Teacher Responses to Pay-for-Performance Policies: Survey Results from Four High-Poverty, Urban School Districts

Paper to be presented at the AERA Annual Meeting

John Wells

April 2011

Westat®

Teacher Responses to Pay-for-Performance Policies: Survey Results from Four High-Poverty, Urban School Districts

Abstract

Policymakers are increasingly adopting "pay-for-performance" policies in which teachers are compensated based on their performance as measured by classroom evaluations and/or student achievement test results. Prior research has produced largely inconclusive findings concerning support among teachers for these policies and their effects on teachers. This paper analyzes teacher survey data drawn from a larger study of four high-poverty, urban school districts as they implemented a pay-for-performance initiative. The paper reports on attitudes about compensation reform in general, support for and perceived benefits and drawbacks of the policy in question, as well as on perceived effects the policy on teachers' working conditions, specifically in the areas of collaboration. Findings indicate high levels of support for pay-for-performance in general but somewhat less support for specific components of the program, such as using student performance to determine teacher pay. Responses were largely mixed regarding the effects of the policy on working conditions such as collaboration. A quasi-experimental design was used to allow for the comparison of responses on a select set of survey items, and teachers' participation in the program generally had no effect on their attitudes about compensation when compared to teachers in similar schools who had not participated in the program. Differences in survey responses based on teacher-level characteristics are also addressed, but these did not explain variations in teachers' responses. Despite teachers' generally favorable perceptions of the pay-for-performance program, their knowledge of the program's details was limited throughout the period of implementation, thus raising serious questions about the program's ability to achieve its goal of bringing about change in teacher behavior.

Teacher Responses to Pay-for-Performance Policies: Survey Results from Four High-Poverty, Urban School Districts

Introduction

Policymakers at the federal, state, and district levels are increasingly adopting pay-for-performance policies as a means to reform the traditional teacher pay structure, to increase teacher effectiveness and productivity, and, ultimately, to improve student achievement. Pay-for-performance policies are based in part on the assumption that if teachers are compensated on the basis of student performance on state-mandated achievement tests, or on the basis of their performance as measured by classroom evaluations, several outcomes will follow, including: teachers will be motivated to work harder to bring about improvements in student learning; effective teachers will be more easily distinguished from less effective ones; and larger proportions of students will achieve at high levels.

Pay-for-performance policies in education take various forms. This study examines the extent to which teachers support policies that provide bonus pay to teachers primarily on the basis of student performance, as measured primarily by student achievement results at the school- and/or classroom-levels, and teacher performance, as determined by classroom evaluations. The study also examines the effects of these policies on teachers, particularly on their levels of collaboration, motivation, and job satisfaction, as perceived by teachers.

This paper draws on teacher survey data collected as part of a five-year, mixed-method evaluation of the Ohio Teacher Incentive Fund (OTIF) program. The OTIF program was established in 2006 by the Ohio Department of Education (ODE) through a grant from the U.S. Department of Education's Teacher Incentive Fund (TIF) program. (Additional information on OTIF and TIF is included in the Policy Background section below.) The following research questions are addressed:

- Are teachers in high-poverty, urban schools supportive of pay-for-performance policies? Which specific aspects of pay-for-performance policies are most supported by teachers?
- In what ways, if any, does support for pay-for-performance policies among teachers differ based on their individual characteristics?

- In what ways, if any, does support differ among teachers in schools that have implemented a pay-for-performance policy for several years compared to teachers in similar schools that have not implemented such a policy?

- In what ways, if any, do pay-for-performance policies affect teachers, particularly in the areas of collaboration, motivation, and job satisfaction (i.e., as perceived by teachers)?

- In what ways, if any, do perceived effects of pay-for-performance on teachers' collaboration, motivation, and job satisfaction differ based on their individual characteristics?

Theoretical Framework

This study is guided by a theoretical framework derived from research in three areas: 1) the issue of support for pay-for-performance policies among teachers; 2) the impact of pay-for-performance policies on teachers; and 3) the potential for variation in responses to policies based on teachers' individual characteristics.

