics and student CRCT scores in reading and language arts. In addition, we conducted interviews with several educators:

- One principal or assistant principal from each of the Griffin-Spalding elementary and middle schools
- The Director of Elementary School Instruction
- The Director of Middle School Instruction
- The Director of Teacher Quality
- The Assistant Superintendent of K-12 Instruction
- Both professional development facilitators

We also collected five web-based survey responses from participating Teacher Leaders.

District Supplied Information

The data requested from the school district included records for the students who were taught by participating teachers as well as other background information. Specifically, the district was asked to provide the following data:

- Student name or unique ID
- Gender
- National School Lunch Program status (proxy for socio-economic level)
- Ethnicity
- Home language
- English learner status
- Disability status (whether or not student is has a disability or is in Special Education rather than the specific condition)

- Date of birth
- Classroom teacher
- Grade
- School the student attends

All student and teacher data having any individually identifying characteristics were stripped of such identifiers, and the data were stored using security procedures consistent with the provisions of the Family Educational Rights and Privacy Act (FERPA).

Achievement Measures

We used two outcome measures: CRCT Reading and CRCT ELA. We also used the CRCT Reading and ELA pretest scores, which were available from the spring 2005 testing. The assessments are criterion referenced. The state of Georgia updated the assessments in 2005 to match the state standards. The scores are also not vertically aligned, meaning that the scores for students at different grades are not on the same scale and cannot be combined for analysis. Following are the cut points:

Performance Level	2005	2006 and 2007
Does Not Meet the Standard	Below 300	Below 800
Meets the Standard	300 to 349	800 to 849
Exceeds the Standard	At or above 350	At or above 850

Observational and Interview Data

In addition to demographics and assessment scores, we also collected the following data over the entire period of the experiment to provide both descriptive and quantitative evidence of the implementation:

- Training observations
- Informal and formal interviews
- Five teacher surveys
- Email exchanges
- Telephone conversations

In general, observational data are used to inform further the nature of the training and the expectations for the Teacher Leaders. Training observations were conducted on January 5, 2007, and May 1, 2007. These data are minimally coded.

Interview data are used to provide an understanding of the program design and a description of the program implementation from the perspectives of the training facilitators, the central office administrators, and the site administrators. Half-hour interviews were conducted either in person at the schools or over the telephone between November 2006 and February 2007.

Survey Data

Surveys were deployed to Teacher Leaders on January 18, February 23, March 19, April 26, and May 23, 2007.

The survey topics were developed to account for the various aspects of teacher actions associated with instruction and learning. In order to characterize the average time teachers spent on specific activities, we used a repeated question strategy. The same questions were asked on all five surveys in order to gain an understanding of variation at different times during the school year. Survey topics consisted of the following:

- Assessment
- Instruction
- Materials
- Planning time
- Redelivery and coaching
- Self-assessment
- Support
- Teaching assignment
- Teaching background
- Unit development

The quantitative survey data were analyzed using descriptive statistics. The free-response portions of the surveys were minimally coded. We calculated response rates as simple percentages based on the ratio of actual received responses to the number of expected responses. There were 11 Teacher Leaders in the study. A total of five surveys were deployed with an overall response rate of 91%.

Characteristics of the *UbD/DI* and Comparison Groups

This section describes the sample that we used to determine the relationship of *Understanding by Design and Differentiated Instruction* to the measured outcomes. The sample consists of two groups of units. As described in the previous section, the program group included all Teacher Leaders who teach reading and/or language arts and who teach grades 4 through 8 in the Griffin-Spalding School District. All 11 Teacher Leaders remained in the program throughout the study.

We used a matching process to select three schools in the same region that had similar demographics and similar reading scores at the focal teacher's grade level in 2005. We inspect the final sample that is available for determining impact and check whether the *UbD/DI* and comparison groups are balanced on important characteristics. (For this accounting, we focus on the data available for CRCT Reading results, which we consider the primary outcome measure.)

After performing the matches, we checked whether there was balance between the treatment and comparison grade levels on the variables we used to form the matches. We found that at each grade level, the average values for the covariates are fairly similar for treatment and control cases. It is important to keep in mind that some of these factors are correlated and that finding the best match when all factors are considered simultaneously does not mean that we will have perfect matches on all factors. However, these matches are optimal in the sense that, if we try to decrease the remaining discrepancies, others will increase by a greater amount.

Table 6. Comparison of Experimental Groups in Grades 4 through 7

	No. of Schools	% Reading proficient	Students-teacher ratio	% Black	% Hispanic	% Low SES
Grade 4						
UbD/DI	3	79.67	16.43	25.47	4.30	53.60
Comparison	9	79.07	16.10	28.58	4.64	52.29
Grade 5						
UbD/DI	5	80.90	16.94	43.56	2.14	59.10
Comparison	15	80.88	17.35	48.11	2.05	60.93
Grade 6						
UbD/DI	2	73.95	17.55	45.00	2.35	61.10
Comparison	6	68.95	16.12	63.58	2.57	72.38
Grade 7						
UbD/DI	1	72.40	10.90	94.90	0.20	88.60
Comparison	3	69.17	16.00	95.97	0.87	89.83

Statistical Equations and Reporting on the Impact of *Understanding by Design and Differentiated Instruction*

Setting Up the Statistical Equation[11]

We put our data for schools into a system of statistical equations that allow us to obtain estimates of the direction and strength of relationships among factors of interest. The primary relationship of interest is the relationship of the program to a measure of achievement. We use SAS PROC MIXED (from SAS Institute Inc.) as the primary software tool for these computations. The output of this process are estimates of effects as well as a measure of the level of confidence we can have that the estimate is true of the population to which the experiment is meant to generalize.

Program Impact

A basic question for the experiment was whether following the intervention, students in *UbD/DI* schools had higher reading or ELA scores than those in comparison schools. Answering this question is not as simple as comparing the averages of the two groups. It is also essential that we understand how much confidence we can have that there really is a difference between the two groups, given our estimate of the size of the difference between the program and comparison groups that we obtain. To appropriately estimate this difference, our equation

[11] The term “statistical equation” refers to a probabilistic model where the outcome of interest is on the left-hand side of the equation and terms for systematic and random effects are on the right-hand side of the equation. The goal of estimation is to obtain estimates for the effects on the right-hand side. Each estimate has a level of uncertainty that is expressed in terms of standard errors or *p* values. The estimate of main interest is for the treatment effect. In this experiment, we model treatment as a fixed effect.

contains a term for *Understanding by Design and Differentiated Instruction* as well as terms for other important factors such as the student pretest score (at the school-level). The student's prior score is, of course, an important factor in estimating his or her outcome score. By including pretest as a term in the statistical equation, we are able to improve the precision of this estimate because it helps to explain much of the variance in the outcomes and makes it easier to isolate the program impact.

