



THE EFFECTIVENESS OF THE TEACHER ADVANCEMENT PROGRAM

Teacher excellence... Student achievement... Opportunities for all

**Revised
April 2007**

by Lewis C. Solmon, J. Todd White, Donna Cohen and Deborah Woo

National Institute for Excellence in Teaching



National Institute for
Excellence in Teaching

The Effectiveness of the Teacher Advancement Program*

by Lewis Solmon, Todd White, Donna Cohen, and Deborah Woo

Executive Summary

The Teacher Advancement Program (TAP) is a comprehensive school reform aimed at restructuring and revitalizing the teaching profession while attaining measurable gains in student achievement. TAP includes multiple elements. Many of these elements have been tried in isolation in the past and have not resulted in student achievement gains. Our innovation changes schools' organizational structure and included key elements to attract, retain, develop, and motivate quality teachers with the ultimate goal of increasing student achievement and closing achievement gaps. TAP's four elements are: **(1) Providing multiple career paths** which enable teachers to advance while staying in classroom, and also providing opportunities for shared instructional leadership—principal cannot do it all alone; **(2) Introducing ongoing applied professional growth** to help all teachers improve instruction by working on their specific needs, as determined by analyzing their classroom performance evaluations and their students' data. We believe even good teachers can become great, and great teachers can become even more effective; **(3) Increasing instructionally focused accountability**. To be fair there are multiple (at least four) evaluations for all teachers by trained and certified evaluators (master teachers as well as mentor teachers and the principal) based on clearly defined scientifically validated teaching rubrics. This type of accountability can identify effective teachers and can also determine who needs to improve; **(4) Providing performance-based compensation** rewards to teachers for hard work if they are successful, for taking on additional responsibilities, for their performance as determined by multiple evaluations, and for the performance of their students as determined by pre- and post-test outcomes. TAP now operates in over 130 schools in 14 states and the District of Columbia.

The purpose of our evaluation paper is to analyze the impacts of TAP. The research question we ask is: If a school implements TAP, are its teachers more likely to outperform—in terms of value-added gains—similar teachers not implementing TAP, and, are TAP schools likely to outperform non-TAP schools? Our evaluation of TAP is multifaceted, first comparing student achievement gains of individual teachers and schools to similar, non-TAP teachers and schools. We also considered adequate yearly progress (AYP) of TAP schools and their states overall, as well as teacher attitudes towards elements of the program.

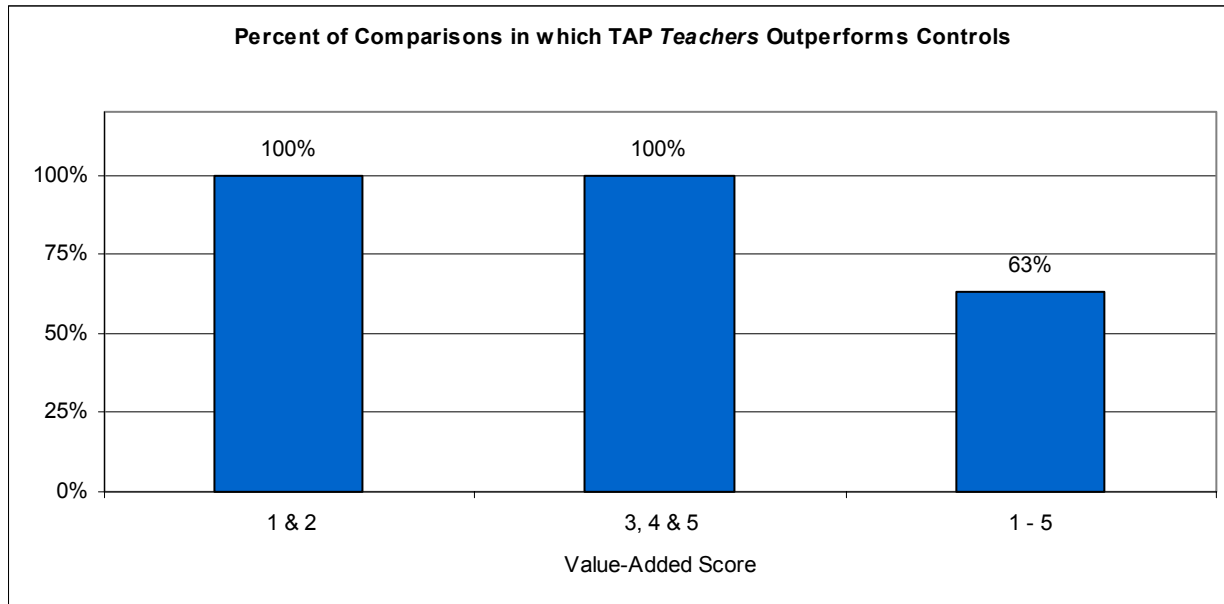
We analyzed the 2004-2005 student achievement gains at two levels of comparison—teacher-to-teacher and school-to-school. SAS® EVAAS®, a system developed by William Sanders and now used by Sanders at the SAS Institute Inc., uses student test score data from TAP schools and control schools to calculate individual teachers' value-added gains in order to determine individual performance bonuses for TAP teachers, and the school-wide gains for school-wide bonuses. A by-product of these calculations is the ability to compare student achievement growth from TAP teachers and schools to such growth from control teachers and schools.

* Authors would like to acknowledge the contributions of June Rivers, William Sanders, Eric Hanushek, Kimberly Firetag Agam, Diana Wardell, Geneva Galloway, and staff of the National Institute for Excellence in Teaching.

In evaluating TAP teachers, and similarly TAP schools, we calculate the effect of each teacher on student progress as assessed by the difference between the actual average scores of the teacher's students and the expected average scores of those students (as derived from previous scores). Then we compare this effect to similar calculations of the difference between the actual and expected average scores of the control group. By dividing the individual teacher effect by the associated standard error we can determine how many standard error units a particular teacher's effect is from the growth average, and then can place each teacher in one of five categories—below the average teacher's estimate (score of 1 and 2) or at or above the average teacher's estimate (score of 3-5). The standard error units calculated for each teacher enable us to determine what proportion of the teachers (TAP and otherwise) do *statistically* significantly better than average and what proportion do *statistically* significantly worse than the growth average as determined by the control teachers. In other words, we examined whether or not the growth a teacher makes with her students is different from the average amount of growth and with how much confidence we can say so.

Under each of the five categories, we noted which of the two groups, TAP or control, outperformed the other in each state. In categories “1 and 2” the “outperforming” group is the one with the smaller of the two percentages, meaning that fewer teachers produced less than an average year's growth. In categories “3, 4, and 5” we noted which group had the higher of the two percentages, meaning that more teachers produced an average year's growth or more in their students' achievement. This is documented in the following summary charts.

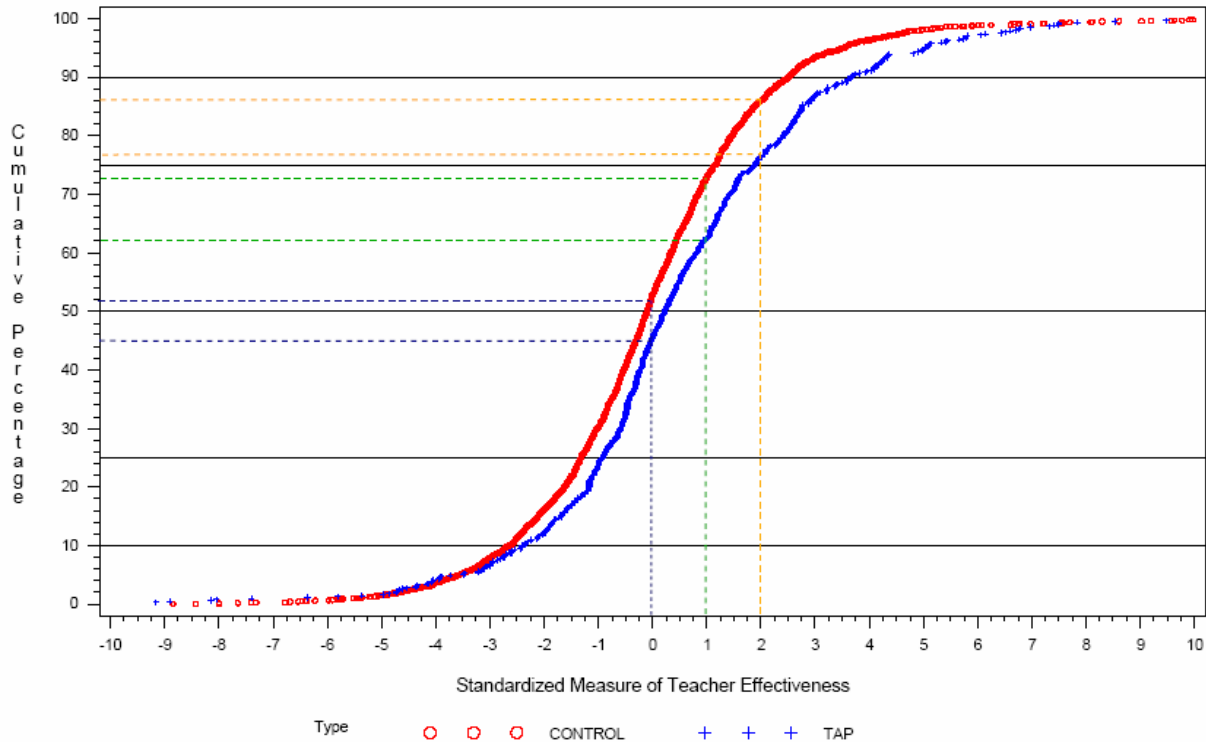
In almost two thirds (63%) of the comparisons of whether TAP teachers outperformed control teachers in each separate growth level (1-5), TAP teachers came out on top across the six states in the study. All states have a smaller percentage of TAP teachers scoring a “1 or 2” than controls, which means that fewer TAP teachers were significantly less effective in raising their students' scores than control teachers. To clarify, fewer TAP teachers had students whose progress was below average. Additionally, we found that in all states a higher percentage of TAP teachers scored a “3 or above” than their controls, meaning more TAP teachers were significantly more effective in raising their students' scores than control teachers and more TAP teachers had students scoring at or above the average year's growth.



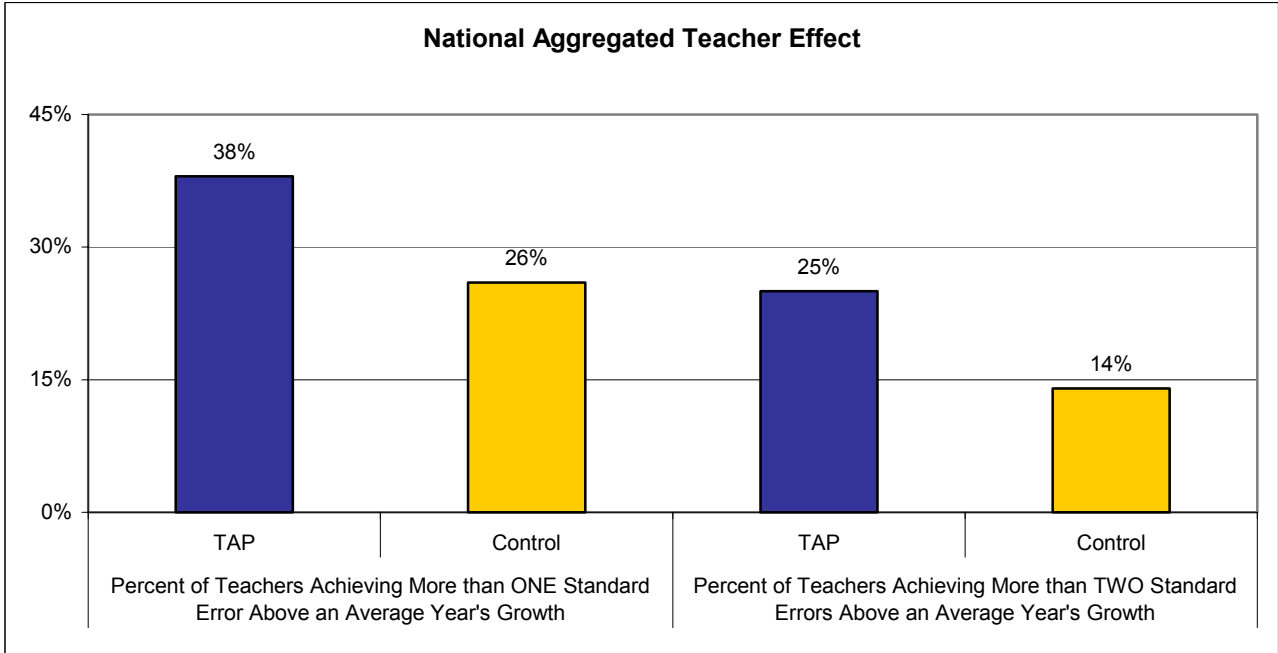
These results are very positive, clearly demonstrating that TAP teachers produce higher student achievement growth than similar teachers not in TAP schools.

Next, SAS® EVAAS® calculated a standardized measure of teacher effectiveness that includes all 610 TAP teachers from the six states in the study and 2,337 control teachers from the same states. They then produced a cumulative distribution that is shown below.