Examining support for and the effects of pay-for-performance policies is especially important given the rapidly increasing number of states and school districts that have begun adopting these types of policies. Despite numerous experiments with pay-for-performance policies in recent years, it is not fully clear to what extent teachers are supportive of these policies or how teachers are affected, if at all, or under what circumstances.

The issue of support is especially important when one considers the literature in the field of education policy implementation, which identifies teacher support, or "buy-in," as a critical factor in successful implementation. A series of analyses of various types of school reforms in the 1980s determined that few change efforts succeeded unless teacher commitment to the policy was developed (Firestone, 1989; Fuhrman, Clune, & Elmore, 1991; Odden & Marsh, 1988). Similarly, Odden (1991) found that teacher commitment and effort were critical especially for implementation at the school level, and that changes intended by policies rarely occurred without these factors. More recent studies of school reforms show that they are more likely to be successful when supported by teachers (Berends et al., 2002; Berends et al., 2001). With regard to pay-for-performance policies in particular, several studies in the private sector indicate that support is critical to implementation (Cooper, Dyck, & Frolich, 1992; Gross & Bacher, 1993; Welbourne & Gomez-Mejia, 1995), and, within the field of K-12 education, at least one study has shown that widespread perceptions of unfairness and lack of support among teachers led to the failure of a set of pay-for-performance policies across the country (Cornett & Gaines, 1994).

Prior research has demonstrated that teachers are often skeptical of and express limited support for new programs, due in part to their frustration with constant shifts in educational policy (Milanowski, 2006). With regard to pay-for-performance policies (i.e., in which pay is linked to student and/or teacher performance), studies have reported wide variations in levels of support. For example, Azordegan, Greenman, and Coulter (2005), Ballou and Podgursky (1993), and Heneman and Milanowski (1999) all found that teachers generally considered the possibility of receiving a bonus tied to student and/or teacher performance a desirable outcome. In contrast, Farkas, Johnson, and Foleno (2003) found that only 38 percent of teachers support financial incentives for teachers whose students score higher on standardized tests, and Goldhaber, DeArmond, and DeBurgomaster (2007) found that only 17 percent favored this type of policy.[1] A recent set of evaluation studies of the ProComp program in Denver and the Texas Educators Excellence Grant (TEEG) has found more consistent support for such policies (Springer et al., 2009; Wiley, Spindler, & Subert, 2010).

The research literature from a variety of professional fields, including business and government, suggests that pay-for-performance policies have the potential to alter individuals' behavior and levels of motivation and performance (Nelson, 2008). However, within the context of education, research generally has not documented negative effects on teachers. For example, studies by Ritter et al. (2008), Schacter et al. (2004), and Winters et al. (2008) all found that performance pay did not create a less collegial school environment.

Although a fundamental purpose of pay-for-performance policies is to help improve student achievement, it is important to determine the effects of such policies specifically on teachers because such effects could have implications for opportunities and outcomes for students. For example, if teachers consider their working conditions less favorable, and since research has shown that teachers' choices are influenced heavily by working conditions (Ingersoll, 2001; Smith & Ingersoll, 2004), this could result in a greater inclination among teachers to leave the teaching profession. Increases in teacher attrition, already an important factor in high-poverty schools (Darling-Hammond, 2004; Ingersoll, 2004), is especially important when one considers that a large body of research has demonstrated that teacher qualifications are among the most important school-based factors in determining student performance levels (Clotfelter, Ladd, & Vigdor, 2007; Goldhaber, 2002; Hanushek, Kain, & Rivkin, 2004; Wenglinsky, 2000).

A substantial body of research has demonstrated the potential for variations in responses to school reform policies at both the teacher- and school-levels (Berman & McLaughlin, 1978; Berends,

[1] It should be noted that, in some cases, there are methodological factors that limit the interpretation of the studies' findings. For example, one survey had a response rate of only 20 percent (Jacob & Springer, 2007).

Bodilly, & Kirby, 2002; Berends, Chun, Schuyler, Stockly, & Briggs, 2001; McLaughlin, 1989). Similar variations have been observed within the specific context of pay-for-performance policies in education. For example, several studies have found a relationship between teachers' years of experience and their support for pay-for-performance policies, with veteran teachers (i.e., more than 20 years of experience) expressing lower levels of support than teachers with less than 5 years of experience (Ballou & Podgursky, 1993; Goldhaber et al., 2007; Jacob & Springer, 2007; Milanowski, 2006). Lastly, several studies have found that elementary level teachers were less supportive of pay-for-performance policies than secondary level teachers (Ballou & Podgursky, 1993; Farkas et al., 2003; Goldhaber et al., 2007; Jacob & Springer, 2007).