Fixed and Random Effects

The covariates in our equations measure either 1) fixed characteristics that take on a finite set of values (e.g., there are only two levels of gender) or 2) a set of characteristics that is assumed to have a distribution over a population and where we treat the values that we measure as though they were a random sample from that larger population. The former are called "fixed effects," the latter, "random effects." Because a small number of schools is involved, in our equations for this study, we have chosen to treat all the variables as fixed (except what is called the "error term" and shows on our results tables as the "residual"). Fixed effects produce less uncertainty but also limit the extent to which we can generalize our results.

The choice of terms for each statistical equation is not rigid but depends on the context and the importance of the factors for the question being addressed. The tables reporting the estimates resulting from the computation will provide an explanation of these choices in table notes where necessary for technical review.

Reporting the Results

When we run the computations on the data, we produce several results: Among them are effect sizes, the estimates for fixed effects, and *p* values. These are found in all the tables where we report the results.

Effect sizes

We translate the difference between program and comparison groups into a standardized effect size by dividing the average group difference by the amount of variability in the outcome. The amount of variability is also called the "standard deviation" and can be thought of as the average distance of all the individual scores from the average score (more precisely, it is the square root of the average of squared distances). Dividing the difference by the standard deviation gives us a value in units of standard deviation rather than units of the scale used by the particular test. This standardized effect size allows us to compare the results we find 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. When possible we also report the effect size of the difference after adjusting for pretest score and other fixed effects, since that adjustment provides a more precise estimate of the effect by compensating for differences in the average pretest of the program and comparison groups. Theoretically, with many replications of the experiment, these differences would wash out; therefore we would expect the adjusted effect size on average to be closer to the true value.

Estimates

We provide estimates to approximate the actual effect size. Any experiment is limited to the small sample of students, teachers, and schools that represent a larger population in a real-world (or hypothetical) setting. Essentially we are estimating the population value. When we report an estimate in a table, the value refers to the change in outcome for a one-unit increase in the associated variable. For example, since we code participation in the comparison group as 0 and participation in the program group as 1, the estimate is essentially the average difference that we expect in going from the comparison to the program group (while holding other variables constant).

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. Roughly speaking, it tells us the risk of concluding that the intervention has had an effect when in fact it has not. This mistake is also known as a "false-positive" conclusion. Thus a *p* value of .1 gives us a 10% probability of drawing a false-positive conclusion. This is not to be confused with a common misconception about *p* values: that they tell us the probability of our result being true.

We can also think of the *p* value 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 $p > .20$.

In reporting results with *p* values higher than conventional statistical significance, our goal is to inform the local decision makers with useful information and provide other researchers with data points that can be synthesized into more general evidence.

Results

Teacher-Level Implementation Results

In this section we describe the aspects of the implementation that characterize this intervention. First we list some key indicators of the extent of program implementation at the *UbD/DI* schools. Next we present the implementation data that we gathered for each indicator. In addition, we provide data regarding the implementation in terms of English/language arts instruction and regarding implementation over time. Finally, we summarize these results.

Key Indicators of the Extent of *UbD/DI* Implementation

Implementing new educational programs in schools is always a challenge. The implementation of the program that combines *Understanding by Design and Differentiated Instruction* in Griffin-Spalding may be particularly challenging because it is implemented across an entire school district for all schools, grade levels, and academic subjects. Additionally, this particular program poses its own set of challenges because it requires a deep understanding on the part of the teachers: Unlike other programs that prescribe what to teach and how to teach, this program requires that teachers determine what and how to teach, based on the state standards.

This type of program also poses challenges in terms of evaluating implementation, since neither the district nor the publisher describe a small set of explicit, measurable actions that would reveal an adequate or ideal implementation. However, from our review of the program literature and from our interviews with training facilitators and school and district administrators, we learned about some key indicators of the extent of program implementation. (In the cases where there were differences in the responses among the interviewees, the responses of the Assistant Superintendent of K-12 Instruction were taken as the definitive response.)

We list these indicators here and we then describe the implementation data for each of these indicators. The list, however, does not completely describe the actions that are needed for adequate or ideal implementation. Again, the program is dependent on each teacher's ability to determine specific actions she/he needs to take in order to meet the needs of her/his unique classroom. Therefore, these indicators suggest only whether the implementation reaches a minimum threshold.

Teacher Leader Indicators:

- Participate in all training sessions.
- Work together to deliver the content of the trainings to the teachers at their schools (also referred to as retraining or redelivery).
- Encourage the teachers in their schools to develop units and differentiate instruction.
- Support the teachers by answering their questions.
- Become competent users of *UbD* and *DI.*
- Develop two *UbD* units through all three stages in 2005-2006.
- Develop one *UbD/DI* unit through all three stages in 2006-2007.
- Deliver standards-based instruction in their classrooms by teaching *UbD* units and by differentiating instruction in their classrooms on the basis of student interest, readiness, and learning profile. The instructor should differentiate an equal amount in each of these three areas.

Indicators of Improvement:

In addition to the indicators listed above, we learned from our interviews that the expectation is not a set level of implementation, but rather increased implementation over time. Therefore we also track some of these indicators month by month.

Indicators of Implementation for English/language arts:

Since we measure outcome scores for ELA, we also attempt to understand the implementation in terms of ELA instruction.

Implementation Data by Indicator

Here we present the implementation data according to the key indicators.

Key indicator: Teacher Leaders participate in all training sessions

The average Teacher Leader in this study participated in 13 of the 14 training sessions or 94% of the training days. As displayed in Figure 1, all but one of the focal teachers attended at least 13 days of the training.