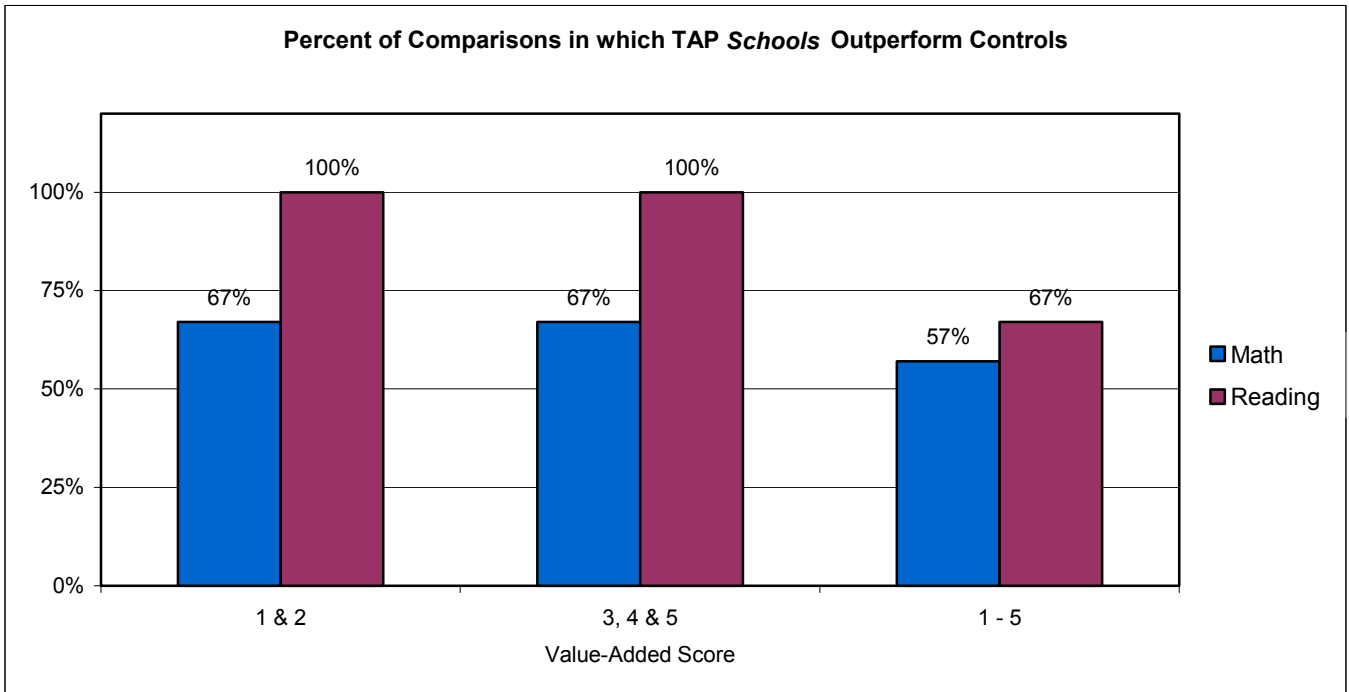
TAP Teachers vs Control Teachers Cumulative Distribution Comparative Plot Standardized Teacher Effectiveness Estimates



By drawing a vertical line from any point on the horizontal axis (which indicates the level of teacher effectiveness) to either of the cumulative distribution lines, we can see what percentage of TAP or control teachers achieved that level of effectiveness *or less*. The one standard error point on the horizontal axis indicates that when applying one standard error to the teachers' estimates, 62% of TAP teachers and 74% of control teachers had estimates that indicated their average student progress was *at or below* the average gain. It is then easy to calculate that 38% of TAP teachers as compared to 26% of control teachers had estimates that indicated their average student progress was *above* the average gain. Using the same method, the two standard errors point on the horizontal axis indicates that when applying two standard errors to the teachers' estimates, 25% of TAP teachers as compared to 14% of control teachers had estimates that indicated their average student progress was *above* the average gain. This is illustrated in the following chart.



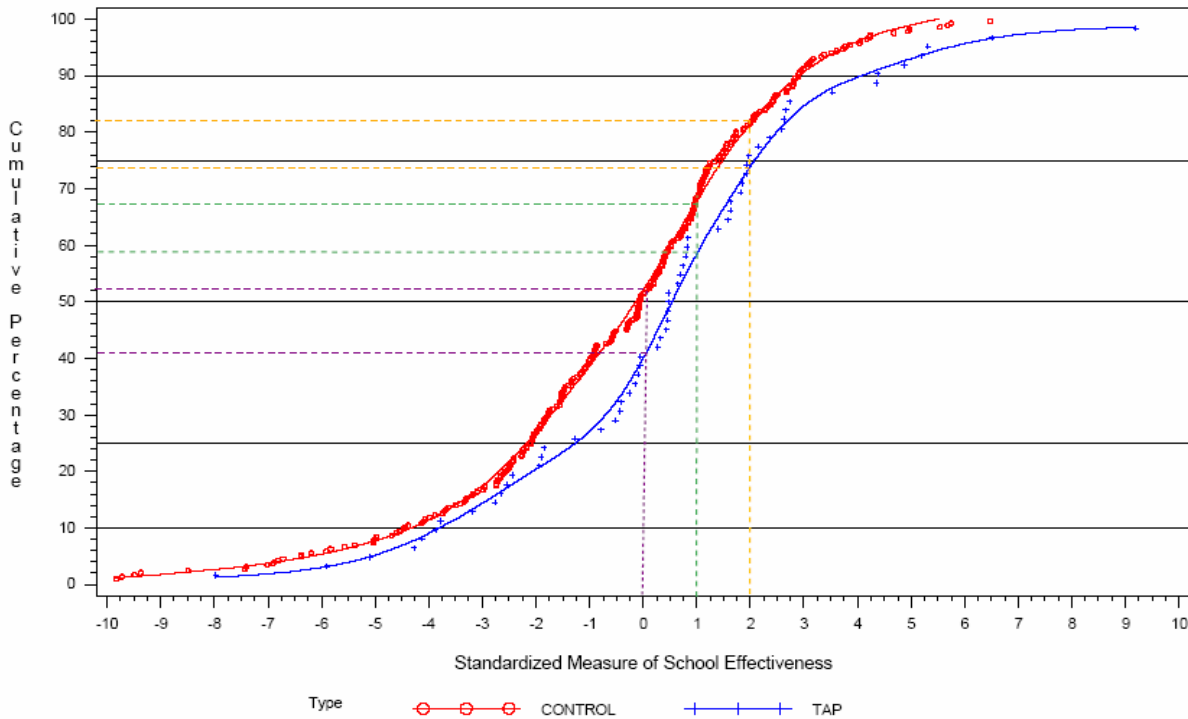
SAS® EVAAS® also provides *school-wide* gains of TAP schools as compared to control schools. TAP schools outperformed their controls in 57% of the individual categories (1-5, by state) in math and in 67% of the categories in reading. In the comparison of the percentage of schools scoring below the average, and the percentage of schools scoring at or above the average, TAP schools outperform their controls in 67% of the categories in math and in 100% of the categories in reading.



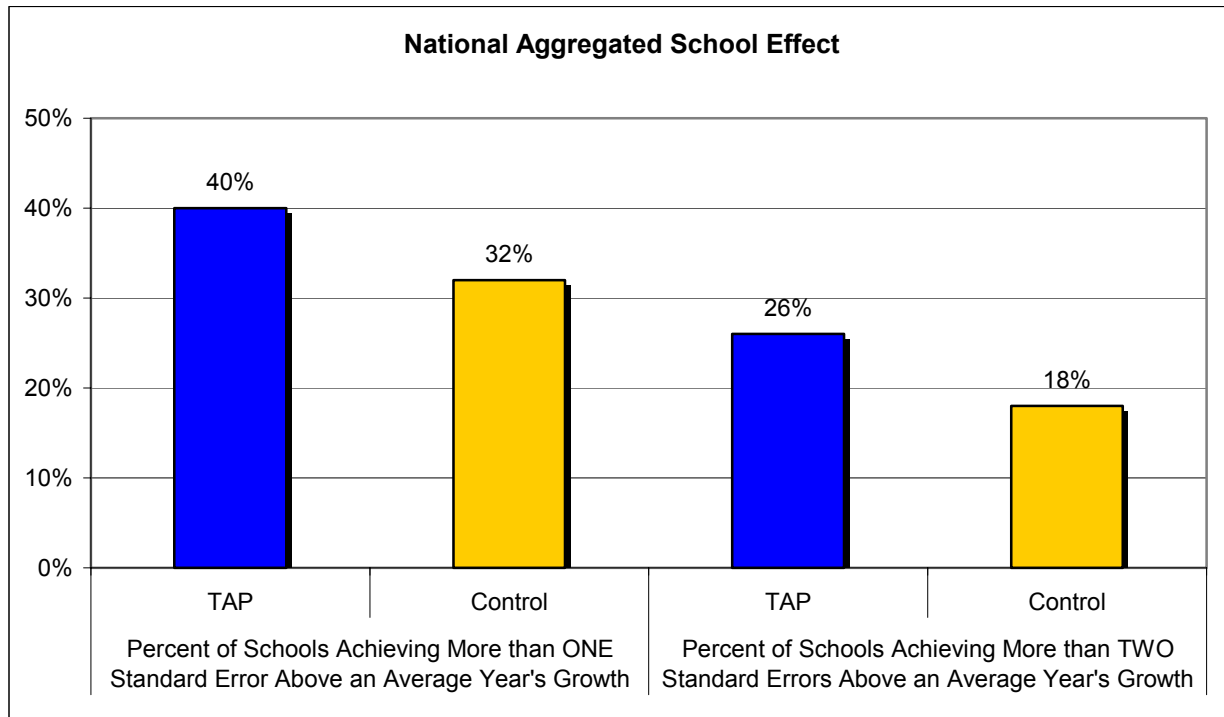
Similar to the teacher-to-teacher comparison, more TAP schools outperformed similar non-TAP schools in producing an average year’s growth or more in both reading and math achievement.

An aggregate analysis of 61 TAP schools in six states compared to 285 control schools in the same states is illustrated below.

**TAP Schools vs Control Schools
Cumulative Distribution Comparative Plot
Standardized School Effectiveness Estimates**



From the above plot, we can conclude that when applying one standard error to their estimates, 40% of TAP schools had estimates that indicated their average student progress was above the average gain, whereas 32% of control schools using the same criteria had that result. When we look at the even higher standard of applying two standard errors to their estimates, 26% of TAP schools and 18% of controls had estimates that indicated their average student progress was above the average gain. This is illustrated in the following chart.



To put our results in context, the RAND study of Comprehensive School Reform (CSR) schools concluded that 50% of CSR schools out-performed their controls in math; and 47% outperformed their controls in reading, although CSR had been operating for a substantially longer period of time than TAP.¹ It is important to remember that even though TAP has been operating in schools since 2000, the majority of schools have joined in the last two to three years. Generally, scholars who study comprehensive school reform contend that one should not expect student achievement results to materialize for at least three years and, in many cases, five years.²

Next, we analyzed adequate yearly progress (AYP) results for the 2004-2005 and 2005-2006 school years in TAP schools as compared to statewide averages. In most cases an equal or higher percentage of TAP schools in the six states make AYP than all schools in their states, despite TAP schools having more students receiving free or reduced-price lunch. When this was not true, TAP schools usually had more high-need students. **We are pleased with this success TAP schools have had, particularly when poverty is taken into account.**

For the 2005-06 school year, Stewart Street Elementary in Gadsden County, Florida ranked #15 of the top 100 elementary schools in the state, gaining an outstanding 88 points from the previous year. Similar elementary schools in Gadsden County gained/decreased from 44 points to -15 points. Stewart Street Elementary's school grade increased from an "F" to a "C" on Governor Bush's A+ plan.

¹ Berends, M., Bodilly, S. and Kirby, S.N., "Looking Back Over a Decade of Whole-School Reform: The Experience of New American Schools," *Phi Delta Kappan*, October (2002): 170.

² Berends, M., Chun, J., Schuyler, G., Stockly, S., and Briggs, R. J., "Challenges of conflicting school reforms: Effects of New American Schools in a high-poverty district," (Santa Monica, CA: RAND Corporation, 2002). /

<http://www.rand.org/publications/MR/MR1483/>

Fullan, M. *Leading in a culture of change*, (San Francisco, CA Jossey-Bass, 2001).

Gray Middle School in Lake County, Florida ranked #18 of the top 75 middle schools in the state, gaining an impressive 71 points. Similar middle schools in Lake County gained from 57 points to 4 points. Gray Middle School rose from a “C” to an “A” on the state’s A+ plan.

Finally, in examining TAP teacher attitudes we have found that overall TAP teachers support the four elements of TAP, and that their support grows the longer they are in the program. We also examine other national teacher surveys and compare attitudes about teaching among those respondents and TAP teachers.

One of the major attitudinal themes of TAP is that the program provides teachers with high-quality professional development and strong teacher collaboration and support. TAP teachers also found their professional development to be more useful in improving their effectiveness in the classroom than teachers nationwide. The most striking difference between TAP professional development and that of other programs is the amount of support and collaboration teachers experience.

The other major theme from the survey results is that, contrary to popular belief, performance pay has neither led to competition nor susceptibility to principal bias in TAP schools. Clearly, as TAP shows, collaboration can remain strong despite the implementation of performance pay, and principal bias need not distort performance pay decisions. This is in sharp contrast to teachers who have not experienced TAP.

Overall, we find that TAP teachers compared to non-TAP teachers experience higher quality professional development as well as more opportunities for collaboration and collegiality, and ways to improve their effectiveness in the classroom.

Our summary conclusion from the large and varied amount of data analyzed is that TAP has been very successful in its first five years. It has improved teaching with the result of better student achievement, and teachers, for the most part like the program. This explains its growth.

TAP in Florida, Louisiana, Ohio, South Carolina and Texas has been funded in part through a \$1,984,000 grant to the National Institute for Excellence in Teaching from the U.S. Department of Education’s Fund for the Improvement of Education (FIE).

Introduction

One of the reasons that ineffective reforms in K-12 education have been implemented and keep getting funded is the absence of appropriate evaluation of programs that have been attempted. Evaluations in education often have considered whether or not those involved (teachers, administrators, parents, students) *like* a program. Furthermore, determining whether a reform is effective usually is attempted in an incomplete way³. For these reasons the Institute for Education Sciences of the U. S. Department of Education advocates a “medical model,” wherein schools or students should be randomly assigned to treatment or control groups in order to get a “scientific” answer the extent to which various reforms are effective.

The gold standard in epidemiology is the randomized, controlled clinical trial. In these studies, large numbers of people are assigned by chance to separate groups to compare different treatments of health conditions objectively. The participants must agree to be part of the study but cannot choose the group to which they are assigned. Neither the participants nor those conducting the study know which treatment they are receiving until it has been completed and the data has been analyzed. As Dr. Kenneth Shine, the former president of the Institute of Medicine, has stated, even in medicine such trials “are difficult to execute and expensive to complete...but they are essential because bad studies only lead to bad information.”⁴

The purpose of this paper is to analyze the impacts of the Teacher Advancement Program⁵ (TAP). TAP is a comprehensive program that includes multiple elements. Many of these elements have been tried in isolation in the past and have not resulted in student achievement gains. Our innovation changes schools’ organizational structure and includes key elements to attract, develop, motivate and retain effective teachers with the ultimate goal of increasing student achievement and closing achievement gaps. TAP’s four elements are: **(1) Providing multiple career paths** which enable teachers to advance while staying in classroom, and also providing opportunities for shared instructional leadership—principals cannot do it alone; **(2) Introducing ongoing applied professional growth** to help all teachers improve instruction by working on their specific needs, as determined by analyzing their classroom performance evaluations and their students’ data. We believe even good teachers can become great, and great teachers can become even more effective; **(3) Increasing instructionally focused accountability**. To be fair there are multiple (at least four) evaluations for all teachers by trained and certified evaluators (master teachers as well as mentor teachers and the principal) based on clearly defined and scientifically validated teaching rubrics. This type of accountability can identify effective teachers and can also determine who needs to improve; **(4)**

³ When students in smaller classes get higher test scores, some claim small class size *causes* higher student achievement. Some studies ignore the fact that small classes are more likely in affluent districts where kids are more likely to have home and family advantages that are complementary to achievement. Some studies also ignore that many of the best teachers want to be in affluent schools with small classes, and those teachers would get high student achievement regardless of class size. Recognizing that correlation does not necessarily mean causation, some evaluators attempt to control for the effects of other variables statistically. The most common of such efforts are regressions that look at the partial correlation of the treatment variable (e.g. class size) holding other factors (family income, teacher quality) constant. The debate then turns to whether or not the proper factors are being controlled for. Is there an unmeasured factor affecting the relationship of concern, which if included would render the correlation between class size and achievement insignificant?