Policy Background

It is important to define what is meant by the term "pay-for-performance" within the context of this study because these policies take various forms. This study focuses on policies that provide bonus pay to teachers based primarily on student achievement as measured at the school and/or classroom level. Brief descriptions of the federal and state-level programs that provide the funding for these policies, as well as specific details on the pay-for-performance models used in the four school districts included in this study, are provided below.

The U.S. Department of Education established the TIF program in 2006 to support efforts to develop and implement performance-based teacher and principal compensation systems in high-need schools. The federal program has four primary goals:

- Improve student achievement by increasing teacher and principal effectiveness;
- Reform teacher and principal compensation systems so that teachers and principals are rewarded for increases in student achievement;
- Increase the number of effective teachers teaching poor, minority, and disadvantaged students in hard-to-staff subjects; and
- Create sustainable performance-based compensation systems.

As part of the initial cohort of TIF grants, ODE was awarded a $20.5 million grant in 2006 to implement and evaluate the OTIF program, which creates incentives to secure the best-qualified teachers for schools with the greatest need and lowest academic performance. The OTIF program, implemented over a five-year period (i.e., the 2006-07 school year through 2010-11), seeks to ensure

that high-quality teachers and school leaders are recognized and promoted, have access to ongoing professional development, work in collaborative environments, and are compensated appropriately based on skills, knowledge, responsibilities, and student performance. OTIF was implemented in four large, urban districts—Cincinnati, Columbus, Cleveland, and Toledo.

Although the four districts share similar performance compensation models, each district is unique in how it approached performance compensation. One purpose of this approach was to test alternative models of performance compensation. The System for Teacher and Student Advancement, known as TAP, was utilized in two of the four districts, Cincinnati and Columbus. In both cases, the TAP model was selected in collaboration between district and union leaders. Cleveland and Toledo implemented home-grown models developed by the local teachers' unions, boards of education, and district staff.

The TAP model is used as a comprehensive school improvement strategy in high-need schools. The program contains four primary elements:

- Multiple career paths, in which teachers have the opportunity to advance to master or mentor teacher positions through a rigorous process that considers a teacher's skills, knowledge, and interests;

- Ongoing professional development, which requires that professional development based on the identified needs of teachers and students be provided, primarily in the form of weekly cluster groups within individual schools that are led by master or mentor teachers;

- Instructionally focused accountability, which requires each teacher to be evaluated four to six times per school year by multiple evaluators to assess teacher effectiveness; and

- Performance-based compensation, in which teachers are compensated based on the quality of instructional performance in the classroom and performance of their students. Calculation of incentive pay in the two TAP districts is based on building-level student performance and acquisition of skills and knowledge (i.e., as measured by teacher evaluations, observations, portfolios, etc.) Teachers are typically eligible for as much as $2,000, although individual payouts can be somewhat larger in some schools.[2]

The models being utilized in the two non-TAP districts contain many of the same key features as the TAP model. Teachers are provided with professional development with the aim of increasing subject matter knowledge and understanding of research-based practices. Some teachers are

[2] Award funds are established annually for each school based on $2,000 per teacher (i.e., if all eligible teachers meet the criteria, all teachers would receive $2,000); however, since schools are not required to return unspent incentive funds, an individual teacher's payout amount can be influenced by the relative performance of their colleagues within their school.

identified as lead teachers, whose roles may include implementing instructional strategies to help meet school performance goals, participating in curriculum and instructional development activities, serving as mentors, or accepting teaching assignments in schools identified as high-need or other difficult-to-fill teaching assignments. Lead teachers receive an annual stipend of up to 15 percent of their base salary, depending on which role they fulfill.