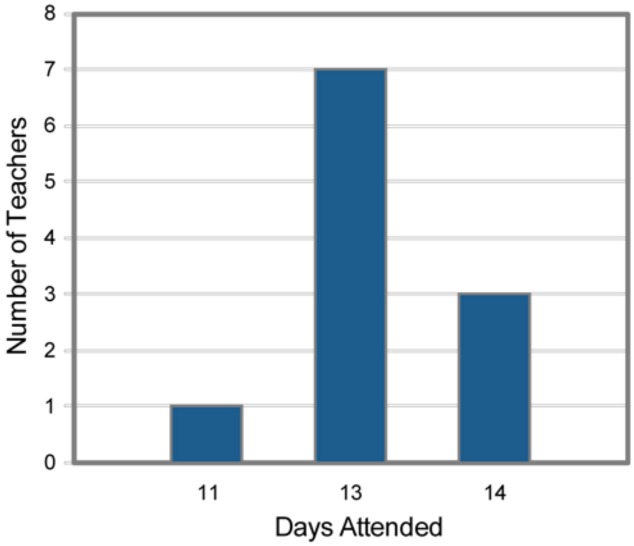


Figure 1. Teacher Leader Attendance at ASCD Training

Key indicator: Teacher Leaders work together to redeliver the content of the trainings to the teachers at their schools

During our interviews with site administrators in January 2007, one respondent indicated that the Teacher Leader (TL) at her pr his school who is participating in this study is not comfortable doing redelivery but that the other TLs at the school do provide training. The administrators from all other schools reported that their study TLs had been providing redelivery during the 2005-2006 academic year and during the first half of the 2006-2007 academic year. They described participants providing training at various times, including professional development days, common planning periods, faculty meetings, winter break, and after school.

Key indicator: Teacher Leaders encourage the teachers in their schools to develop units and differentiate instruction and support the teachers by answering their questions

During the principal/assistant principal interviews in January 2007, we asked whether the study Teacher Leaders provide coaching. We defined coaching as providing one-on-one assistance for implementing the ASCD program.

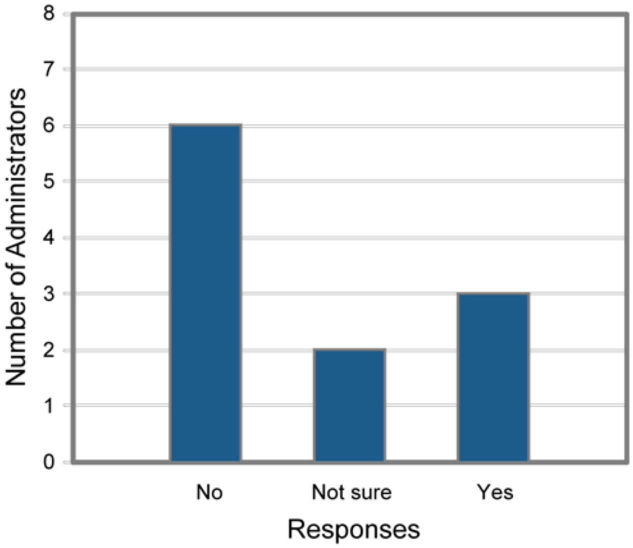


Figure 2. Administrator Response to Interview Question Regarding Whether Teacher Leaders Provide Coaching

As indicated in Figure 2, only 27% or three of the 11 administrators responded “yes.” In one case a school had a full-time coach until January of 2007. One of the district administrators responded that she did not believe that the TLs were providing coaching.

All Teacher Leaders indicated on their surveys that they provide some coaching, with the average TL providing less than half an hour of coaching per week of coaching. Between January and June 2007, the weekly average amount of coaching provided by each respondent ranged from 0.1 hours to 0.9 hours.

Table 7. Teacher Leader Reports Regarding Time Spent Coaching and Number of Teachers Coached

Teacher leader	No. of teachers coached	Mean weekly hours coaching
1	35	0.2
2	8	0.1
3	12	0.2
5	8	0.7
6	0	0.3
7	2	0.9
9	3	1.0
10	10	0.1
12	30	0.6
13	2	0.2

In apparent contradiction to this indication of time spent coaching, one of the same respondents indicated that the number of teachers she/he had coached was "0," as displayed in Table 7. In addition, one would expect that when there are large differences between the number of teachers coached (i.e., from 0 to 35), that there would likewise be correspondingly large differences between the time TLs spent coaching. However, as displayed in Table 7, there is no apparent pattern. The correlation coefficient between the number of teachers coached and the weekly average time coaching is -.21.

Discrepancies regarding coaching may be due to differing definitions of coaching. Principals who responded that their teachers provide coaching were careful to elaborate that the TLs either work with their grade level teams and, therefore, have the opportunity to work individually with teachers or that the TLs were frequently available to answer teachers' questions. It is possible that principals who responded "no" or "not sure" were using a more precise definition of coaching. For example, one site administrator who responded that his/her TL does coaching, was careful to clarify, "but it is not organized."

Key Indicator: Teacher Leaders become competent users of *UbD* and *DI*

While we are not able to evaluate the extent to which the TLs become competent users of *UbD/DI,* the data on the extent to which they create their own instructional units and to which they differentiate instruction do shed some light in this area.

Key Indicator: Develop two *UbD* units through all three stages in 2005-2006 and one unit through all three stages in 2006-2007

Only one of the ten Teacher Leaders who responded to the March survey indicated that she/he had developed fewer than three instructional units through all three stages, as displayed in Table 8.

Table 8. Teacher Leader Response to Survey Question Regarding Number of Units Developed

Teacher Leader	No. of units developed through Stage 3	No. of ELA units developed	No. of units developed and taught
1	2	2	2
2	4	3	2
3	6	3	0
5	7	7	7
7	3	0	2
8	2	2	2
9	3	3	2
10	3	1	3
12	3	0	2
13	4	4	2

Key Indicator: Deliver standards-based instruction in their classrooms by teaching *UbD* units and by differentiating instruction in their classrooms on the basis of student interest, readiness, and learning profile. The instructor should differentiate an equal amount in each of these three areas.

Also in Table 8, the majority of the TLs indicated teaching two of the units they had developed, although one TL reported teaching seven and another TL reported that she/he had not taught any of the units developed.

The surveys asked teachers each month whether they had differentiated instruction in terms of Readiness, Interest, and Learning Profiles. All TLs reported differentiating instruction based on student Readiness and Interest and all but one Teacher Leader reported differentiating based on student Learning Profiles. Participants reported differentiating the most in terms of Readiness and the least in terms of Learning Profiles.