⁴ Markel, H. "Ideas & Trends: Mixed Medical Messages; So What’s a Responsible Sun Worshiper to Do?" *New York Times Magazine*, August 25, 2002.

⁵ For more information on TAP please see Appendix D.

Providing **performance-based compensation rewards** to teachers for hard work if they are successful, for taking on additional responsibilities, for their performance as determined by multiple evaluations, and for the performance of their students as determined by pre- and post-test outcomes.

The challenge in analyzing the impact of TAP is that it is difficult to randomly assign schools to it. Indeed, one of the lessons we have learned through our studies of performance pay is that teacher buy-in is essential to any program's success. Moreover, Dr. Shine acknowledges that even the best-designed clinical trial can produce misleading data—"science is imperfect and dependent on an accumulation of information. . . rather than relying on a single study, we need to draw from the ever-increasing body of knowledge." Hence our evaluation is multifaceted, looking at student achievement gains, adequate yearly progress (AYP) and teacher attitudes.

It is generally accepted that the next best alternative to random assignment is to compare a treatment group to a similar "control group." Here the question becomes, if we observe changes in student achievement in schools adopting TAP, can we attribute these changes to TAP? By comparing changes in student achievement in TAP schools to control schools that are identical other than the fact that they are not implementing TAP, this question is, in theory, answerable. The next question is, how similar to the TAP schools are the control schools in any comparison? It is rarely the case when we can compare two identical schools. Even if we find schools with similar socioeconomic and achievement profiles, they might differ from TAP schools in more subtle ways such as the implementation of new programs or curricula, the stability of the faculty or the quality of the school's leadership. However, if the control group is large enough, such differences among control schools might balance out so that the average characteristics of the control schools might resemble the characteristics of the TAP schools. Interpreting results must take into account both the similarities and differences among TAP schools and their control schools.

This paper analyzes three aspects of the impact of TAP. First we analyze student achievement gains using 2004-2005 student data at two levels of comparison—teacher-to-teacher (TAP teachers compared to teachers in non-TAP schools) and school-to-school (TAP schools versus non-TAP schools). Dr. June Rivers of SAS Institute Inc., utilizes SAS[®] EVAAS^{®6} which is a system that uses student test score data from TAP schools and control schools to calculate individual teachers' value-added gains in order to determine performance bonuses for TAP teachers, and the school-wide gains for school-wide bonuses. A by-product of these calculations is the ability to compare student achievement growth from TAP teachers and schools to such growth from control teachers and schools.

The results are encouraging. When comparing TAP teachers to controls, in every state fewer TAP teachers scored below the average amount of student achievement growth than controls, and more TAP teachers scored at or above the average amount of student achievement growth than controls. When comparing TAP schools to controls, based on the percentage of schools scoring below the

⁶ SAS[®] EVAAS[®] for K-12 builds on the Tennessee Value-Added Assessment System (TVAAS) methodology developed by Dr. William L. Sanders and his colleagues at the University of Tennessee. SAS[®] EVAAS[®] provides the most precise and reliable way to measure schooling influence. SAS[®] EVAAS[®] supplies a precise measurement of student progress over time, as well as a reliable diagnosis of opportunities for growth. Over the course of a school year, SAS[®] EVAAS[®] reporting shows if all students are making reasonable progress. The diagnostic reports tell teachers if students at varying achievement levels have experienced growth. Based on the reporting provided, teachers can understand two things: the incoming levels of academic preparedness of their students and the growth patterns within their classrooms. (<http://www.sas.com/govedu/edu/services/effectiveness.html#overview>)

average as well as at or above the average, TAP schools outperformed controls in 67% of the categories in math and in 100% of the categories in reading.

An aggregate analysis of all TAP teachers compared to control teachers shows that when applying one standard error to their estimates, 38% of TAP teachers as compared to 26% of control teachers had estimates that indicated their average student progress was *above* the average gain. Also, when applying two standard errors to the teachers' estimates, 25% of TAP teachers as compared to 14% of control teachers had estimates that indicated their average student progress was *above* the average gain. A similar comparison of all TAP *schools* compared to control schools shows that when applying one standard error to their estimates, 40% of TAP schools had estimates that indicated their average student progress was above the average gain, whereas 32% of control schools using the same criteria had that result. When we look at the even higher standard of applying two standard errors to their estimates, 26% of TAP schools and 18% of controls had estimates that indicated their average student progress was above the average gain.

To put our results in context, the RAND study of Comprehensive School Reform (CSR) schools concluded that 50% of CSR schools out-performed their controls in math; and 47% outperformed their controls in reading, although CSR had been operating for a substantially longer period of time than TAP.⁷ It is important to remember that even though TAP has been operating in schools since 2000, the majority of schools have joined in the last two to three years. Generally, scholars who study comprehensive school reform contend that one should not expect student achievement results to materialize for at least three years and, in many cases, five years.⁸

Second, we analyze adequate yearly progress in TAP schools as compared to statewide averages. In many cases TAP schools are equally or more successful in making AYP than other schools in their respective states.

Third, we examine TAP teacher attitudes and have found that overall TAP teachers support the four elements of TAP, and that their support grows the longer they are in the program. We also examine other national teacher surveys and compare attitudes about teaching among those respondents and TAP teachers. Overall, we find that TAP teachers experience higher quality professional development as well as more opportunities for collaboration, collegiality, and ways to improve their effectiveness in the classroom.

Student Achievement Gains

Teacher-to-teacher comparisons. In order to determine individual value-added bonuses for eligible TAP teachers, SAS[®] EVAAS[®] first estimates reference population growth averages for each grade and subject using the student test scores (from the same state test) of both TAP teachers and similarly matched non-TAP teachers. The reference population growth average is the average amount

⁷ Berends, M., Bodilly, S. and Kirby, S.N., "Looking Back Over a Decade of Whole-School Reform: The Experience of New American Schools," *Phi Delta Kappan*, October (2002): 170.

⁸ Berends, M., Chun, J., Schuyler, G., Stockly, S., and Briggs, R. J., "Challenges of conflicting school reforms: Effects of New American Schools in a high-poverty district," (Santa Monica, CA: RAND Corporation, 2002). /

<http://www.rand.org/publications/MR/MR1483/>

Fullan, M. *Leading in a culture of change*, (San Francisco, CA Jossey-Bass, 2001).

of growth a teacher is expected to make with his or her students in a particular grade and subject in a year.

SAS[®] EVAAS[®] then identifies which TAP teachers influenced student progress at a rate that is statistically different from the average teacher in a particular grade and subject. All TAP teachers are classified into five categories, as having their effect estimates: 1) more than two standard error units below the average teacher's estimate, 2) between one and two standard error units below, 3) within one standard error unit above or below the average teacher's estimate, 4) between one and two standard error units above the average, or 5) more than two standard error units above in terms of their students' gains. Teachers with gain estimates at or above the average teacher's estimates (i.e., in groups 3, 4, or 5) are considered to have made one year's growth or more with their students and so, in the TAP model, qualify for the part of the bonus that is based on their own students' gains.⁹

A by-product of this analysis is the ability to place the set of control teachers into the same five categories, which then allows us to make direct comparisons between the effectiveness of TAP teachers and their controls. If TAP teachers fall into the two lowest categories less frequently and fall into the three highest categories more frequently, then we can conclude that TAP has a positive effect on teachers' influence on student achievement.

In evaluating TAP teachers (and similarly below in evaluating TAP schools), we calculate the effect of each teacher on student progress as assessed by the difference between the actual average scores of the teacher's students and the expected average scores of those students (as derived from previous scores). This is based on previous test scores in all subjects and is linked at the student level and analyzed simultaneously to provide the precise estimate of schooling's influence on student progress. By dividing the individual teacher effect by the associated standard error (the measure of uncertainty around the estimate) we can determine how many standard error units a particular teacher's effect is from the growth average, and then can place each teacher in one of the five categories.¹⁰

Each teacher has her own effect estimate derived from the difference between the averages of her students' actual and expected scores, and there is a standard error associated with her effect based on the quantity and quality of her students' data. The numbers represent composites that include both

⁹ A standard error indicates how variable the sample statistic (e.g., the class's average score) is from sample to sample. The standard error of a statistic depends on the sample size; the larger the sample size the smaller the standard error. (Schuyler, H.W., William, C.H., *Reading Statistics and Research, Second Edition* (New York, NY: Harper Collins, 1996). The simplest, most non-technical way to think of the standard error of measurement is the following: If a single student (or in the case of a teacher, her whole class of students) were to take the same test repeatedly (with no new learning taking place between tests and no memory of question effects), the standard error is the measure of the "spread" of the repeated average test scores of that student (or class). It tells us how accurate or representative one score by a student (or class) is. Does one score predict the next? ("Standard Error of Measurement,"

<http://www.tea.state.tx.us/student.assessment/taks/standards/sem.pdf>)

The difference between the "standard deviation of scores on a test" and the "standard error of measurement on a test" is that when one refers to the standard deviation of scores on a test, usually this is referring to the standard deviation of the test scores obtained by a *group* of students on a single test. It is a measure of the "spread" of scores *among* students whereas the standard error of measurement on a test is a measure of the "spread" of scores *within* a single student (from one testing to another).

¹⁰ For example, if the teacher effect is 7.9 points and the standard error is 5.7 points, then the number of standard error units would be 1.38 (7.9/5.7). 1.38 standard error units is between one and two standard errors above the average gain, therefore this particular teacher would receive a score of 4.

teachers teaching math and reading within the same grade as well as individual teachers responsible for instruction in either math or reading. For example, if a teacher has students in the third and fourth grades all taking separate reading and math tests, two different entries are counted (one for her students in each grade with reading and math combined). To calculate individual teacher performance bonuses, each teacher gets one combined report; that is, if a teacher had multiple reports,¹¹ the scores from each would be combined by taking a weighted average (based on the number of students that took each subject test) in order to calculate an overall score.¹²

Standard error units help determine how statistically significant these differences in a particular teacher's average student scores are from the growth average determined by the control teachers. They indicate what proportion of the teachers (TAP and otherwise) do *statistically* significantly better than average and which do *statistically* significantly worse. In other words, the five categories mentioned above indicate whether or not the growth a teacher makes with her students is different from the average amount of growth and with how much confidence we can say so. For example, for a teacher who scores a "5," we can say with 95% confidence that she made more than the average year's growth in her students' achievement. Similarly, for a teacher who scores a "1," we can say with 95% confidence that she made less than the average year's growth in her students' achievement.¹³

Unfortunately, standard errors do not indicate *how much* more or less growth a teacher has made with her students when compared to the average; rather, it tells us how statistically significant a teacher's difference in growth is from the average, whatever the size of that difference may be. However, the statistical significance is a function of both the effect size and the standard error, which in turn depends upon the quantity and quality of the data.

Shrinkage estimation helps to compare the relative magnitude of effect teachers have on their students' achievement, and is another similar way of describing the process of adjusting teacher effects for the accuracy and reliability of their data. In order to qualify for an individual bonus, teachers are required to have scores for at least ten students who have been in their classroom for the full school year. The more student test scores a teacher has and the better the student data (i.e., the smaller the standard error), the more precise the measurement of his or her effectiveness. Thus, teacher effect estimates for teachers with small classes and thus less data to use in calculating their impact or with less accurate data for other reasons, are protected by giving less weight to their individual scores and greater weight to the average scores. Those with more student test scores, or with more reliable individual classroom data for whatever reasons, are weighted more toward their individual scores and less toward the average. Such modifications are called "shrinkage estimates." According to Richard L. Tate of Florida State University, shrinkage estimation, "combine(s) the observed student achievement mean for (a) teacher with the overall average of the student means for all of the teachers...as a possible solution to the problem of very unstable estimates of the effects for, say, teachers with very small

¹¹ A report refers to a single score for each subject per grade.

¹² An overall score—based on the value-added gains individual teachers make with her students—is necessary because that score determines part of the teacher's bonuses.

¹³ For a teacher who scores a "2" or "4," we can say with 68% confidence that she made less or more than the average year's growth in her students' achievement, respectively. For a teacher who scores a "3," the average growth she made in her students' learning is not detectably different from the reference average population growth.

classes.”¹⁴ While it protects teachers with small classes, shrinkage also allows teachers with more students and more complete data to be more correctly and more confidently identified.

Each teacher’s final estimate is a combination of a weighted individual teacher point estimate and a weighted average for all teachers. So if a teacher has only ten students for whom she has reliable test data (resulting in a larger standard error), her final effect estimate will give more weight to the teacher average than to her individual point estimate, moving her final estimate more towards the average, even though her individual point estimate may be much higher or lower than the average. If a teacher has 30 students with reliable test data, her effect estimate will give more weight to her individual point estimate, pulling her estimate away from the average and more towards her individual score. If a teacher with a larger and more reliable set of data (which is reflected in a smaller standard error) has an individual point estimate that is detectably different from the average teacher, her overall estimate will reflect this difference.¹⁵ Thus it is possible for two teachers to have the same individual estimates in a grade and subject, yet have different overall scores (i.e., be placed in different categories 1-5) because of differences in the size of their accompanying standard errors.¹⁶

Because all teachers are considered to be equal in their effectiveness (i.e., at the average level) until the quality and quantity of the data for their classrooms (historical and current year) pull their estimates away from the average, teachers categorized in the tails of the distribution (i.e., being more than two standard errors either below or above the average) have the most accurate estimates (small standard errors) and relatively large effect sizes compared to their standard errors. In other words, all teachers who have scores that are above or below the average have them because they had relatively large differences between the scores of their students and the mean score, *and* enough *reliable* data (small standard errors) to pull them away from the average. Otherwise, teachers are determined to be closer to the average because of less reliable data, or small differences from the mean, or both. Shrinkage estimates are conservative estimates of teacher effectiveness because it assumes that each teacher performs at the average level unless a teacher has enough data that tell otherwise and can pull his or her estimate away from the average.¹⁷

In their analysis, SAS[®] EVAAS[®] does not literally match controls to TAP teachers. Rather, they estimate each teacher’s effects across a sufficiently large distribution of teachers (i.e., at least 20 teachers including both TAP and non-TAP teachers)¹⁸ in order to get a more precise estimate of how a teacher performs relative to a comparable set of teachers. Control teachers are chosen from within the same state as the TAP teachers (with the exception of Colorado)¹⁹ because the comparisons for value-added are based on student growth on the same state test.