In the non-TAP districts, schools develop targeted improvement goals and determine the level of growth required for each selected goal based on rigorous criteria. There are three school-level goals—two academic goals (i.e., increases in student achievement outcomes, including a decrease in the gap in the school's performance index, a state measure used to determine a school's level of performance relative to others, and a decrease in the gap in the school's overall math percentage) and one related improvement goal (i.e., increase in graduation rate or attendance). In Cleveland, those teachers whose school meets all three goals receive $2,000 in bonus pay, and those whose school meets two of the three goals receive $1,000 (no bonus is provided if a school meets only one of the three goals). In Toledo, teachers in schools that meet all three established goals receive a $1,000 bonus, and those in schools that meet two of the three goals receive a pro-rated amount.

Methodology

This study analyzes three years of survey data collected from classroom teachers[3] in four high-poverty, urban school districts in the State of Ohio during the 2007-08, 2008-09, and 2010-11 school year.[4] The survey was designed to obtain longitudinal data on a variety of issues pertaining to compensation system reform, including teachers' levels of support for and understanding of the program, attitudes about which factors should be important in determining pay, and in what ways, if any, levels of collaboration among teachers and other working conditions have changed as a result of the implementation of the OTIF program. A quasi-experimental design is used, which allows for the comparison of responses among teachers in schools that have implemented a pay-for-performance policy with teachers in similar schools that have not implemented such a policy on a select set of survey items.

---

[3] Lead teachers in all four districts were also surveyed as part of the evaluation of the OTIF program; however, due to the small number of lead teachers in the sample, they could not be included in the primary methods of data analysis used in this study and thus were excluded from this study altogether.

[4] Although the period of implementation was from the 2006-07 school year through the 2010-11 school year, no survey data were collected in 2006-07 since it was considered a pilot year in many schools. No survey data were collected during the 2009-10 school year, for two reasons: 1) the lack of variation in survey results from the 2007-08 school year to the 2008-09 school year and 2) the evaluation's emphasis during that year on collecting case study data.

Sample Selection

The study utilizes a two-tiered sampling approach for treatment schools because the four districts involved in OTIF fall into two categories based on the number of schools participating in the program. In the two districts in which a select set of schools are participating in the program (i.e., Cincinnati and Columbus), surveys were administered to all classroom teachers in participating schools, and in the two districts in which all schools are participating (i.e., Cleveland and Toledo), surveys were administered to a random sample of classroom teachers from 10 schools in each district. In addition, surveys were administered to a set of comparison schools in Cincinnati and Columbus (i.e., where not all schools are participating) during the 2010-11 school year, primarily to gauge their attitudes about which factors should be important in determining teacher pay.

Every effort was made to ensure that the sample consisted of the same individual teachers throughout all three years of the survey's administration. In years 2 and 4, we verified our lists of teachers with school coordinators to determine if teachers from the prior year were still in the school. However, due to retirements, transfers, and turnover, a substantial number of replacements of teachers based on random selection were necessary in each year. Of the 305 teachers who responded in year 2, at least 179 (59 percent) also responded in year 1. Of the 268 teachers who responded in year 4, 136 (51 percent) had also responded in year 2, and 89 (33 percent) had also responded in year 1. A total of 80 teachers responded to the survey in all three years.

Response Rates

The survey response rates for treatment schools were high, ranging year to year from 77 percent to 85 percent, as shown in Table 1. Similarly, the within-district response rates ranged from 71 to 90 percent, while the overall response rate for the comparison group was 60 percent.

Table 1. Survey response rates, by district

District	Treatment Schools, Year 1 (n=304)	Treatment Schools, Year 2 (n=305)	Treatment Schools, Year 4 (n=268)	Comparison schools, Year 4 (n=81)
Cincinnati.....	84.9	78.7	79.4	55.4
Columbus.....	77.2	89.9	71.0	67.3
Toledo	72.0	90.0	81.0	NA
Cleveland	75.0	82.2	77.0	NA
Total.....	**77.2**	**85.0**	**77.7**	**60.0**

NA = not applicable.