We also surveyed Teacher Leaders to determine the extent to which they use assessment data to plan lessons. As displayed in Table 9, of the nine TLs who responded to the survey, two-thirds responded that they plan for both student groupings and instructional strategies based on diagnostic assessment data before at least half of their units.

Table 9. Number of Teacher Leaders by Response to Survey Question: How frequently do you develop a plan for both student groupings and instructional strategies based on diagnostic assessment data?

Before some of my units	Before about half of my units	Before most units	Before each unit or more frequently
3	2	3	1

Teacher Leaders also indicated that they tend to have a moderate level of comfort implementing the various aspects of the ASCD program. Table 10 shows that, among the nine respondents to this survey, only one TL reported a low level of comfort for the activities.

Table 10. Teacher Leaders' Self-Rating Regarding Their Comfort in Doing the Following Activities

Activity	None	Low	Moderate	High
Creating standards-based units using backward design	0.0%	11.1%	55.6%	33.3%
Assessing student readiness	0.0%	11.1%	66.7%	22.2%
Assessing student interests	0.0%	11.1%	66.7%	22.2%
Assessing student learning profiles	0.0%	11.1%	66.7%	22.2%
Modifying content based on assessment data	0.0%	11.1%	66.7%	22.2%
Modifying instructional strategies based on assessment data	0.0%	11.1%	55.6%	33.3%
Modifying products based on assessment data	0.0%	11.1%	55.6%	33.3%
Differentiating instruction in order to meet the learning needs of all your students	0.0%	11.1%	55.6%	33.3%

Note. Nine teachers responded to this survey question

Indicators of Reading/Language Arts Implementation

Teacher survey responses reveal that two of the Teacher Leaders did not develop any English/language arts units and only five of the TLs developed three or more ELA units. The numbers of Teacher Leaders who differentiated reading/language arts lessons at least once are as follows:

- 10 (91%) differentiated based on Readiness
- 11 (100%) differentiated based on Interest
- 9 (82%) differentiated based on Learning Profiles.

Indicators of Implementation over Time

There is evidence that little redelivery took place after the administrator interviews, during the second half of the 2006-2007 school year. The focus of the ASCD training in 2006-2007 was *DI.* However, trainers and administrators explained that, because the *DI* training was so new, they did not expect the TLs to be ready for redelivery until the end of the school year. At the January training, facilitators instructed the Teacher Leaders to not redeliver the *DI* content yet. In each of the five surveys that were administered during this period of time, we asked participants to estimate the number of hours of retraining they had provided during the past 10 full days of instruction. Only two TLs reported providing any retraining: one conducted one hour of redelivery in February and one reported six hours in March.

The coaching appeared to occur irregularly. When we asked TLs the number of hours of coaching they provided during the past two weeks, the most common response and the median response were both "0." Table 11 lists the weekly average hours of coaching by month. The average amount of coaching was highest in February, and no TLs reported providing any coaching in April or May.

Table 11. Weekly Average Hours of Coaching by Month

Jan	Feb	March	April	May
0.6	1.1	0.3	0	0

Table 12 displays the weekly average number of times teachers reported differentiating by month. The number of lessons the Teacher Leaders differentiated decreased in April and May.

Table 12. Weekly Average Number of Times Differentiating as Reported in TL Surveys

Differentiation	Jan	Feb	March	April	May
Readiness	3.0	1.6	2.1	0.6	1.0
Interest	1.8	1.5	1.9	0.6	0.8
Learning Profile	2.1	0.8	0.8	0.2	0.2

As explained previously, because neither the district nor the publisher describes a small set of explicit, measurable actions that would reveal an adequate or ideal implementation, we are unable to determine the extent of program implementation over time. While the data presented here might appear to indicate that the implementation has decreased over time, we do not have multiple measures of implementation over the course of the program's two years that would allow for triangulation. The administrator interview data reveal their perceptions based on practices during the first year and a half of the program, while TL survey data report their practices during the last half year. We also do not know whether redelivery would have increased if the trainers had not told the Teacher Leaders to not redeliver in January. Moreover,, the decrease in coaching and differentiation might be due to Spring Break, standardized testing, and end-of-year activities. In fact, several TLs reported on their surveys that this was the case. Therefore, here we analyze interview and survey data to attempt to understand the program's trajectory.

Interviewees and survey participants consistently reported that it is extremely important for teachers to differentiate instruction and to have access to a large bank of well designed units that teachers have created using backward design. Respondents state that these tasks are necessary for the delivery of the state standards. In fact, some of the schools report that they had begun training in these processes before the 2005-2006 school year. For example, one administrator explained that her/his school had purchased *UbD* books previously and that pieces of the program had begun to be implemented in 2003-2004. Another administrator revealed a high level of buy-in when she/he stated that the faculty embrace differentiation; that they have multiple levels of learning happening in the class at the same time; that instruction is not so teacher-driven; that teachers share with each other within grade levels; that they are creating student portfolios and centers; and that the students are helping each other.

At the same time, others express concerns. One TL wrote on the April survey, "I'm no longer sure about where our school is going with *UBD.* We are waiting for our principal to confirm where we are going before we continue with our faculty in this project." Another TL wrote the following comment on one of the surveys:

> *UbD* is a model of a teaching unit which incorporates many aspects important to the education of all students; however, the very best written *UbD* units will never solve the problem of the teacher who does not have "presence" in his/her classrrom [sic]. *UbD* provides a script or plan for a teacher, but it can never "teach" or "provide" all of the hidden secrets of being an effective teacher in a classroom. Until we embrace the problems that exist when teachers do not know how to relate to students, we will not begin to improve education. While *UbD* is a prescriptive teaching unit useful for a new teacher, it cannot "teach" the teacher how to deliver the information to the students. The most brilliantly written *UbD* units are only words on paper without effective classroom teachers to carry out the plan.

The biggest concerns revealed by study participants had to do with time and competing programs. When we asked the administrators about their biggest challenges, five of the site administrators and one of the directors listed time constraints. Similarly, a TL wrote the following on one of the surveys, “Time has and will always be an issue for teachers.”

Interviewees explained that the middle schools and at least two of the elementary schools have other new programs, and that teachers are challenged with deciding where to focus their time and energy. In addition, there are times when other programs conflict with *UbD/DI*. Most notably, one site administrator explained that she/he sends teachers to another training program that has a different template for creating units and that the teachers do not want to create the units in the manner prescribed by ASCD.