¹⁴ Tate, R. L. “A Cautionary Note on Shrinkage Estimates of School and Teacher Effects.” *Florida Journal of Education Research*, 42 (2004): 1-21.

¹⁵ Another way of looking at this is a teacher’s point estimate is reduced when divided by the standard error of that estimate. That is, if a teacher data yields a large standard error, the point estimate will be reduced and drawn closer to the mean.

¹⁶ Source from email from June Rivers to Lewis Solmon, November 7, 2006

¹⁷ In TAP, performance bonuses are paid for teachers and schools in categories three, four and five. Thus, unless teachers or students are more than one standard error unit below average growth they will receive that part of the bonus. We err on the side of providing too many payouts.

¹⁸ TAP teachers are included in the control group because both groups contribute to the reference population average.

¹⁹ We do not have information on the control group for Colorado’s TAP teachers since SAS[®] EVAAS[®] has access only to the results from Northwest Evaluation Association (NWEA)—a nationally normed test—and not the Colorado Student Assessment Program (CSAP).

Based on their long history of delivering teacher reports, SAS[®] EVAAS[®] has determined that analyses that include control groups of at least 20 teachers per grade and subject is sufficient upon which to draw inferences. Using at least 20 teachers allows for the assumption that each control group is representative of all teachers within a grade and subject area, and for the expectation that the subgroup of TAP teachers, had they not been in a TAP school, resembles the control group. Any divergence from the average teacher effect would thus be attributed to TAP. SAS[®] EVAAS[®] analyses identify the part of student progress attributable to schooling influences across a broad population of teachers or schools. Some could argue that a district effect is confounded in the teacher effect estimates, but based on previous statewide analyses, SAS[®] EVAAS[®] has found the district effect to be relatively small compared to the teaching effects within schools within districts.

Some could also argue that TAP's effect on student achievement is overestimated because teachers who are already highly effective are more likely to choose to participate in TAP.²⁰ In other words, there may be ways in which TAP teachers are different from non-TAP teachers that affect student achievement and are not controlled for in the analysis. If so, this could bias the results to make TAP look more effective than it actually is. However, by following both TAP and non-TAP teachers over time, with more observations it will be possible to ascertain more precisely how much the participation in TAP influences the ultimate effectiveness of these teachers.

As explained above, teacher effectiveness is measured by the value-added growth of each teacher's students which is determined by the difference in the averages of her student's actual and expected scores. SAS[®] EVAAS[®] requires at least three years of test scores for each student in order to form a more precise estimate of each student's expected achievement. This requirement also allows each student to serve as his or her own control, because other covariates to achievement, such as socio-economic status, previous achievement and other personal characteristics, remain the same from year to year for each student.

Table 1²¹ (see Appendix A) compares the number of math and reading teachers in TAP schools to the number of control teachers who achieved gains of various levels of statistical significance with their students. Each teacher is counted separately for each grade taught; if multiple subjects are taught in a grade, these are combined. As mentioned before, all teachers—TAP and control—are classified into five categories according to how statistically significantly different from the reference population growth average a teacher's measure of effectiveness is. These results provide the frequencies of TAP and control teachers in each category. Since SAS[®] EVAAS[®] compared TAP teachers to a control group of similar TAP and non-TAP teachers within the same state, and since the set of controls are redefined annually, we would expect TAP teachers to have results similar to control teachers. Any divergence could thus be attributed to TAP.

Then, under each of the five categories, we noted which of the two groups, TAP or control, outperformed the other. *In categories "1 and 2" the "outperforming" group is the one with the smaller*

²⁰ Although there is no evidence that TAP is operating disproportionately more in schools that start with highly effective teachers (indeed the opposite can be a motivator for schools to implement TAP), it is likely that highly effective teachers are attracted to TAP schools, which is one of the goals of TAP.

²¹ All data in Tables 1 and 2 are from the 2004-05 school year and are collected from TAP schools and control schools and analyzed by SAS[®] EVAAS[®].

of the two percentages, meaning that fewer teachers received a score that is one or more standard error units below the average teacher’s estimate. In categories “3, 4, and 5” we noted which group had the higher of the two percentages, meaning that more teachers had estimates that indicated that their average student growth was at or above the average gain. This is documented in the summary table below.

Summary of Table 1: Type of teacher that outperformed in each category								
State**	# of Categories TAP Outperformed Controls	1	2	3	4	5	Total % in 1& 2	Total % in 3, 4 & 5
AR	4 out of 5*	TAP	TAP	TAP	TAP	Control	TAP	TAP
IN	3 out of 5	TAP	Control	TAP	Control	TAP	TAP	TAP
MN	3 out of 5	TAP	TAP	Control	Control	TAP	TAP	TAP
SC	3 out of 5	TAP	TAP	Control	Control	TAP	TAP	TAP
FL	4 out of 5	TAP	TAP	TAP	TAP	Control	TAP	TAP
LA	2 out of 5	TAP	Control	TAP	Control	Control	TAP	TAP
% TAP	63%						100%	100%

* This figure tells us that of the five possible comparisons of TAP teachers to controls, one at each achievement category (1-5), Arkansas TAP teachers outperformed their controls in 4 of the 5 categories. This same interpretation can be applied to each state in this table.

The last two columns in this table tell us that in every state, fewer TAP teachers than controls scored below the average amount of student achievement growth, and more TAP teachers than controls scored at or above the average amount of student achievement growth.

** This table excludes information about Colorado TAP teachers.

Results in this summary table show that in almost two thirds (63%) of the comparisons of whether TAP teachers outperformed control teachers in each growth level (1-5), TAP teachers came out on top across six states. The last two columns tell us that when combining the percentages in categories “1 and 2,” and combining the percentages in categories “3, 4, and 5,” in every state, fewer TAP teachers than controls scored below the average amount of student achievement growth, and more TAP teachers than controls scored at or above the average amount of student achievement growth in every comparison. In other words, in 100% of these comparisons, TAP teachers came out on top.

To illustrate this in greater detail, Table 1 (in Appendix A) shows that in Arkansas, 95% of TAP teachers had estimates that indicated that their average student growth was at or above the average gain (score of “3, 4 and 5”), as compared to 75% of non-TAP teachers. In the corresponding summary table above, under Arkansas, “TAP” is noted in the “3, 4 and 5” column, indicating that more TAP teachers scored “3, 4 and 5” in comparison to similar, non-TAP teachers.

As has been mentioned earlier, the methodology used for these ratings is more reliable at distinguishing between the significantly worse (score of “1”) and significantly better (score of “5”) performing teachers, and less so at distinguishing the performance of teachers in the middle.²² Using this criterion, all states have a smaller percentage of TAP teachers scoring a “1” than controls, which means that fewer TAP teachers were statistically significantly less effective in raising their students’

²² To be categorized as a “1” or “5” a teacher must have a large effect size (scores relatively far from the mean) as well as have a small standard error compared to the effect size.

scores than control teachers.²³ To clarify, fewer TAP teachers had estimates that indicate that their average student growth was below the average gain. At the other end, half of the states (Indiana, Minnesota and South Carolina) had a higher percentage of TAP teachers scoring a “5” than controls, which means that these three states had a higher percentage of TAP teachers who were statistically significantly more effective at raising their students’ scores than control teachers. Additionally, we see that in all states a higher percentage of TAP teachers scored a “3 or above” than their controls.²⁴ These comparisons are very positive, clearly demonstrating that, on average, TAP teachers produce higher student achievement growth than similar teachers not in TAP schools.

In the case of Colorado, because Eagle County uses the Northwest Evaluation Assessment (NWEA), a nationally normed test, the applicable one year’s growth for students is determined by looking at the growth of similar students and their teachers across the country, which accounts for more teachers than a state test alone. Then each teacher is placed in one of the five categories. In this case, the comparison can be with the one year growth level. Without TAP, most teachers should fall into category three—with a lesser percentage falling below or above. In other words, we would expect a normal distribution around the mean (category 3). However with TAP, 7.5% of TAP teachers fell into categories 1 and 2 and 46.7% fell into category 5 alone. Thus, we conclude that TAP in Eagle County was highly successful.

To determine how large an overall effect, or how much more or less value-added, TAP teachers achieved with their student compared to control teachers, SAS EVAAS took all the data for teachers (and later for schools) across schools, districts and states and attempted to standardize the effects for all the different state tests utilized. They then produced a Comparative Cumulative Distribution Plot that displays the distribution of the composited measure of teacher effectiveness for TAP teachers and for non-TAP teachers. The composite teacher effect, as calculated by SAS EVAAS, can be considered to be a standardized measure of teacher effectiveness in terms of student outcomes across the subjects of math and reading taught by a teacher.

Figure 1 (see Appendix A) displays the national aggregated standardized teacher effectiveness estimates, and is based upon 610 TAP teachers and 2,337 control teachers. One way to look at the data is to consider the proportion of TAP and control teachers who achieve a certain level of effectiveness or less. By drawing a vertical line from any point on the horizontal axis (which indicates the level of teacher effectiveness) to either of the cumulative distribution lines, we can see what percentage of TAP or control teachers achieved that level of effectiveness *or less*. The one standard error point on the horizontal axis indicates that when applying one standard error to the teachers’ estimates, 62% of TAP teachers and 74% of control teachers had estimates that indicated their average student progress was *at or below* the average gain. It is then easy to calculate that 38% of TAP teachers as compared to 26% of control teachers had estimates that indicated their average student progress was *above* the average gain. Using the same method, the two standard errors point on the horizontal axis indicates that when applying two standard errors to the teachers’ estimates, 25% of TAP teachers as compared to 14% of control teachers had estimates that indicated their average student progress was *above* the average gain.

²³ In Arkansas, no TAP teachers scored a 1, while 8.6% of control teachers did. In Indiana, there were 6.8% of TAP teachers versus 7.9% of controls; in Minnesota, 7.8% versus 14.3%; in South Carolina, 27.9% versus 34.0%; in Florida, 3.3% versus 9.1%; and in Louisiana, there were 1.6% of TAP teachers who scored a 1 versus 11.5% of control teachers.

²⁴ All TAP schools outperformed their controls. Most notably, 95.4% of Arkansas TAP teachers scored average or above as compared to 74.9% of their controls—a 20.5% percentage point difference.

Comparison of school-wide gains. SAS[®] EVAAS[®] also provides data on school-wide gains of TAP schools as compared to control schools in order to determine the school-wide portion of the teacher’s performance bonus. SAS[®] EVAAS[®] requires a minimum of ten schools with matched testing across years in order to do its analysis.²⁵

The results of the school-wide comparisons can be found in Table 2 (see Appendix B) and a summary table, similar in format to the summary table for the teacher-to-teacher comparisons, is below. Table 2 illustrates the percentage of TAP and control schools in each TAP state scoring in each of the five categories on math, reading, and math and reading combined. This table composites math and reading across grades served at a school, but reports on math and reading both separately and combined. Once again, all schools are classified into five categories, in this case indicating various numbers of standard errors at, above or below the average school’s estimate of average school-wide student gains. Then for each subject and state, we added up the percentage of schools scoring a “1” and “2” to see which group (TAP or control) had a lower percentage in this category (“1 and 2”). In this case, the group with the *lower* percentage is more effective because it produced statistically significant below average gains less frequently. The same calculation was made for the category “3, 4, and 5” combined to see which group had a *higher* percentage, meaning it produced a statistically significant average gain or more in a subject and state more often. The results are noted in the following summary table.

²⁵ When comparing the scores in the teacher-to-teacher analysis to the school-to-school analysis, the discrepancies are due to the fact that the school-wide analysis compares achievement scores of *all* the students who took the tests in a school whereas the teacher-to-teacher analysis uses the test scores of students of *only eligible teachers* (i.e., those that had at least ten students in his or her classroom for the full school year). Given that these estimates are shrinkage estimates, the larger pool of students’ scores used to calculate the school-wide gains are more likely to pull the estimate away from the average as compared to the smaller pool of a classroom of students. As a result, it is possible for a school to have scored a “5” while none of its teachers did and vice versa.

Summary of Table 2: Type of school that outperformed in each category									
State*	Subject	# of Categories TAP Outperformed Controls	1	2	3	4	5	Total % in 1&2	Total % in 3, 4&5
AR	Math	5 out of 5**	TAP	TAP	TAP	TAP	TAP	TAP	TAP
	Reading	4 out of 5**	TAP	TAP	TAP	TAP	Control	TAP	TAP
FL	Math	2 out of 5	TAP	Control	Control	TAP	Control	TAP	TAP
	Reading	3 out of 5	TAP	TAP	TAP	Control	Control	TAP	TAP
IN	Math	2 out of 5	Control	TAP	Control	TAP	Tied	Control	Control
	Reading	3 out of 5	Control	TAP	TAP	Tied	TAP	TAP	TAP
LA	Math	2 out of 5	Control	Control	Control	TAP	TAP	Control	Control
	Reading	4 out of 5	TAP	TAP	TAP	TAP	Control	TAP	TAP
MN	Math	4 out of 5	TAP	TAP	TAP	Control	TAP	TAP	TAP
	Reading	3 out of 5	TAP	Control	Control	TAP	TAP	TAP	TAP
SC	Math	2 out of 5	Control	TAP	TAP	Control	Control	TAP	TAP
	Reading	3 out of 5	TAP	TAP	Control	Control	TAP	TAP	TAP
% TAP	Math	57%						67%	67%
	Reading	67%						100%	100%

* This table excludes information about Colorado TAP schools.

** These figures tell us that of the five possible comparisons of TAP schools to controls at each achievement category (1-5) in Math and Reading, Arkansas TAP schools outperformed their controls in 5 of the 5 categories in Math and in 4 of the 5 categories in Reading. This same interpretation can be applied to each state in this table.

The last two columns in this table tell us that overall, fewer TAP schools than controls scored below the average amount of student achievement growth and more TAP schools than controls scored at or above the average amount of student achievement growth.