Data Analysis

This study utilizes a combination of factor analysis, regression analysis, and item-level analysis. With regard to factor analysis, this approach has several advantages; most notably, it identifies relationships among individual survey items, thereby allowing for a more systematic assessment of the survey data on the issues addressed in the study. For the teacher survey, multiple questions were developed around key areas to obtain data about implementation efforts (e.g., professional development) and teacher outcomes (e.g., support for OTIF) of the program. Although item-level results are informative, it is sometimes difficult to describe survey findings in key areas holistically. Factor analysis allows for the representation of many observed variables in a few unobserved factors (also referred to as scales) and can be used not only to verify the relationships among items and factors, but also to create weighted scales driven by theory and empirical data. This process resulted in the creation of a set of five factors (described below) that allowed for the examination of the degree to which responses are consistent across similar items. The scales can be used both descriptively and as variables in statistical models. More detailed information on the procedures for scale development is included in the Appendix.

In brief, the development for the scales involved the following steps:

- Exploring the item-factor relationship with both theory-driven and data-driven analyses. The theory-driven analysis mapped survey items to program implementation features and outcomes. At the same time, relationships among items using an exploratory factor analysis (EFA) technique were examined. Then, the results between item mapping and EFA were compared and any discrepancies between the two methods were resolved.

- Confirming the link between the mapping and the data, using the confirmatory factor analysis (CFA) technique. If the observed data support the proposed theory-driven model, the relationship between the items and factors are substantial and statistically significant, and the model-data fit index will fall into a range that indicates that the model fits the data well. All of the factor models constructed in this study demonstrated a reasonable model-data fit.

- Ensuring the factors that are measuring the same traits between the groups (i.e., districts, year), using the multi-sample analyses (MSA) in the context of CFA. MSA is a powerful technique in the context of CFA that will be used to test mainly whether 1) the factor structures (i.e., number of factors) are group invariant, and 2) the patterns of factor loadings are group invariant.

- Establishing the final group-invariant factor models and calculating weighted scales. We calculated individual teacher scores by multiplying the item response with the factor loading. Therefore, the scales were weighted by the proportion of items the scale contributed to the factor. For presentation, we standardized the scale scores on a range of 1 to 100, with 100

representing the highest possible score. For example, a teacher with 100 in "support" means that she selected the highest category of response for all of the items related to support.

These steps produced five factors, which portray different aspects of implementation and outcomes from the viewpoint of teachers. Two of the factors pertain to implementation:

- Knowledge about OTIF. This factor addresses the extent of teachers' knowledge about the payment portion of the program (e.g., payout amounts, criteria used to determine payouts).[5]
- Attitude about supplemental pay. This factor examines how important various factors should be in determining a teacher's supplemental pay.

To introduce these two implementation factors and show their relationship to the teacher survey items, a crosswalk of the survey items with their corresponding factors and subfactors is presented in Exhibit 1.

Exhibit 1. Item-to-scale mapping—implementation factors

Scale	Survey Items
Implementation Factor 1: Knowledge about OTIF	Composed of all teacher knowledge items, which in some cases are unique to each respective district. Note: For ease of presentation, "Teacher Knowledge" is presented as a factor but is calculated as an index of all correct responses to all relevant items.
Implementation Factor 2: Attitude about Supplemental Pay	Highest academic degree earned
	Student performance on standardized tests, as measured at the classroom level (i.e., value-added analysis)
	Student performance on standardized tests, as measured at the school level
	Teacher performance, as determined by principal evaluations, observations, teaching portfolios, etc.
	Specific subject being taught by teacher
	Level of participation in professional development
	Fulfillment of additional roles (e.g., serving as mentor to other teachers)
	Teaching in "hard-to-staff" schools

The three other factors pertain to outcomes:

- Support for OTIF. This factor addresses the extent of teachers' opinions toward the program in general.
- Collaboration. This factor covers the questions about collaboration and interpersonal support among teachers.
- Perceived changes resulting from OTIF. This factor looks at the extent of changes teachers indicated as a result of the program's implementation.

[5] Knowledge about OTIF was not developed using the steps listed above; rather, it is an index of the percentage of correct answers.

For the three outcomes factors, a crosswalk of the survey items with their corresponding factors and subfactors is presented in Exhibit 2.