Other interviewees also indicated that they did not believe it is necessary to follow the *UbD* unit template. One administrator said that she/he will always design units and differentiation, but may not always use the ASCD template; she/he complained that the template is redundant: “If you put it on one page you have to say it again someplace else.” In addition, while several administrators reported that they embrace the concepts and practices promoted by ASCD, they believed that their school should move at a slow pace in order to gain a deep understanding rather than rush to implement every aspect.

One administrator stated that when observing teacher instruction, she/he looks for best practices, not for whether the teachers are implementing the units they designed. This administrator did not define best practices, but she/he was not referring exclusively to practices that had been trained for by ASCD. Similarly, another administrator stated that rather than looking for the designed units, he/she looks for the instruction of the standards. A third administrator stated that she/he doesn’t require a *UbD* lesson plan but is happy to see the teachers implementing some of the best practices from *UbD*.

Summary

The survey and interview data presented here suggest that the key indicators were in place at the majority of the study schools for at least part of the two years of the training program:

- 91% of the Teacher Leaders attended at least 13 of the 14 training sessions
- All but one of the TLs redelivered the training at their schools during the 2005-2006 school year and the first half of the 2006-2007 school year.
- The extent to which study participants provided one-on-one support to teachers in their schools is ambiguous, but all TLs claim to have provided a small amount of coaching.
- All but one of the TLs designed the requisite number of units and all but one of the TLs taught some of the units they had designed.
- All study participants report differentiating instruction based on student Readiness and Interest and all but one reported differentiating based on student Learning

Profiles. The amount of differentiation within each of the three categories was not exactly equal, but only one of the TLs failed to differentiate in all three areas.

- All participants report developing a plan for both student groupings and instructional strategies based on diagnostic assessment data before at least *some* of their units.
- All but one of the TLs indicated a moderate or high level of comfort implementing the program.
- Not all TLs implemented *UbD/DI* for English/language arts.
- There is no evidence that implementation increased over time. In fact, participants' concerns regarding time limitations and conflicting programs and their unwillingness to create units using the *UbD* template could point to a decrease in implementation.

Implementation at Comparison Schools

In this quasi-experiment, the comparison schools are not asked to participate in the study. Therefore, we are mostly limited to data that we can obtain from the Internet. However, researchers did attempt to reach an administrator at each comparison school by telephone to ask whether the school uses a backward design model. The reason for this was information that the state had adopted this model, which is a component of the ASCD program.

Table 13 reveals the result of the telephone calls. Most administrators (57%) indicated that they do use such a model. Respondents, however, did not all provide a strong indication that their teachers were faithfully implementing a backward design program. For example, one administrator stated that the teachers were "supposed" to use this design and that they had training during the 2006-2007 school year. Another answered that teachers had received training during the past two years from both the county and the state. A third administrator answered that she/he encourages teachers to write their units using backward design, that they had extensive training from the county, and that the administrator thought the backward design model was common all over the state. The two administrators who were coded as responding "maybe" indicated that their teachers had received training a couple years ago and that it was not clear how many of them were actually using this method. While backward design is a component of the program under study, there is little evidence that it was used extensively in the comparison schools. Insofar as it was a part of comparison school programs, it makes the current study a more stringent test of the unique characteristics of the ASCD program.

Table 13. Number of Comparison Schools Using a Backward Design Model, Based on Administrator Response to Question by Telephone

Yes	No	Maybe	No Response
16 (57%)	2 (7%)	2 (7%)	8 (29%)

Comparison of Programs Using School- and Student-Level Results

Overview

The primary goal of our study was to understand the relationship of *Understanding by Design and Differentiated Instruction* and student reading and English language arts achievement. For each of these areas we estimate the difference between *UbD/DI* and comparison schools: Within each content area we show whether being in a *UbD/DI* or a comparison class makes a difference for the average school. The units we compared are schools but more precisely, within each school, we focused on the grade level within which a Teacher Leader worked. As shown in Table 6, we matched comparison schools to these schools on the basis of their scores in the relevant grade and we used data only for that relevant grade in comparing the groups. In many of the *UbD/DI* schools, the TL taught all students in the grade. In some of the schools, the TL taught only one class. In all cases, however, we use the results for the whole grade in comparing to the corresponding whole grade in the comparison schools.

The basic unit of analysis is the school, which is the level at which we have results for the comparison group. However, we also estimate the association between student performance and program status. We express this association in units of standard deviation of student scores, along with the level of confidence that we have in the result. That is, in addition to figuring whether there is an association between average school performance and program status, we also consider whether there is an association between student performance and program status. The first of these is expressed in terms of the spread in average school scores; the latter of these is expressed in terms of the spread in student scores.

Association of *UbD/DI* and reading Achievement

Association of School-Level Outcomes With Program Status

We first address*Understanding by Design* and *Differentiated Instruction* outcomes using the CRCT Reading scale. Table 14 provides a summary of the sample we used and the results for the comparison of CRCT Reading scores for students in *UbD/DI* and comparison groups. The "Unadjusted" row gives information about the schools using the posttest means without any statistical adjustment. This shows the mean of the percentages of students meeting or exceeding the state standard as measured across schools within each condition, In the next column are the standard deviations of the percentages, also measured across schools in each condition This is followed by the counts for the number of schools in each condition. The last two columns provide the effect size, that is, an estimate of the difference in proportions for *UbD/DI* and comparison groups in standard deviation units[12]. Also provided is the *p* value, indicating 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. The "Adjusted" row is based on the same sample of schools. The means, and therefore the effect size, are adjusted to take into account the previous year's scores; hence, these statistics are adjusted for imbalances on the pretest between the two samples[13].

[12] There are different ways to compare proportions, and therefore different kinds of 'effect size'. To be consistent with our other reports, we have used a conventional approach of dividing the difference by the pooled standard deviation. This effect size is known as Hedges' g. Because we are using averages of proportions, by the Central Limit Theorem, we assume that average scores are approximately normally distributed. .

[13] In a quasi- experiment, inclusion of the pretest as a covariate in the statistical equation serves to increase the precision of the effect estimate but also potentially to reduce selection bias. The bias-reducing effect has been demonstrated consistently across studies (Glazerman, Levy, and Myers, 2003). With a relatively small sample of grade-level teams, such as the one we have to work with, overloading a regression model with many covariates to 'net out' their effect is infeasible. The number of potentially useful covariates that are available from public websites is small. Further, we have already established matches in terms of those variables through the initial matching process, so that modeling them explicitly may not add much (we verified that there is balance between the program and comparison cases on these covariates.) For the reasons noted above, our 'adjusted' result is based on a relatively simple model which includes the pretest, as noted above, as well as fixed effects for the matched groups (i.e., a separate indicator for each of the 11 sets of one program school and its three comparison cases). We block on matching groups to increase the precision of the estimate of the program effect by eliminating between-matched-group variability.