There are two ways to summarize the data in Table 2 (in Appendix B).²⁶ The first is to look at each of the five categories (summarized in the column labeled “# of Categories TAP Outperformed Controls” in the table above), where each discipline is examined separately. In this comparison, TAP schools outperformed their controls in 57% (17/30) of the categories in math and in 67% (20/30) of the categories in reading. The second is to look at the last two columns which show the percentage of schools scoring at or above the average (scores 3-5) and the percentage of schools scoring below the average (scores 1-2). In this comparison, TAP schools outperform their controls in 67% (4/6) of the categories in math and in 100% (6/6) of the categories in reading.

To illustrate the effects of TAP in more detail, Table 2 shows that in Minnesota, 83% of TAP schools had estimates that indicated their average student progress was at or above the average gain as compared to 71% of non-TAP schools. In math, 83% of TAP schools in Minnesota had estimates that indicated their average student progress was at or above the average gain as compared to 62% of non-TAP schools. In the summary table these same cells (“1 & 2” and “3, 4 & 5” under Minnesota) are marked with “TAP”, meaning that fewer TAP schools scored below the average school’s estimate, and

²⁶ When subjects along with grade levels are combined (the R&M rows in Table 2, see Appendix B), the confidence intervals for each score differ, therefore slightly modifying the results so R&M cannot be derived from simply adding R and M together. When examining the percentage of schools scoring at or above the average school in the representative group and the percentage of schools scoring below the average school for reading and math combined, TAP schools outperform their controls in 50% of the categories.

more TAP schools scored above the average school's estimate in math and reading, indicating that TAP teachers generally outperform their controls.²⁷

The results of the national aggregated school effect shows that TAP schools are again more effective than control schools. Using the same approach as was used for the aggregate teacher effect (i.e., for Figure 1 in Appendix A), Figure 2 (see Appendix B) shows that when applying one standard error to their estimates, 40% of TAP schools had estimates that indicated their average student progress was above the average gain, whereas 32% of control schools using the same criteria had that result. When we look at the even higher standard of applying two standard errors to their estimates, 26% of TAP schools and 18% of controls had estimates that indicated their average student progress was above the average gain.

Similar to the TAP evaluation, the RAND study of Comprehensive School Reform (CSR) analyzed the performance trends across the set of CSR schools. The study focused on whether schools made gains in test scores relative to the other schools in their jurisdictions. To put our results in context, the RAND evaluation concluded that 50% of CSR schools outperformed their controls in math, and 47% outperformed their controls in reading. Further, it is important to note that even though TAP has been operating in schools since 2000, the majority of schools have joined in the last two to three years, whereas the CSR program emerged in the early 1990s. Generally, scholars who study comprehensive school reform contend that one should not expect student achievement results to materialize for at least three years and, in many cases, five years.²⁸ Thus, not only has TAP resulted in greater test score gains compared to controls than CSR, but it also has produced these gains in a shorter amount of time.

Adequate Yearly Progress

With the passage of the No Child Left Behind Act (NCLB), the U.S. Department of Education now requires that every state make adequate yearly progress (AYP) on the path to ensuring 100% “proficiency” of all students in reading and math by 2014. According to NCLB, by 2014 all students are to be “proficient” or above as defined by their percentile rank on their state’s test, the rigor and

²⁷ The comparisons of schools would benefit from descriptive statistics about control and TAP schools (student demographics, achievement levels, etc.) to make sure that control schools are similar to TAP schools. SAS[®] EVAAS[®] attempts to mitigate the differences by comparing TAP schools to at least ten control schools for a particular grade and subject—with a larger the number of control schools, the more likely it is that similarity between TAP and non-TAP schools exists. In states where the same test scale is used in all grades within a subject, comparisons of each teacher’s estimated effect are made to an exogenous progress rate, such as established state norms of average growth for a student per grade (South Carolina, Colorado, and Texas). In those states with mixed scale testing, comparisons of each teacher’s estimate are made to the average teacher in the distribution of all teachers within a control group of at least ten schools for a particular grade and subject. Examples of these include Louisiana, Arkansas, Indianapolis Archdiocese, Minnesota and Florida. As states move toward consistent testing in all grades and subjects, we will be able to compare each teacher’s estimate to an exogenous standard for growth (reference gain).

²⁸ While these school-level measures allow the comparison of performance in CSR schools with that of the district as a whole, they are subject to important limitations. For example, these aggregated measures may fail to capture changes in the tails of the distribution, or they may miss some significant achievement effects that could be captured if comparable student-level data were available across jurisdictions.

Berends, M., Chun, J., Schuyler, G., Stockly, S., and Briggs, R. J., “Challenges of conflicting school reforms: Effects of New American Schools in a high-poverty district,” (Santa Monica, CA: RAND Corporation, 2002).
<http://www.rand.org/publications/MR/MR1483/>

definition of which vary substantially from state to state. Each state is allowed to determine the progress that has to be made each year. As a result, some states have required equal increments of growth in the percent reaching at least the proficiency level each year and others have delayed raising the percent that must be proficient until later years. Since the AYP standard varies across states and across time for each state, comparisons among schools and across states in terms of their success in meeting AYP are not possible. However, it is reasonable to look at how TAP schools are doing in making AYP compared to all other non-TAP schools in their respective states as a whole, because standards are the same for all schools in a state.

Table 3 (see Appendix C) presents AYP results for the 2004-2005 and 2005-2006 school years.²⁹ Results for the whole state serve as the comparison group, but we must take into account the possibility of there being different types of students in TAP schools compared to the state as a whole. Specifically, TAP schools may have more or fewer high-need students than is typical in the state. We also show the percentage of students who qualify for free or reduced-price lunch in TAP schools and statewide in 2005.³⁰

In 2004-2005 we see that in Arkansas and Florida a higher percentage of TAP schools than all schools in the state made AYP despite teaching more “high-need” students. The same was true in Colorado, but these TAP schools had fewer high-need students. In Louisiana, TAP schools achieved about the same level of AYP as schools statewide although they had more high-need students. In Minnesota and South Carolina, TAP schools were below the state overall in making AYP, but in both cases they had substantially more high-need students. We also include Ohio and Texas in the analysis, even though in 2004-2005 TAP had not yet been implemented in those states. Before implementing TAP, only one of the four soon-to-become TAP schools made AYP in Ohio, but all the TAP schools had higher percentages of students eligible for free or reduced-price lunch than the state as a whole. In contrast, both soon-to-be TAP schools in Texas³¹ had made AYP before they started TAP even though they too had higher percentages of students eligible for free or reduced-price lunch than all schools in Texas.

We would expect the percentage of schools in each state making AYP to decline each year when the percent of each subgroup of students required to meet the proficiency standard goes up. Also, if more schools joining TAP are at the lowest achievement levels, this will add to the disadvantage TAP schools have compared to all schools in the state because they will have further to go to get to the “proficient” level.

In the 2005-06 school year, a higher percentage of TAP schools in Florida and Arkansas continued to make AYP as compared to their statewide average, despite having more students receiving free or reduced-price lunch. In 2005-2006 these were the only two TAP states in this study that raised the bar for AYP. In Arkansas, the percent of all schools making AYP in the state fell from 73% to 30%, a decrease of 43 percentage points, but TAP schools remained much higher at 93% in

²⁹ All data on adequate yearly progress and student demographics is from the National Center for Education Statistics, “Search for Schools, Colleges, and Libraries,” <http://nces.ed.gov/globallocator/>

³⁰ The percent of students who qualify for free or reduced-price lunch in a school is unlikely to change significantly from year to year.

³¹ The third TAP school in Texas opened in 2005, therefore did not report AYP status and was not included in our calculations.

2006, meaning TAP schools making AYP decreased by a mere 7 percentage points.³² The percent of Florida schools making AYP also fell, from 36% to 28%, while the percentage of Florida TAP schools remained much higher with two TAP school failing to make AYP³³ (60% did make AYP compared to 100% the year before).

Louisiana did not change its AYP targets between 2005 and 2006. Louisiana is the only state in which the percentage of TAP schools making AYP increased from 2004-05 to 2005-06. The share of all Louisiana schools making AYP remained the same at about 85%, but the share of TAP schools rose from 83% to 93%. It is important to note that the number of TAP schools in Louisiana increased from six to 32, but the share of students in Louisiana schools on free or reduced-price lunch remained about the same from 73% to 74%. The AYP data for Louisiana's analysis, excludes the new TAP charter schools in the Algiers section of New Orleans as well as two TAP schools in Jefferson Parish, so we would expect that the percentage of free or reduced-price lunch in all Louisiana TAP schools would be even higher than 74%.³⁴ This is in comparison to a state average of 62%.

All other TAP states in this study also did not change their standards for AYP between 2005 and 2006. Colorado was the only state in our comparison that increased the percentage of schools making AYP from 2005 to 2006 *statewide*, while all other states saw decreases in the average percentage of schools making AYP. Colorado's results in TAP schools were equivalent to the previous year's results. In Colorado, 73% of TAP schools made AYP in both years, although the state's share rose from 59% to 75%, bringing the state average to the same percentage as the TAP schools.

Ohio and Texas had a few schools in TAP that started in 2006. Statewide, Ohio saw a big drop in the percent of schools achieving AYP from 2005 to 2006. Nevertheless, one additional TAP school achieved AYP compared to the year before entering TAP (i.e., one of four became two of four making AYP). In Ohio, in 2005, 26% of all students statewide were eligible for free or reduced-price lunch, compared to 71% of students in TAP schools. In Texas only two TAP schools reported and one of them made AYP. In Texas, TAP schools had a higher percentage of high-need students than the state as a whole, with 62% of students in TAP schools eligible for free or reduced-price lunch compared to 47% statewide in 2005.

We should note that in South Carolina, the share of TAP schools making AYP statewide fell from 47% to 37% between 2004-05 and 2005-06. The case of South Carolina stands out, because with only 12% of South Carolina TAP schools making AYP in the first year and 9% doing so in the second, TAP schools in this state seem to be doing particularly badly. South Carolina is well-known for its rigorous academic standards. However, there are other reasons why only one out of nine South Carolina schools made AYP in 2004-05, and it is instructive to provide a detailed explanation. In South Carolina as in every state, every student subgroup (indicators), including minority, special

³² One Arkansas TAP school went from making AYP in 2005 to not making AYP in 2006. In this elementary school, 92% of the students are eligible for free or reduced-price lunch. All Arkansas TAP schools made AYP in 2005.

³³ One TAP school, Gray Middle School in Lake County which will be discussed later, made provisional AYP. A provisional AYP is assigned if a school did not meet AYP, but received a school grade of A or B on Governor Bush's A+ Plan. School grades are based on school scores on Florida's Comprehensive Assessment Test (FCAT), progress of low performing students as measured by the FCAT and the percentage of test takers in each subgroup. In our case, Gray Middle School received an A school grade.

³⁴ Schools in Jefferson Parish were exempted from having to make AYP due to the hurricanes in 2005.

education, poor and ESL students, all must reach the goal for the year in both reading and math if a school is to be deemed as making AYP. So it is notable that four schools were within one or two indicators of making AYP.³⁵

1. West Hartsville Elementary School met 19 out of 21 indicators for AYP and missed its target for disabled students' performance on the language arts and math portions of the test by 10% and 8.1%, respectively. Based on the number of students identified, this equates to seven students moving proficiency levels in English, Language Arts (ELA) and five in math. If a disabled student takes an off-grade level test, the score is automatically reported as below proficient rather than not tested.
2. Brunson Dargan Elementary School also met 19 out of 21 indicators for AYP and missed their target for disabled student's performance on ELA and math by an even narrower margin than West Hartsville. In ELA, Brunson Dargan was 2.1% away which would represent fewer than two students based on their total number of disabled students (85). In math, Brunson Dargan was 0.5% of students prevented them from meeting that indicator (one student or one proficiency level).
3. MS Bailey Elementary School met 14 out of 15 indicators for AYP and missed its target for the scores of students receiving subsidized meals on the ELA portion of the test only. The target was missed by 2% of the 106 students who qualified for free or reduced-price lunch.
4. EB Morse Elementary school met 18 out of 19 objectives and missed its target for the scores of students receiving subsidized meals on the ELA portion of the test only. The school was less than two percentage points away from meeting this target.

Thus, if 15 to 20 students in the previously mentioned TAP schools increased one level, five out of nine of the TAP schools would have met AYP.

Additionally, of the remaining four schools in South Carolina that did not reach AYP and were over three indicators away, their status as middle schools is not coincidental. All of the schools that either met AYP or were one to two indicators away received substantial Title I funds to implement improvements, such as reducing class size and paying for materials and technology. However, three of the four middle schools did not receive Title I funds because districts only allocate these funds to the elementary schools. Further, the four TAP middle schools that did not meet AYP draw students from very low socioeconomic status communities, and most of the students reach sixth-grade reading far below grade level. In order to meet the needs of these students, most of the TAP middle schools have been focusing on pulling students from the bottom level (below basic) to the 'passing' level (basic). However, only the movement of students into the higher category of 'proficiency' is counted by AYP. On the other hand, the state uses movement into all levels (below basic I, below basic II, basic, proficient, advanced) in giving schools their state grades. Even though these middle schools are making great strides in improving student achievement, which is reflected in the state's report card, these accomplishments are not recognized by the AYP requirement.³⁶

There are other factors that affect why schools in a particular state may or may not make AYP in a specific year—namely, the rigor of the state's proficiency standard, and the proportion of students

³⁵ Source from email from Jason Culbertson to Todd White, July 19, 2006.

³⁶ Source from email from Jason Culbertson to Todd White, July 19, 2006.

in each category required to reach them in a particular year. However, the most important factor in explaining differences among TAP schools and all schools in the state is the difference in the percent of high-need students that the schools serve. The correlation between the percent of schools making AYP and the percent of students in the state receiving free or reduced-price lunch among the six states is very strong (-0.71).³⁷ The negative correlation indicates that as the percentage of high-need students in schools increases, the percent of schools making AYP decreases. Of the six states we are looking at, TAP schools in South Carolina had the largest difference in students eligible for free or reduced-price lunch from the state as a whole—81% compared to 51% statewide. As we saw in South Carolina, there are also specific explanations as to why the TAP schools did not make AYP as well. Nevertheless, on average, in the 2004-05 school year more TAP schools in the six states made AYP, despite having more students receiving free or reduced-price lunch.

In sum, we are pleased with the success TAP schools have had in terms of AYP compared to their states as a whole, particularly when poverty is accounted for.