Exhibit 2. Item-to-scale mapping—outcomes factors

Scale	Subscale	Survey Items
Outcomes Factor 1: Support for OTIF		I feel I will be better rewarded financially for what I do as a teacher under the OTIF program than in prior years.
		The amount of the incentive will be large enough to motivate me to examine my teaching practices more closely.
		The OTIF program neglects to measure important aspects of my teaching performance.
		The potential monetary amount teachers can receive is pretty small in comparison to what we are being asked to do in the OTIF program.
		The OTIF program will encourage teachers to work harder than in prior years to get more pay.
		The program is unlikely to be sustained after the grant, so teachers are pessimistic about lasting results.
		I support implementing the OTIF program at my school.
		Most teachers in my school support implementing the OTIF program.
		The OTIF program can be insulting to teachers since it implies they are not doing a good job of teaching students already.
Outcomes Factor 2: Collaboration		The prospect that teachers at my school can earn a bonus encourages staff to work together.
		When teachers and administrators in a school work collaboratively, student achievement improves.
		My input is valued at my school.
		I feel supported by other teachers at my grade level.
		I am becoming a better teacher because of the support and collaboration at my school.
Outcomes Factor 3: Perceived Changes Resulting from OTIF	Positive Changes Among Teachers	I would describe teachers at this school as a more satisfied group.
		I like the way things are run at the school better.
		Teachers can be counted on more often to help each other with their teaching, even though it may not be part of their official assignment.
	Negative Changes Among Teachers	The stress and disappointments involved in teaching at this school are much greater.
		I think about transferring to another school/district more often.
		I have noticed increased resentment among teachers.
		Teachers seem more competitive than cooperative.
	Changes in Working With Students	Teachers more often expect students to complete every assignment.
		Teachers more often encourage students to keep trying even when the work is challenging.

Multiple regression analysis was also conducted on each factor. The purpose of this analysis was to explore the extent to which variations in teachers' responses to survey items could be explained by their individual characteristics. The characteristics, or covariates, included in the analysis are years of teaching experience, grade level taught, subject taught, receipt of a bonus payment (i.e., had or had not received a payment under the program), and extent of knowledge about the OTIF program. Teachers' responses within year 2 and year 4 were examined as part of the regression analysis; however, given the substantial number of replacements in the survey sample from year to year, the

extent to which changes in teachers' responses from year to year could be explained by these characteristics were not part of the regression analysis.

Several levels of analysis of the survey data at the item level were also conducted. Descriptive analyses at the item level were conducted within each of the three years. Statistical tests of significance at the item level were conducted to determine change over time from year 1 to year 2; however, tests of significance at the item level were not conducted to determine change from year 1 to year 4, for two reasons; 1) very few significant differences in teachers' responses from year 1 to year 2 were observed at the item level and 2) very few significant differences in teachers' responses from year 1 to year 4 were observed at the factor level. Therefore, the item level results presented in this paper are descriptive and focus only on findings within year 4.

Limitations

Several methodological limitations should be noted. The first pertains to the issue of generalizability. The findings presented in this paper reflect only what has occurred within four large, high-poverty, urban school districts and the individual schools included in the study. Although one purpose is to better understand the impacts of pay-for-performance policies, the results are not necessarily representative of schools in other types of districts.

The fact that the data used in the paper are comprised of teachers' perceptions of the types of events that have transpired in their respective schools and classrooms presents several limitations. First, some of the concepts measured in this paper, such as collaboration, cannot be quantified or otherwise measured with any significant degree of precision. Also, in using surveys of teachers as a data source, the paper relies on teachers' capacity to describe their own responses to the policy, with its potential for subjectivity. In addition, by relying on teachers' perceptions, the paper does not seek to establish cause and effect, or any other direct relationship between pay-for-performance policies and subsequent changes in behavior. Nevertheless, the perceptions of teachers are of considerable importance, since they are key players in the implementation of the policy and their perceptions can be expected to influence responses to the policies.

Finally, the study does not fully take into account the various contextual factors at the district- and school-levels that may help explain teachers' responses. As Grant (2001) and others have found, educators' behavior and policy responses are likely to be influenced by a range of factors, and acknowledging that educators are subject to multiple influences provides a richer understanding of what is occurring. This is especially important when one considers that many of the areas addressed

in this study, such as support and effects on collaboration and motivation, have been shown to be influenced by a wide range of factors. For example, research has shown that school leadership represents a strong influence on all of these areas; teachers in school characterized as having "effective" or "transformational" leadership are more likely than those in other schools to have a deeper commitment to the school and its students, and to working with others to improve performance (Leithwood et al., 2002; Leithwood, 2001; Leithwood & Jantzi, 1999; Quinn, 2002; Ross & Gray, 2006). It is possible that these types of school-level differences may influence educators' perceptions of and responses to the same policy.