Table 14. Overview of Sample and Relationship Between of *UbD/DI* and Reading Achievement

	Condition	Means	School-level standard deviations	No. of schools	Effect size	*p* value[a]
Un-adjusted	***Comparison***	79.40	12.07	33	0.62	.01
	UbD/DI	86.27	5.52	11		
Adjusted	***Comparison***	86.00[b]	As above		0.47	.06
	UbD/DI	91.20				

[a] Because of the relatively small number of treatment schools in our sample we also checked the statistical significance of the results using a non-parametric test – the Wilcoxon Test. We compared the school averages of the percent of students meeting or exceeding the standard in the program and comparison schools, as well as the gains in these percentages (the former of these corresponds to the 'unadjusted' result above and the latter of these is similar to the 'adjusted' result given above, in the sense that it controls for the pretest). The former comparison yielded a *p* value of .25, the latter test yielded a *p* value of .12. Note that this result also does not figure in uncertainty due to variability in performance among students.

[b] The estimate of comparison group performance in the adjusted outcome is the median average value from among the matched cases.

Figure 3 is a view of the first two lines of Table 14 showing the average proficiency levels across the two groups of schools based on the result reported for spring 2007. The values in the table are for the combination of *Exceeded* and *Met the Standards* (the top two segments). These differences are then adjusted in a statistical equation that includes pretest resulting in the adjusted comparison.

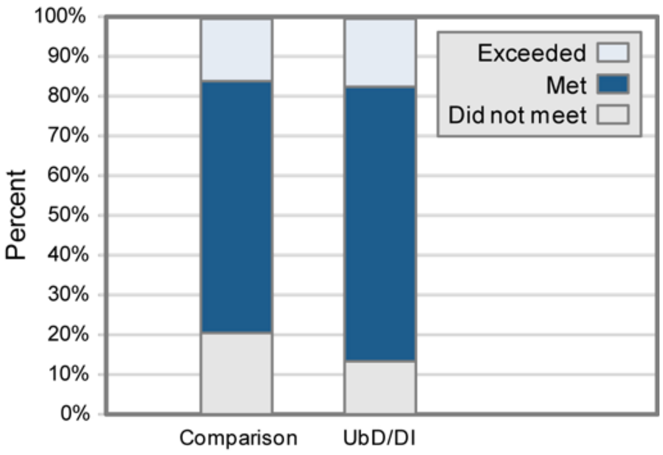


Figure 3. Unadjusted Comparison of the Comparison and *UbD/DI* Schools Showing the Three Levels of Proficiency

Table 15 shows all the parameter estimates (except for the fixed effects for clusters of matched cases) in the statistical equation that was used to compute the adjusted effect size. Included is the estimated difference between *UbD/DI* and comparison in the percentage of students achieving the standard in reading as measured by CRCT Reading.

Table 15. Association of *UbD/DI* and CRCT Reading Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for the comparison school with an average pretest	86.00	13.54	1	0.69	<.01
Change in outcome for the comparison school for each unit-increase on the pretest	0.97	0.19	1	5.20	<.01
Effect of *UbD/DI* for a school with an average pretest	5.20	2.70	1	1.93	.06
Random effects[b]	**Estimate**	**Standard error**	**DF**	***z* value**	***p* value**
Residual	58.96	14.98	31	3.94	<.01

[a] Although we also modeled differences among clusters of matched cases we do not exhibit these estimates in the table. By modeling these fixed effects the intercept value always represents the average for the comparison group for a particular matched cluster. We selected the cluster with the median value.

The estimate associated with *UbD/DI* is 5.20. This shows a small positive difference associated with *UbD/DI.* The *p* value of .06 indicates that we can expect to see a difference, as large as or larger than the absolute value of the estimate, 6% of the time when there truly is no effect. Using the criteria outlined earlier in the report, we conclude that we have some confidence that the association we are estimating is different from zero.

Figure 4 is a visual display of results from Table 15. It shows estimated performance on the posttest for the two groups based on a statistical equation that adjusts for students' pretest scores and other fixed effects. We added 80% confidence intervals to the tops of the bars in the figure. The lack of overlap in these intervals further indicates the level of confidence that we have that the true difference is not zero.

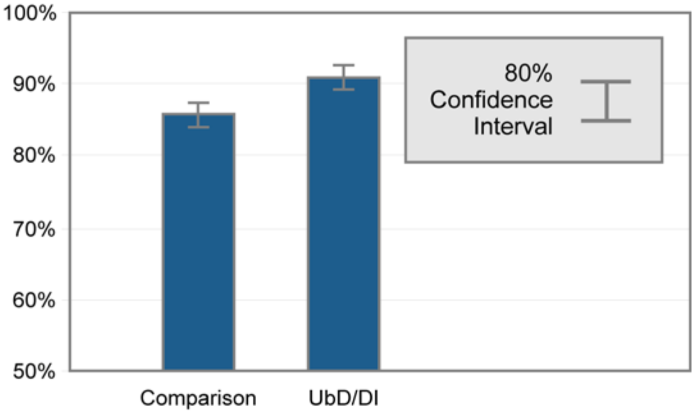


Figure 4. Relationship to CRCT Reading Achievement: Adjusted Means for Comparison and *UbD/DI*

Association of Student-Level Outcomes with Program Status

The calculations reported in Table 14 and Table 15 used the school-level data that was available from the state website. We also had individual student-level data for the Griffin-Spalding schools but this was aggregated to get a school level average. It is preferable to work with student-level data across the board because it is then possible to use information about clustering of students within the schools The resulting estimates are usually more conservative than the estimates based only on the school aggregate because they reflect the uncertainty due to a potential re-sampling students as well as schools. For this reason, we recalculated the results shown in the first two lines of Table 14 using an approximation of what student-level data would look like[14]. An effect size of .18 was thus obtained as the estimated difference in performance between students in the program and comparison condition divided by an estimate of the spread of student scores (i.e., the standard deviation) within schools. A *p* value of .18 was calculated considering both the variation in average school scores as well as the variation among student scores within schools. The *p* value of .18 gives us limited confidence that the true value we are trying to estimate is different from zero. While, we do not include a result that adjusts for pretest as is reported in Table 15, this approximation for the "unadjusted" result illustrates a reason to be cautious in the interpretation of the results from the school-level data available for this study.