Specific TAP Schools. There are also some results from specific TAP schools that confirm their progress. In Florida during the 2005-2006 school year, Stewart Street Elementary in Gadsden County ranked #15 of the top 100 elementary schools in the state, gaining an outstanding 88 points from the previous year. Similar elementary schools in Gadsden County gained or decreased from 44 points to -15 points, respectively. Stewart Street Elementary's school grade increased from an "F" to a "C" on Governor Bush's A+ Plan. Gray Middle School in Lake County ranked #18 of the top 75 middle schools in the state, gaining an impressive 71 points. Similar middle schools in Lake County gained from 57 points to 4 points. Gray Middle School rose from a "C" to an "A" on the A+ Plan. The principals of both these schools received special recognition by the State Board of Education due to outstanding performance as an educational leader in raising their school's grade.³⁸

In Columbus, Ohio during the 2005-06 school year, South High and Barrett Middle made AYP after years of falling significantly below the threshold. This accomplishment put South High and Barrett Middle on the state's Continuous Improvement list. Barrett Middle School outperformed two other middle schools with similar demographics in the same district.³⁹ Barrett increased the number of students passing in reading by 12 percent while the other two schools decreased. In math, Barrett increased the number of students passing by 10 percent while one school showed neither improvement nor decline and the other increased by only 5 percent. South High also outperformed two other high schools with similar demographics in the same district.⁴⁰ South High increased the number of students passing in math by 10 percent while one similar school increased by 2 percent and the other decreased by 2 percent. In reading, the school increased the number of students passing by 2 percent while both of the other schools demonstrated a decrease of 12 percent.

These are true success stories, and we could add many more.

³⁷ This is compared to the correlations we calculated across states, between rigor of standards and percent making AYP (-0.08), and between "height of the bar" and percent making AYP (0.31). Neither of these weak relationships makes sense in terms of sign and clearly are confounded by student socioeconomic status.

³⁸ Source from email from Ruthe Hardy to Lew Solmon, August 31, 2006.

³⁹ Source from email from Greg Paulmann to Lew Solmon, September 3, 2006.

⁴⁰ Source from email from Edna Thomas to Lew Solmon, September 22, 2006.

TAP Teacher Attitudes

In addition to looking at various measures of student achievement growth, TAP conducts an annual teacher attitude survey to determine teacher satisfaction levels with the four elements of the program. In the 2005 annual report,⁴¹ we provide a factor analysis of the survey results, which uses several related questions on the four elements to summarize overall teacher attitudes about each element as a whole.⁴² That analysis does not allow the full range of opinions on the components of each factor to be captured. In order to examine the nuances of perspectives and to try to delineate the reasons behind teachers' opinions, here we have separated out each survey question and taken a closer look at the frequencies of the range of responses (which run from one to five, one representing "not at all" and five representing "very much"). We also looked at other national teacher surveys in order to compare the attitudes of TAP teachers to the attitudes of teachers in general on performance pay and professional development.⁴³

Professional Development. One of the major attitudinal themes of TAP is that the program provides teachers with high-quality professional development and strong teacher collaboration and support. According to the Public Agenda's 2003 national survey of public school teachers, the most demanded topic for professional development was new teaching techniques by 42% of respondents.⁴⁴ In TAP schools, 76% of respondents stated that "learning to effectively use new instructional techniques" was somewhat to very frequently the focus of their cluster group sessions. This is in comparison to National Center for Education Statistics' (NCES) results of its most recent national survey of public school teachers regarding professional development,⁴⁵ which states that 72% of those who responded participated in professional development for new methods of teaching during the 1999-2000 school year, but most (59%) of them only received one to eight hours of such training. In contrast, TAP teachers participate in cluster groups for one to 1 ½ hours every week.

Accordingly, TAP teachers also found their professional development to be more useful than did teachers nationwide. Sixty-one percent of TAP teachers who responded agreed or very much agreed that their professional development experiences were very useful for their roles as teachers, as opposed to only half of the respondents in the Public Agenda survey who felt that their professional development made them a better teachers.

⁴¹ Firetag Agam, K., Reifsneider, D. & Wardell, D. (2005). "The Teacher Advancement Program (TAP): National Teacher Attitudes." Unpublished manuscript. (<http://www.talentedteachers.org/publications.taf>)

⁴² The survey results are drawn from 1,784 TAP teachers who responded to questions regarding the 2004-2005 school year.

⁴³ We reference three national surveys: (1) The Milken Educator Award Survey is a joint study conducted by the Milken Family Foundation and the Teacher Advancement Program regarding Milken Educator opinions on a variety of issues related to teacher quality. Milken Educators are considered to represent the highest quality teachers in the nation. (2) Stand by Me: What Teachers Really Think about Unions, Merit Pay and Other Professional Matters". The Public Agenda, 2003. (3) Teacher Preparation and Professional Development. NCES, 2000. We do not provide all of the questions and relevant percentages upon which this discussion is based. However, we are happy to make the data available upon request.

⁴⁴ Farkas, S., Johnson, J. and A. Duffett, (2003) "Stand by Me: What Teachers Really Think about Unions, Merit Pay and Other Professional Matters." The Public Agenda. This survey is based on a national mail survey of 1,345 public school teachers conducted in the spring of 2003. This represents an overall response rate of 27%, with a margin of error of plus or minus three percentage points.

⁴⁵ Parsad, B., Lewis, L., and Farris, E., *Teacher Preparation and Professional Development: 2000*, NCES 2001-088 (Washington, DC: U.S. Department of Education, National Center for Education Statistics, 2001). This survey was conducted in 2000 using the Fast Response Survey System (FRSS). Questionnaires were mailed to a national representative sample of 5,253 public school teachers. The weighted overall response rate was 75%.

The most striking difference between TAP professional development and that of other programs is the amount of support and collaboration teachers experience. Seventy-seven percent of TAP teachers who responded agreed or very much agreed that they feel support from their grade level, the typical group with whom they have their cluster groups. Additionally, 57% of TAP respondents either agreed or very much agreed that they are becoming better teachers because of the support and collaboration at their schools. Of those who responded to the NCES survey, 69% said that they participated in “regularly scheduled collaboration with other teachers, excluding meetings held for administrative purposes.” However only 31% of them participated at least once a week, and only 24% stated that it improved their teaching a lot. Furthermore, only 23% of NCES’ respondents were mentored by another teacher; only 35% of the mentored teachers met with their mentors at least once a week, and only 37% stated that it improved their teaching a lot. All TAP teachers, on the other hand, hold cluster meetings with other teachers, which are led by either a mentor or master teacher, on a weekly basis.

Performance-Based Compensation. The other major theme from the survey results is that, contrary to popular belief, performance pay has not led to a perception of either competition or susceptibility to principal bias in TAP schools. Only thirty-eight percent of TAP respondents agreed or very much agreed that “performance-based compensation programs encourage competition rather than collaboration among teachers.” Similarly, only about one third (33%) of TAP respondents agreed or very much agreed that “the performance pay system [would] reduce the sense of community among teachers at [their] school.” In contrast, 63% of Public Agenda’s respondents (almost double the percentage of TAP teachers) agreed that “instead of cooperation, there would be unhealthy competition and jealousy among teachers” if some form of merit pay for teachers was implemented at their schools. Regarding principal bias, only 18% of TAP respondents agreed or very much agreed that “performance-based compensation [had] more to do with being well liked than teaching well,” and only 25% agreed or very much agreed that certain teachers were given preferential treatment without good reason at their schools. In contrast, 52% of Public Agenda’s respondents agreed that “principals would play favorites and reward teachers who are loyal to them or who don’t rock the boat” if merit pay was implemented at their schools. Clearly, as TAP teacher responses show, collaboration can remain strong despite the implementation of performance pay, and principal bias need not distort performance pay decisions. This is in sharp contrast to teachers who have not experienced TAP.

On the other hand, similar to other public school teachers nationwide most TAP teachers do not support measuring teacher effectiveness by, or basing performance payouts on, student scores on standardized tests. This is despite the fact that 64% of TAP respondents agreed or very much agreed that more effective teachers should be paid more. This percentage is comparable to the percentage of the Milken Educator Award recipients’ survey, in which 59% of respondents agreed or highly agreed that effective teachers should be paid more than ineffective teachers.⁴⁶ However, only 29% of TAP respondents agreed or very much agreed with the statement, “I endorse measuring *and rewarding* teachers based on the student achievement gains individual teachers produce.” In TAP schools, teacher performance is measured using student achievement scores as well as multiple evaluations of teacher effectiveness in the classroom.

⁴⁶ “2005 Milken Educator Survey Results,” November 2005. This survey was mailed to 1,659 Milken Educators in August 2005. 1,106 surveys were returned, making the overall response rate 67%.

What these seemingly disparate responses tell us is that while teachers like the general idea of rewarding effective teachers, they do not agree with basing those rewards *solely* on test scores. In fact, when we took a closer look at the data from the TAP teacher survey, the lack of support for performance pay was positively correlated with the lack of support for using student test scores to measure teacher effectiveness. In other words, we found a statistically significant relationship between how teachers felt about performance pay and standardized tests.⁴⁷ Instead of effectiveness as measured by student test scores, more teachers support effectiveness as measured by a qualitative evaluation of their performance; 57% agreed or very much agreed with the statement, “I endorse the skills, knowledge and responsibilities standards and rubrics my school has set.”⁴⁸ Additionally, 62% of Public Agenda’s respondents somewhat to strongly favored giving financial incentives to teachers who consistently receive outstanding evaluations by their principals, while only 38% somewhat to strongly favored giving financial incentives to teachers whose kids routinely score higher than similar students on standardized tests.⁴⁹ This reinforces the importance to teachers of basing teacher performance compensation on multiple measures of performance.

The reason for this difference in acceptance of performance ratings and test scores is clear. Only six percent of TAP respondents agreed or very much agreed that standardized test scores accurately represent the academic achievement of students. Similarly, only 14% of Public Agenda’s respondents believed that “standardized tests are necessary and valuable—they are a reliable yardstick for measuring student performance,” while 62% believe that “standardized tests are a necessary evil—ultimately, the schools need some kind of standardized assessment.”

It is likely that the lack of support for performance-based pay incentives is not a reflection of the lack of support for TAP, or an unwillingness of teachers to be held accountable, as it is a reflection of teachers’ disapproval and distrust of current standardized tests. Although, according to the Public Agenda’s survey, most teachers believe that standardized tests are a “necessary evil,” teachers have many criticisms about the structure, implementation and consequences of the tests. Additionally, teachers may dislike their lack of control over background and socioeconomic traits or random factors (e.g., a student who becomes ill on a test day or whose family is faced with a sudden trauma) that may influence a student’s performance on a single test, yet says little about the student’s actual ability. They often do not understand how value-added analysis that requires multiple years of test data can account for these factors, which is one very important reason that districts and states should work to ensure that teachers understand new performance compensation systems. Teachers clearly feel they have more control over classroom evaluations of their teaching.

⁴⁷ We found a statistically significant negative correlation ($r=-.282$ at the .01 level) in support between the performance pay factor (as measured by a series of questions that ask for the respondents’ level of support about performance pay) and the statement, “There is no way of measuring the impact of a specific teacher on student learning.” We also found a statistically significant positive correlation ($r=.317$ at the .01 level) in support between the performance pay factor and the statement, “Standardized test scores accurately represent the academic achievements of students.”

⁴⁸ The Skills, Knowledge and Responsibilities Performance Standards and Rubric define the expected skills, knowledge, and responsibilities for each level teacher in the career path. More often called the TAP Instructional Standards and Rubric, these are used by the evaluation team (made up of master and mentor teachers and administrators) to formally evaluate each teacher’s performance. The TAP Instructional Standards and Rubric are also used throughout the year to guide professional development at the school level.

⁴⁹ Please note that this reflects the percentage of teachers who support incentives based on student achievement *levels* (e.g., the number of students who reach proficient, regardless of where they first performed at the beginning of the school year) not student achievement *growth*. TAP also does not support incentives based on student achievement levels.

As mentioned earlier, only 29% of TAP respondents endorsed measuring *and rewarding* teachers based on test score gains. However, 48% of Public Agenda’s respondents and 43% of Milken Educator Award respondents agree that using student achievement gains is a good way of measuring teacher effectiveness. Although some teachers agree that using test score gains is a good way of *measuring* teacher effectiveness, *this says nothing about how they feel about test score gains affecting their pay*. In other words, it may be that they are fine with looking at these data as long as their pay is not impacted by what they show.

Also, the lack of understanding teachers have about the statistical methodology behind how student achievement gains are measured may influence their opinions about the use of such scores. Only half of the TAP respondents understand how the student achievement attributed to the teacher is measured, and only 43% understand how school achievement is measured. As a methodology, value-added is generally believed to be a sound way to measure the effect a teacher has on his or her students while taking account of other factors such as student SES. However, only about half of TAP teachers understand how value-added is calculated, which may influence their confidence in the use of student achievement gains to measure their effectiveness in the classroom. Again, this points to an important need for districts to involve teachers in the development of performance pay systems and keep them continually informed about how the system works

It is also important to note that the annual TAP teacher survey is conducted towards the end of the school year (end of March and through May). Because standardized tests are administered in the spring, it is likely that many teachers are focused on preparing their students for testing and are therefore feeling stressed about their students’ performance. The coincidence of the survey with testing might cause teachers to express more negative feelings about testing in the survey than they otherwise would had the survey been conducted at a different time of the school year. Spring is also the time when the school year is wrapping up, when teachers are often exhausted from the school year and when students grow fidgety with the impending summer break, all of which can negatively influence how teachers respond in the survey. More important, teachers do not receive their bonus payouts until the beginning of the following school year due to delays in test results being sent to the schools by the states. This probably causes some teachers to grow frustrated and this is likely to be reflected in the survey. In other words, not seeing the “fruits” of their labor in a timely manner may negatively influence the TAP teachers’ perspectives about performance-based pay.

Overall, the Teacher Advancement Program has provided participating teachers with high quality professional development, opportunities for collaboration and collegiality, and ways to improve their effectiveness in their classrooms. Our surveys demonstrate strong support for these elements. There is still resistance to basing some portion of performance pay on student test scores, however, by providing multiple measures of teacher effectiveness TAP has found important ways to mitigate this concern. Low support for performance-pay based on test scores presents some challenges for the future implementation of performance pay programs. First, the lack of support for basing payouts on test score gains sheds light on the larger problem of the lack of support for standardized tests. This suggests we address the concerns and criticisms teachers have and break some of the negative myths about standardized tests, acknowledge the teachers’ legitimate concerns and work towards addressing the limitations. More specifically, the lack of teacher understanding about value-added urges us to evaluate and improve how we explain it to teachers so that they fully understand the methodology. Lastly, we should consider administering the survey after the teachers receive their payout to see if

there are differences in their perspectives about performance pay. If there are considerable differences, it would imply that our current results may not necessarily be the most accurate reflection of TAP teachers' perspectives on performance pay.

Conclusion

The purpose of this paper is to analyze the impacts of the Teacher Advancement Program (TAP). Our evaluation is multifaceted, looking at student achievement gains, adequate yearly progress (AYP) and teacher attitudes.

First we analyze student achievement gains at two levels of comparison—teacher-to-teacher and school-to-school. Dr. June Rivers of SAS Institute Inc, utilizes SAS[®] EVAAS[®] which is a system that uses student test score data from TAP schools and control schools to calculate individual teachers' value-added gains in order to determine performance bonuses for TAP teachers, and the school-wide gains for school-wide bonuses. A by-product of these calculations is the ability to compare student achievement growth from TAP teachers and schools to such growth from control teachers and schools.

In evaluating TAP teachers (and similarly below in evaluating TAP schools), we calculate the effect of each teacher on student progress as assessed by the difference between the actual average scores of the teacher's students and the expected average scores of those students (as derived from previous scores). By dividing the individual teacher effect by the associated standard error we can determine how many standard error units a particular teacher's effect is from the average teacher's estimate, and then can place each teacher in one of five categories—below (score of 1 and 2), at (score of 3), or above the average teacher's estimate (score of 4 and 5). Standard error units indicate what proportion of the teachers (TAP and otherwise) do *statistically* significantly better than average and what proportion do *statistically* significantly worse than the growth average determined by the control teachers. In other words, we examine whether or not the actual amount of average growth a teacher makes with her students is different from the expected amount of average growth and with how much confidence we can say so.

All states have a smaller percentage of TAP teachers scoring a “1” than controls, which means that fewer TAP teachers were significantly less effective in raising their students' scores than control teachers. To clarify, fewer TAP teachers had estimates that indicate that their average student growth was below the average gain. At the other end, half of the states (Indiana, Minnesota and South Carolina) had a higher percentage of TAP teachers scoring a “5” than controls, which means that three states had a higher percentage of TAP teachers who were highly effective at raising their students' scores than control teachers. Additionally, we see that in all states a higher percentage of TAP teachers scored a “3 or above” than their controls. These comparisons are very positive, clearly demonstrating that TAP teachers produce higher student achievement growth than similar teachers not in TAP schools.

The results are encouraging. When comparing TAP teachers to controls, in every state fewer TAP teachers demonstrated statistically significant below average amount of student achievement growth than controls, and more TAP teachers demonstrated statistically significant at or above average amount of student achievement growth than controls. TAP *schools* outperformed their controls in 57% of the categories (1-5, by state) in math and in 67% of the categories in reading. In the comparison of

the percentage of schools scoring at or above the average (scores 3-5) and the percentage of schools scoring below the average (scores 1-2), TAP schools outperform their controls in 67% of the categories in math and in 100% of the categories in reading.

An aggregate analysis of all TAP teachers compared to control teachers shows that when applying one standard errors to their estimates, 38% of TAP teachers as compared to 26% of control teachers had estimates that indicated their average student progress was *above* the average gain. Also, when applying two standard errors to the teachers' estimates, 25% of TAP teachers as compared to 14% of control teachers had estimates that indicated their average student progress was *above* the average gain. A similar comparison of all TAP *schools* compared to control schools shows that when applying one standard error to their estimates, 40% of TAP schools had estimates that indicated their average student progress was above the average gain, whereas 32% of control schools using the same criteria had that result. When we look at the even higher standard of applying two standard errors to their estimates, 26% of TAP schools and 18% of controls had estimates that indicated their average student progress was above the average gain.

To put our results in context, the RAND study of Comprehensive School Reform (CSR) schools concluded that 50% of CSR schools outperformed their controls in math; and 47% outperformed their controls in reading, although CSR had been operating for a substantially longer period of time than TAP. It is important to remember that even though TAP has been operating in schools since 2000, the majority of schools have joined in the last two to three years. Generally, scholars who study comprehensive school reform contend that one should not expect student achievement results to materialize for at least three years and, in many cases, five years.

We also analyze adequate yearly progress in TAP schools as compared to statewide averages. In most cases TAP schools are as or more successful in making AYP than other schools in their respective states. Results for the whole state serve as the control group, but we must take into account the possibility of there being different types of students in TAP schools compared to the state as a whole. Specifically, TAP schools may have more or fewer high-need students than is typical in the state. We also show the percentage of students who qualify for free or reduced-price lunch in TAP schools and statewide in 2005. In sum, we are pleased with the success TAP schools have had in terms of AYP compared to their states as a whole, particularly when poverty is accounted for.

There are also some results from specific TAP schools that confirm their progress. In Florida during the 2005-06 school year, Stewart Street Elementary in Gadsden County ranked #15 of the top 100 elementary schools in the state gaining an outstanding 88 points from the previous year. Similar elementary schools in Gadsden County gained/decreased from 44 points to -15 points. Gray Middle School in Lake County ranked #18 of the top 75 middle schools in the state gaining an impressive 71 points. Similar middle schools in Lake County gained from 57 points to 4 points.

In Ohio during the 2005-06 school year in South High and Barrett Middle in Columbus made AYP after no one thought that they would *ever* accomplish this. This accomplishment put South High and Barrett Middle on the state Continuous Improvement list. Barrett Middle School outperformed two other middle schools with similar demographics in the same district. Barrett increased the number of students passing by 12 percent in reading while both of the other schools demonstrated a decrease. In math, Barrett increased the number of students passing by 10 percent while one school showed

neither improvement nor decline and the other increased by only 5 percent. South High also outperformed two other high schools with similar demographics in the same district. South High increased the number of students passing in math by 10 percent while one similar school increased by 2 percent and another decreased by 2 percent. In reading, the school increased the number of students passing by 2 percent while both of the other schools demonstrated a decrease of 12 percent. These are true success stories.

Finally, we examine TAP teacher attitudes and have found that overall TAP teachers support the four elements of TAP, and that their support grows the longer they are in the program. We also examine other national teacher surveys and compare attitudes about teaching among those respondents and TAP teachers. Overall, we find that TAP teachers experience higher quality professional development as well as more opportunities for collaboration and collegiality.

One of the major attitudinal themes of TAP is that the program provides teachers with high-quality professional development and strong teacher collaboration and support. TAP teachers also found their professional development to be more useful in improving their effectiveness in the classroom than teachers nationwide. The most striking difference between TAP professional development and that of other programs is the amount of support and collaboration teachers experience.

The other major theme from the survey results is that, contrary to popular belief, performance pay has neither led to competition nor susceptibility to principal bias in TAP schools. Clearly, as TAP shows, collaboration can remain strong despite the implementation of performance pay, and principal bias need not distort performance pay decisions. This is in sharp contrast to teachers who have not experienced TAP.

Similar to other public school teachers nationwide, most TAP teachers do not support basing performance payouts on student scores on standardized tests. Yet TAP respondents agreed that more effective teachers should be paid more. What these seemingly disparate responses tell us is that while teachers like the general idea of rewarding effective teachers, they do not agree with basing those rewards solely on test scores. The reason for this difference in acceptance of performance ratings and test scores is clear. Only six percent of TAP respondents agreed or very much agreed that standardized test scores accurately represent the academic achievement of students. It is likely that the lack of support for performance-based pay incentives is not so much a reflection of the lack of support for TAP, or unwillingness of teachers to be held accountable, as it is a reflection of teachers' misunderstanding, disapproval and distrust of current standardized tests. They often do not understand how value-added analysis can account for these factors.

Our final conclusion from the large and varied amount of data analyzed is that TAP has been very successful in its first five years. It has improved teaching with the result of better student achievement, and teachers, for the most part like the program. This explains its growth to more than 130 schools. We shall do even better in the future.

Appendix A—TABLE 1: TAP vs. Control Teachers, 2004-05

NOTE 1: This table depicts teachers with multiple reports in addition to all teachers whose students have valid data, including temporary teachers.

NOTE 2: "Number of Teacher Reports" column: A report is for each grade taught. If both math and reading are taught, scores are aggregated by grade.

NOTE 3: Colorado's results below are only a partial analysis, with respect to determining teacher performance bonuses, because only the data provided by NWEA (and not CSAP) are represented.

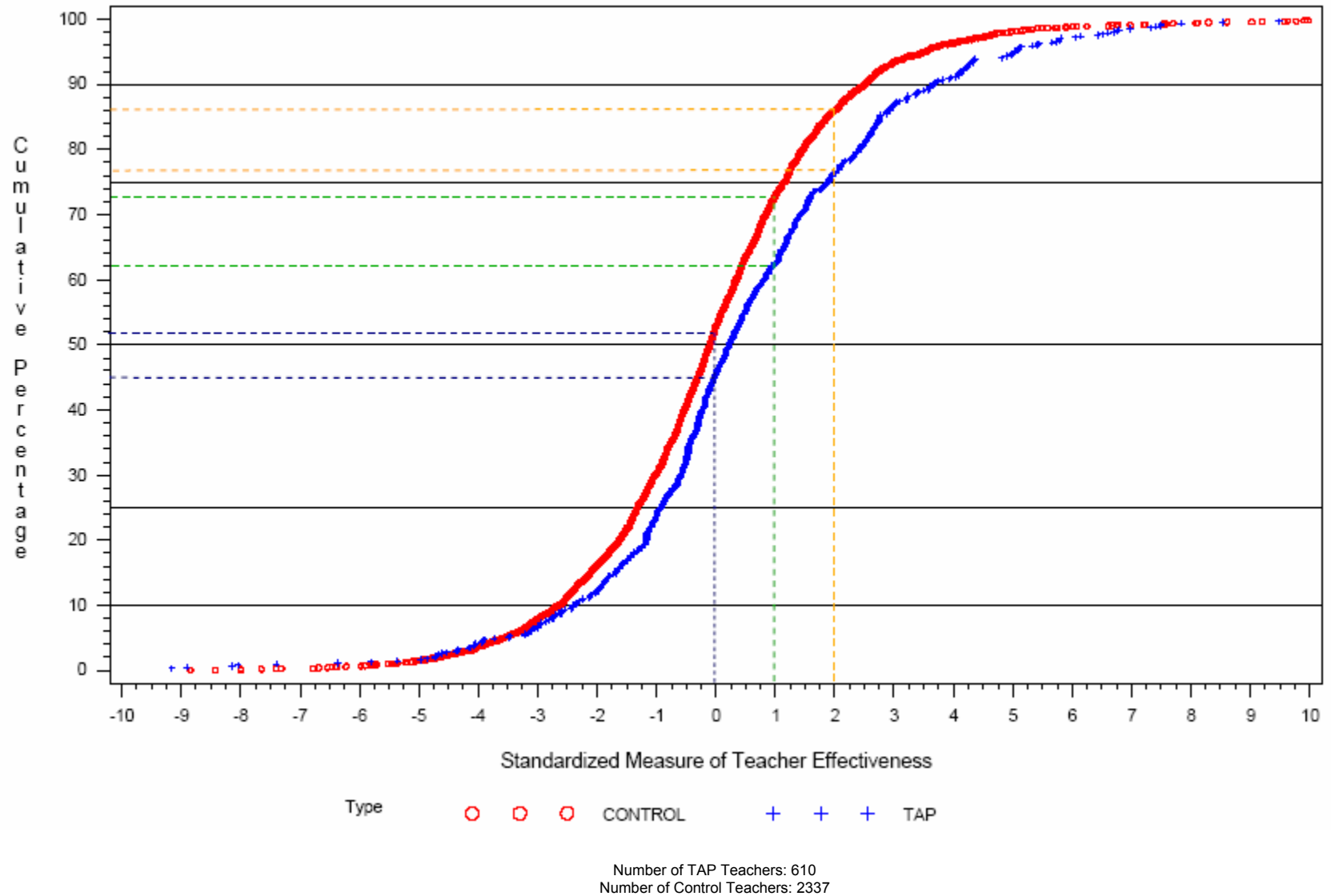
NOTE 4: Reports from the 6th grade at Crestworth Middle Magnet School, LA, were excluded because SAS[®] EVAAS[®] determined that the data showed inconsistencies due to lack of historical data and advised that these reports not be used.

State	TAP or Control	# of Teacher Reports	Percent Score 1	Percent Score 2	Percent Score 3	Percent Score 4	Percent Score 5	Percent Score 1 & 2	Percent Score 3 & 4 & 5
AR	TAP	108	0.0%	4.6%	75.0%	14.8%	5.6%	4.6%	95.4%
AR	Control	243	8.6%	16.5%	53.1%	13.6%	8.2%	25.1%	74.9%
IN	TAP	132*	6.8%	12.1%	57.6%	15.9%	7.6%	18.9%	81.1%
IN	Control	89	7.9%	11.2%	53.9%	21.3%	5.6%	19.1%	80.9%
MN	TAP	51	7.8%	13.7%	39.2%	7.8%	31.4%	21.6%	78.4%
MN	Control	659	14.3%	15.9%	44.5%	12.9%	12.4%	30.2%	69.8%
SC	TAP	179	27.9%	10.6%	22.9%	9.5%	29.1%	38.5%	61.5%
SC	Control	485	34.0%	12.4%	23.1%	9.9%	20.6%	46.4%	53.6%
FL	TAP	61	3.3%	11.5%	62.3%	16.4%	6.6%	14.8%	85.2%
FL	Control	406	9.1%	12.8%	51.2%	16.0%	10.8%	21.9%	78.1%
LA	TAP	125	1.6%	16.0%	73.6%	6.4%	2.4%	17.6%	82.4%
LA	Control	453	11.5%	13.7%	45.5%	12.4%	17.0%	25.2%	74.8%
CO	TAP	120	3.3%	4.2%	25.0%	20.8%	46.7%	7.5%	92.5%

* In Indiana there are more TAP teacher reports than their controls because control groups comprise of TAP teachers and non TAP teachers, where each teacher can serve as more than one control.

- In 6 out of 6 (100%) states, fewer TAP teachers received 2 or more standard deviations below the average teacher in the representative group, as compared to control teachers (Score of 1).
- In 6 out of 6 (100%) states, fewer TAP teachers received 1 or more standard deviations below the average teacher in the representative group, as compared to control teachers (Score of 1 and 2).
- In 6 out of 6 (100%) states, more TAP teachers received equal or more standard deviations above the average teacher in the representative group, as compared to control teachers (Score of 3, 4 and 5).
- In 3 out of 6 (50%) states, more TAP teachers received 2 or more standard deviations above the average teacher in the representative group, as compared to control teachers (Score of 5).