Association of*UbD/DI* and English language arts Achievement

Association of School-Level Outcomes with Program Status

We address the outcomes for CRCT ELA in the same way as for reading. Table 16 provides a summary of the sample we used and the results for the comparison of CRCT ELA scores for students in *UbD/DI* and comparison groups. The "Unadjusted" row gives information about the schools using the posttest means without any statistical adjustment. This shows the mean of the percentages of students meeting or exceeding the state standard as measured across schools within each condition, In the next column are the standard deviations of the percentages, also measured across schools in each condition This is followed by the counts for the number of schools in each condition. The last two columns provide the effect size, that is, an estimate of the difference in proportions for *UbD/DI* and comparison groups in standard deviation units. Also provided is the *p* value, indicating 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. The "Adjusted" row is based on the same sample of schools. The means, and therefore the effect size, are adjusted to take into account the previous year's scores; hence, these statistics are adjusted for imbalances on the pretest between the two samples

[14] The student-level standard deviation was approximated using the variance formula for a binomially distributed random variable. The effect size is obtained by dividing the mean difference in percent by this student level standard deviation. To obtain the *p* value, we assumed that 10% of the variance in outcomes is between-schools, which we consider a plausible value given the partitioning of variance in this study and conventionally accepted values of the intraclass correlation.

Table 16. Overview of Sample and Relationship Between *UbD/DI* and English Language Arts Achievement

	Condition	Means	Standard Deviations	No. of Schools	Effect Size	*p* value[b]
Un-adjusted	***Comparison***	81.45	8.68	33	0.34	.22
	UbD/DI	84.27	5.59	11		
Adjusted [a]	***Comparison***	82.00[b]	As above		0.12	.63
	UbD/DI	83				

[a] Because of the relatively small number of treatment schools in our sample we also checked the statistical significance of the results using a non-parametric test – the Wilcoxon Test. We compared the school averages of the percent of students meeting or exceeding the standard in the program and comparison schools, as well as the gains in these percentages (the former of these corresponds to the 'unadjusted' result above and the latter of these is similar to the 'adjusted' result given above, in the sense that it controls for the pretest.) The former comparison yielded as *p* value of .42, the latter test yielded a *p* value of .65. Note that this result also does not figure in uncertainty due to variability in performance among students.

[b] The estimate of comparison group performance in the adjusted outcome is the median average value from among the matched cases.

Figure 5 is a view of the first two lines of Table 16 showing the average proficiency levels across the two groups of schools based on the result reported for spring 2007. The values in the table are for the combination of *Exceeded* and *Met the Standards* (the top two segments). These differences are then adjusted in a statistical equation that includes pretest resulting in the adjusted comparison.

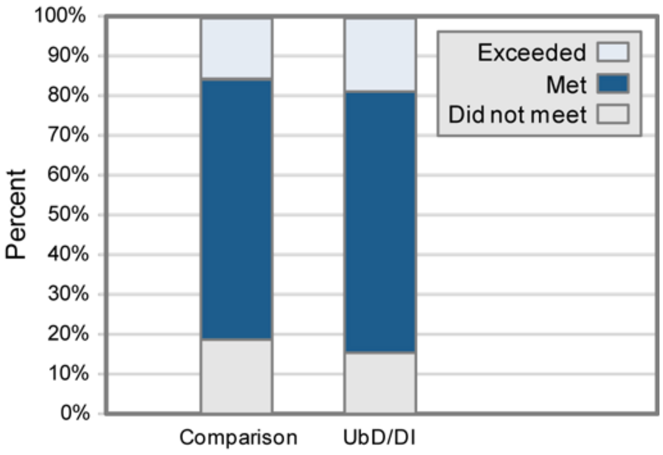


Figure 5. Unadjusted comparison of the comparison and *UbD/DI* schools showing the three levels of proficiency

Table 17 shows all the parameter estimates (except for the fixed effects for clusters of matched cases) in the statistical equation that was used to compute the adjusted effect size. Included is the estimated difference between *UbD/DI* and comparison in the percentage of students achieving the standard in ELA as measured by CRCT Reading.

Table 17. Association of *UbD/DI* and CRCT ELA Achievement

Fixed effects[a]	Estimate	Standard error	DF	*t* value	*p* value
Outcome for the comparison school with an average pretest	82.00	8.21	1	4.79	<.01
Change in outcome for the comparison school for each unit-increase on the pretest	0.58	0.12	1	4.65	<.01
Effect of *UbD/DI* for a school with an average pretest	1.00	2.04	1	0.49	.63
Random effects[b]	**Estimate**	**Standard error**	**DF**	***z* value**	***p* value**
Residual	32.98	8.38	31	3.94	<.01

[a] Although we also modeled differences among clusters of matched cases we do not exhibit these estimates in the table. By modeling these fixed effects the intercept value always represents the average for the comparison group for a particular matched cluster. We selected the cluster with the median value.

The estimate associated with *UbD/DI* is 1.00. This shows a small positive difference associated with *UbD/DI.* The *p* value of .63 indicates that we can expect to see a difference as large as or larger than the absolute value of the estimate 63% of the time when there truly is no effect. Using the criteria outlined earlier in the report, we conclude that we have no confidence that the association we are estimating is different from zero.

Figure 6 is a visual display of results from Table 17. It shows estimated performance on the posttest for the two groups based on a statistical equation that adjusts for students' pretest scores and other fixed effects. We added 80% confidence intervals to the tops of the bars in the figure. The overlap in these intervals further indicates that we have no confidence that the true difference is non-zero.