Appendix A—Figure 1: TAP Teachers vs Control Teachers
Cumulative Distribution Comparative Plot
Standardized Teacher Effectiveness Estimates



Appendix B—TABLE 2: TAP vs. Control Schools (Subject Composites), 2004-05

NOTE 1: Three high schools (one from each of the following states: AR, IN, and MN) were not included in this analysis due to limited available data.

NOTE 2: The confidence intervals for each score differ when subjects are combined.

NOTE 3: Colorado's results below are only a partial analysis, with respect to determining school performance bonuses, because only the data provided by NWEA (and not CSAP) are represented.

* Grade levels are combined within each of these subjects (across grades: math separate, reading separate)

**Grade levels as well as subjects are combined (across grades: math & reading combined)

State	TAP or Control	Subject	Total # of Schools	Percent of schools (TAP/Control) scoring at each quintile on each subject						
				1	2	3	4	5	1 & 2	3, 4 & 5
AR	TAP	M*	13	0%	8%	54%	23%	15%	8%	92%
AR	Control	M	60	12%	15%	47%	17%	10%	27%	73%
AR	TAP	R*	13	8%	8%	62%	23%	0%	15%	85%
AR	Control	R	60	15%	12%	47%	17%	10%	27%	73%
AR	TAP	R & M**	13	8%	0%	62%	15%	15%	8%	92%
AR	Control	R & M	60	23%	10%	35%	13%	18%	33%	67%
FL	TAP	M	5	0%	20%	20%	40%	20%	20%	80%
FL	Control	M	28	25%	11%	25%	11%	29%	36%	64%
FL	TAP	R	5	0%	20%	80%	0%	0%	20%	80%
FL	Control	R	21	10%	29%	29%	14%	19%	38%	62%
FL	TAP	R & M	5	20%	0%	60%	20%	0%	20%	80%
FL	Control	R & M	28	18%	18%	29%	11%	25%	36%	64%
IN	TAP	M	7	43%	0%	14%	29%	14%	43%	57%
IN	Control	M	21	14%	5%	48%	19%	14%	19%	81%
IN	TAP	R	7	14%	0%	57%	14%	14%	14%	86%
IN	Control	R	21	5%	19%	52%	14%	10%	24%	76%
IN	TAP	R & M	7	14%	29%	29%	14%	14%	43%	57%
IN	Control	R & M	21	10%	14%	48%	19%	10%	24%	76%

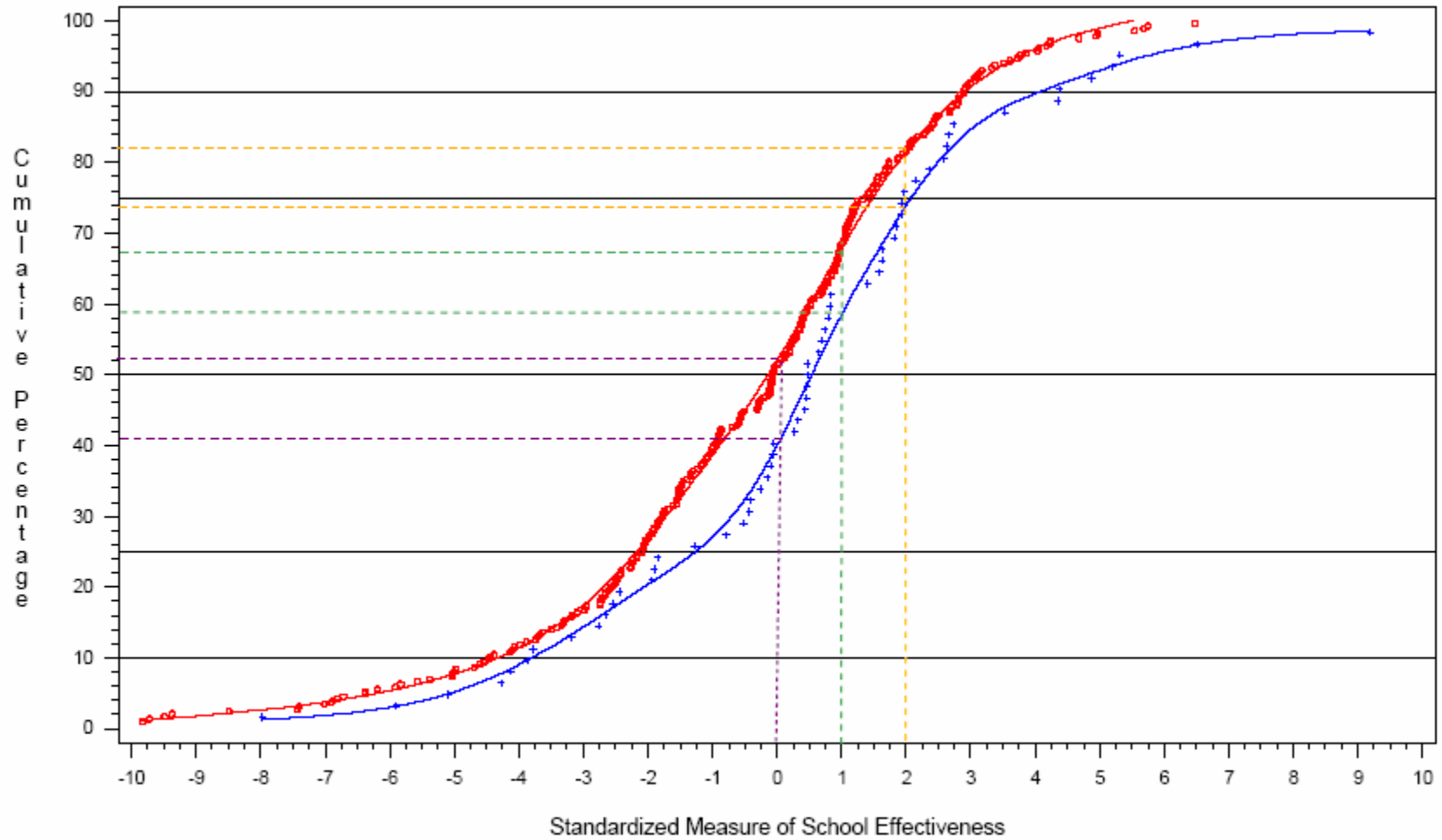
(Appendix B-Table 2 Continued)

State	TAP or Control	Subject	Total # of Schools	Percent of schools (TAP/Control) scoring at each quintile on each subject						
				1	2	3	4	5	1 & 2	3, 4 & 5
LA	TAP	M	6	33%	17%	17%	17%	17%	50%	50%
LA	Control	M	50	12%	16%	40%	16%	16%	28%	72%
LA	TAP	R	6	17%	0%	50%	33%	0%	17%	83%
LA	Control	R	50	18%	16%	30%	14%	22%	34%	66%
LA	TAP	R & M	6	33%	17%	17%	17%	17%	50%	50%
LA	Control	R & M	50	20%	10%	34%	12%	24%	30%	70%
MN	TAP	M	6	0%	17%	33%	17%	33%	17%	83%
MN	Control	M	90	16%	22%	22%	26%	14%	38%	62%
MN	TAP	R	6	0%	17%	17%	33%	33%	17%	83%
MN	Control	R	90	16%	13%	43%	12%	16%	29%	71%
MN	TAP	R & M	6	0%	0%	50%	17%	33%	0%	100%
MN	Control	R & M	90	22%	17%	26%	16%	20%	39%	61%
SC	TAP	M	9	44%	0%	44%	0%	11%	44%	56%
SC	Control	M	36	42%	11%	25%	6%	17%	53%	47%
SC	TAP	R	9	67%	0%	0%	0%	33%	67%	33%
SC	Control	R	36	75%	3%	3%	8%	11%	78%	22%
SC	TAP	R & M	9	78%	11%	0%	0%	11%	89%	11%
SC	Control	R & M	36	69%	8%	8%	8%	6%	78%	22%
CO	TAP	M	8	25%	0%	25%	0%	50%	25%	75%
CO	TAP	M	15	20%	0%	20%	7%	53%	20%	80%
CO	TAP	R	16	6%	0%	31%	25%	38%	6%	94%
CO	TAP	R	15	0%	7%	20%	20%	53%	7%	93%
CO	TAP	R & M	16	19%	0%	25%	19%	38%	19%	81%
CO	TAP	R & M	15	0%	0%	33%	20%	47%	0%	100%

(Appendix B-Table 2 Continued)

- In 4 out of 6 (67%) states, in math, fewer TAP schools received 1 or more standard deviations below the average teacher in the representative group, as compared to control teachers (Score of 1 and 2).
- In 4 out of 6 (67%) states, in math, more TAP schools received equal or more standard deviations above the average teacher in the representative group, as compared to control teachers (Score of 3, 4 and 5).
- In 6 out of 6 (100%) states, in reading, fewer TAP schools received 1 or more standard deviations below the average teacher in the representative group, as compared to control teachers (Score of 1 and 2).
- In 6 out of 6 (100%) states, in reading, more TAP schools received equal or more standard deviations above the average teacher in the representative group, as compared to control teachers (Score of 3, 4 and 5).
- In 3 out of 6 (50%) states, in reading and math combined, fewer TAP schools received 1 or more standard deviations below the average teacher in the representative group, as compared to control teachers (Score of 1 and 2).
- In 3 out of 6 (50%) states, in reading and math combined, more TAP schools received equal or more standard deviations above the average teacher in the representative group, as compared to control teachers (Score of 3, 4 and 5).

**Appendix B—Figure 2: TAP Schools vs Control Schools
Cumulative Distribution Comparative Plot
Standardized School Effectiveness Estimates**



Type ○—○—○ CONTROL +—+—+ TAP

Number of TAP Schools: 61
Number of Control Schools: 285

Appendix C—TABLE 3: Adequate Yearly Progress (AYP)

* The Minnesota Department of Education did not report AYP status for one TAP school; therefore, eleven schools (and not twelve) were included in the calculation of percentage of TAP schools making AYP in Minnesota.

** One TAP school in Texas was a new school as of the 2005-2006 school year therefore did not have an AYP status. Two schools (and not three) were included in the calculation of percentage of TAP schools making AYP in Texas.

State	2004-2005 School Year			2005-2006 School Year			Percent Free/Reduced Lunch, 2005	
	Percent of Schools Making AYP		# of TAP Schools	Percent of Schools Making AYP		# of TAP Schools	Statewide	TAP Schools
	Statewide	TAP Schools		Statewide	TAP Schools			
AR	73%	100%	14	30%	93%	14	52%	56%
CO	59%	73%	15	75%	73%	15	32%	20%
FL	36%	100%	4	28%	60%	5	46%	53%
LA	84%	83%	6	85%	93%	32	61%	73%
MN	87%	75%	7	74%	55%	12*	33%	64%
OH	76%	25%	0	61%	50%	4	26%	71%
SC	47%	12%	9	37%	9%	11	51%	81%
TX	78%	100%	0	81%	50%	3**	47%	62%

Appendix D

What is the Teacher Advancement Program (TAP)?

We all want the best possible education for our children, and research has shown that the single most important school-related factor for student success is having a talented teacher in the classroom. But unless we act now, we will come far short of having the talented teachers required to ensure that all children receive the high-quality education they need and deserve.

To address this problem, the Milken Family Foundation created the Teacher Advancement Program (TAP),⁵⁰ a bold new strategy to attract, develop, motivate and retain talented people to the teaching profession. TAP's goal is to draw more talented people to the teaching profession and keep them there by making it more attractive and rewarding to be a teacher. TAP provides the opportunity for good teachers to earn higher salaries and advance professionally, just as in other careers, without leaving the classroom. At the same time, TAP helps teachers become the best they can be, by giving them opportunities to learn better teaching strategies and by holding them accountable for their performance in the classroom.

TAP is based on four elements:

1. Multiple Career Paths

TAP allows teachers to pursue a variety of positions throughout their careers—career, mentor and master teacher—depending upon their interests, abilities, and accomplishments. As teachers move up the ranks, their qualifications, roles and responsibilities increase—as does their compensation. Multiple career paths allow good teachers to advance without having to leave the classroom.

2. Ongoing Applied Professional Growth

TAP schools must restructure their schedules to provide time during the regular school day for teachers to meet, learn, plan, mentor and share ideas with other teachers so they can constantly improve the quality of their instruction, and hence, increase their students' academic achievement. Ongoing Applied Professional Growth allows teachers to learn new instructional strategies and have greater opportunities to collaborate, both of which will lead them to become more effective teachers. It focuses on identified needs, based on the instructional issues that specific teachers face with specific students. Teachers use data to target these areas of need, instead of trying to implement the latest fads in professional development.

3. Instructionally Focused Accountability

TAP has developed a comprehensive system for evaluating teachers, and it rewards teachers for how well they teach their students. Teachers are held accountable for meeting the *TAP Teaching Skills, Knowledge and Responsibility Standards*, as well as for

⁵⁰ In 2005, the urgent need for teacher quality, coupled with experience and expertise gained from implementing the Teacher Advancement Program, led us to the establishment of an independent entity now known as the National Institute for Excellence in Teaching (NIET). For more information please see our Web site: www.talentedteachers.org.

the academic growth of their students. Every teacher in a TAP school is evaluated at least four times each year by trained and certified evaluators who include master and mentor teachers, as well as the principal.

4. Performance-Based Compensation

TAP changes the current compensation system by compensating teachers according to their roles and responsibilities, their performance in the classroom and the performance of their students. The system also supports districts in offering competitive salaries to those who teach in hard-to-staff subjects and schools.

While each TAP element is a powerful reform in its own right, it is the integration of the four elements that makes the Teacher Advancement Program comprehensive, unique and effective. Unlike other reform measures that narrowly focus on school accountability, teacher learning, or student test scores, the Teacher Advancement Program integrates these goals so that instructional effectiveness is linked directly to and measured by student achievement. For teachers, this means that their instructional decisions are made based on student data, and for students, this means that they receive targeted instruction based on their individual learning needs.

Throughout the TAP process, expert teachers support their colleagues in cluster groups where they teach proven instructional strategies that directly address identified student needs. These experts also provide in-class modeling and support so that all teachers can successfully implement these strategies with their own students. Finally, teachers are rewarded financially for the inevitable success that the data-driven and supported instruction of TAP provides.

To be a TAP school means that the faculty and administrators of the site intentionally put all four elements into action on a daily basis.