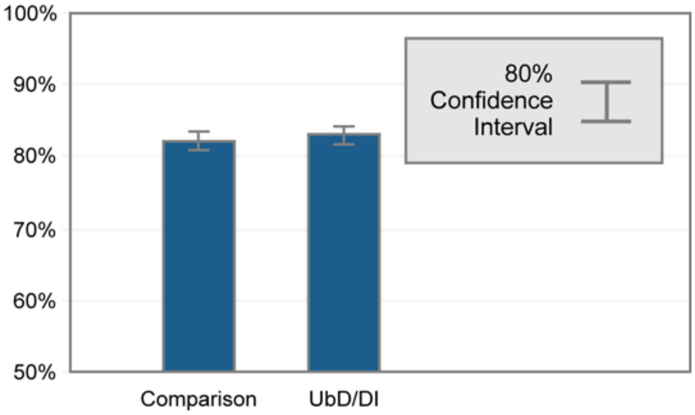


Figure 6. Relationship to CRCT ELA Achievement: Adjusted Means for Comparison and *UbD/DI*

The estimate associated with *UbD/DI* is 1.00. This shows a small positive difference associated with *UbD/DI.* The *p* value of .63 indicates that we can expect to see a difference as large as or larger than the absolute value of the estimate 63% of the time when there truly is no effect. Using the criteria outlined earlier in the report, we conclude that we have no confidence that the association we are estimating is different from zero.

Association of Student-Level Outcomes with Program Status:

As with the results for reading, we also calculated an approximation of the difference it would make if we were to have student-level data for all the schools. The recalculation (without adjustment for pretest) gives us an effect size of 0.07. The *p* value of .58 gives us no confidence that the true value we are trying to estimate is different from zero. As with the reanalysis for the reading results, there is reason to believe that the effect size and our level of confidence would be lower if student level data were available.

Discussion

Our study addresses the effectiveness of the professional development program that combines *Understanding by Design* and *Differentiated Instruction*. Specifically, we began our research by asking whether students in Georgia's Griffin-Spalding County School System whose teachers participated in these programs provided by the Association for Supervision and Curriculum Development (ASCD) were more likely to meet the Georgia state standards as measured by the Criterion Referenced Competency Test (CRCT) in reading or language arts than comparable students in similar districts taught by teachers who had not received the training. We focused on teachers trained as Teacher Leaders considering the more general effect on the students in the grade-level in which they taught.

For this research, we used a comparison group design (also known as a quasi-experiment). Our comparison group was selected by using a matching process that started with identifying other districts in Georgia that shared geographic proximity to Atlanta and went on to take advantage of the characteristics available on the state website—particularly reading scores and demographics. For each of the 11 Griffin-Spalding schools, we selected three matching schools that contained the same grade level as the one taught by the focus teacher leader. Since the implementation of the ASCD program began in the fall of 2005, we used test scores and other demographics from the spring of 2005 for the purposes of finding matches.

Our quasi-experiment yielded two main findings. We found a positive difference in reading that, even with the small sample of schools, provides some limited confidence that the program is associated with more students meeting the Georgia reading standards. In the test of English language arts, however, we found no difference between the Griffin-Spalding schools and the comparison schools in other districts.

The data on the extent of program implementation suggest at least a minimum level of implementation of the program among the Griffin-Spalding schools. Most Teacher Leaders participated in the training, created the requisite number of units, and provided *UbD* training in their own schools. However, the schools indicated concerns, especially regarding the amount of time required for implementation.

While our method took maximal advantage of the available data to find appropriate matches and to perform the appropriate statistical calculations, the comparison group design and the very small sample available in Griffin-Spalding put serious limitations on what we can conclude from this study. We can't conclude that implementation of the program directly caused an improvement in reading achievement that would not have happened over that same period for other reasons. The fact that the district chose to implement the ASCD program may be related to a characteristic of the district that predispose the participants to a more positive outcome. Districts that are similar to this one in other observable respects may not have this critical unobservable characteristic and may not do as well. This is a basic weakness in any comparison group design. There are a multitude of factors that may distinguish Griffin-Spalding from otherwise similar Georgia districts.

We are also limited to a small number of program schools and are limited in not having data on individual students in the comparison schools. A larger and more fine-grained sample, especially one taken from within a single district, would allow evaluations of school, teacher, or student differences that make a difference for the success of the ASCD program. The positive results for reading achievement warrants additional research using stronger controls including a richer set of student and teacher variables, a larger sample, and ideally an opportunity to randomly select schools within a district to implement the program. Assignment of cases to conditions through a random selection process, such as a coin toss, which gives an even chance of any participant joining the program or the comparison group eliminates variables other than the program that can explain a difference in results.

Griffin-Spalding schools adopted the ASCD program with the intention of improving standards-based instruction. Regardless of whether the outcomes in reading are caused by the implementation of *UbD/DI*, the schools are delivering positive results in reading. The implementation of *UbD/DI* so far has not been as extensive as originally envisioned. Our recommendation to Griffin-Spalding, if the district decides to continue this program, is to ensure that all teachers receive the full ASCD training, and that they receive sufficient time and support to fully implement *UbD/DI.*

We do not recommend broader generalization beyond this particular district especially if the population differs in demographics or other standards from the limited sample in this study. The difference in results for reading and ELA suggests that further studies in a variety of jurisdictions with different standards will be important if we are to understand the areas of strength of this program and how it can be implemented to best effect.

References

Brown, John L. (2004). *Making the Most of Understanding by Design.* Alexandria, VA: Association for Supervision and Curriculum Development.

Glazerman, S. Levy, D., & Myers, D., (2003). *Nonexperimental Versus Experimental Estimates of Earnings Impacts*, American Academy of Political & Social Science, Vol. 589, No.1, 63-93.

McTighe, Jay & Wiggins, Grant. (2004). *Understanding by Design Professional Development Workbook.* Alexandria, VA: Association for Supervision and Curriculum Development.

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

Tomlinson, Carol Ann. (2003). *Fulfilling the Promise of Differentiated Classroom: Strategies and Tools for Responsive Teacher.* Alexandria, VA: Association for Supervision and Curriculum Development.

Tomlinson, Carol Ann. (2004). *The Differentiated Classroom: Responding to the Needs of All Learners.* Alexandria, VA: Association for Supervision and Curriculum Development.

Tomlinson, Carol Ann. (2004). *How to Differentiate Instruction in Mixed Ability Classrooms.* Alexandria, VA: Association for Supervision and Curriculum Development.

Tomlinson, Carol Ann & McTighe, Jay. (2006). *Integrating Differentiated Instruction and Understanding by Design: Connecting Content and Kids.* Alexandria, VA: Association for Supervision and Curriculum Development.

Wiggins, Grant & McTighe, Jay. (2005). *Understanding by Design, Expanded 2nd Edition.* Alexandria, VA: Association for Supervision and Curriculum Development.