Findings

This section presents findings from the three years of teacher surveys.[6] We present the survey results within five factors, as outlined above, with each representing a key area and comprising a set of survey items that relate to that area. The first section discusses findings pertaining to the implementation of the program. The following two factors are addressed:

- Knowledge about the OTIF program;
- Attitudes about supplemental pay;

The second section presents results pertaining to outcomes and includes the following three factors:

- Support for the program;
- Collaboration; and
- Perceived changes resulting from the program.

For each of the factors, the findings address how teachers responded within year 1, year 2, and year 4 (i.e., during the fifth and final year of the program's implementation), both overall and by district, as well as the extent to which teachers' responses varied from year to year, both overall and by district. The factor analysis resulted in a score for each factor based on a scale of 1 to 100, and the results are presented largely on the basis of this scale. As noted above, multiple regression analysis was conducted on each factor, and the results for items that pertain to the implementation scales are

[6] It should be noted that although the findings refer to "year 1," "year 2," and year 4," the project was actually in its second year of implementation when the first year of survey data were collected (i.e., during the 2007–08 school year); thus, the "year 4" survey data were collected during the fifth and final year of the program's implementation.

presented at the end of that section, under the heading of Correlational Relationships, and the same is the case with respect to the outcomes scales. With respect to item-level results, since most of the items in the survey utilize a four-point response scale (i.e., strongly agree, agree, disagree, strongly disagree), we present these results as percentages (e.g., the percentage who agreed or strongly agreed); for items that do not include a four-point response scale, we report on the percentages of teachers who responded in a particular way (e.g., the percentage of teacher who responded "yes"). It should also be noted that the item level results presented here are limited to year 4 (as explained in the Methodology section), and they are presented only for those survey items in which it is useful to elaborate on results beyond the factor level.

Findings on Implementation

Knowledge about the OTIF Program

This factor includes all of the survey items that were designed to measure how effectively information about the payment portion of the program had been communicated to teachers. In the survey, teachers were asked to either respond "yes" or "no" to a series of statements, some of which were accurate and some not accurate. In our analysis, we examined the percentage of teachers who correctly labeled each statement as accurate/not accurate.

As shown in Table 1, the results for this factor indicate that in year 4 (i.e., the fifth and final year of implementation), teachers' knowledge was low, with a score of 31 on a scale of 1 to 100, with 100 representing the maximum accuracy (i.e., all teachers correctly labeling all of the statements as accurate/not accurate). The overall score of 49 in year 2 represents a statistically significant gain in knowledge from year 1, in which the score was more than 16 points lower (33 on a scale of 1 to 100). All of the districts with the exception of Cleveland had experienced a significant increase from year 1 to year 2 (Cleveland's scores were consistently lower than those in the three other districts, with the exception of Cincinnati in year 1, which was also very low). Even at its peak in year 2, knowledge was still low in absolute terms, with teachers across the four districts responding correctly only about half the time. In summary, the survey data also show that the gains in teachers' knowledge that had occurred in year 2 had eroded by year 4, in which the overall score went back down to 31.

Table 2. Knowledge about OTIF

Knowledge	Overall	Cincinnati	Columbus	Cleveland	Toledo
Status (year 1)	33.0	21.2	43.9*	26.8	41.1*
Status (year 2)	49.3	61.0	61.3	28.8*	50.7
Change (year 1-year 2).....	16.3*	39.8*	17.4*	2.1	9.6*
Status (year 4)	31.2	28.1	37.9	23.4*	37.4
Change (year 2-year 4).....	-18.1*	-32.8*	-23.4*	-5.5	-13.3
Change (year 1-year 4).....	-1.8	7.0	-6.0*	-3.4	-3.8

* statistically significant difference at the .05 level.

Item-level results from across the three years indicate that in the TAP districts, teachers were more knowledgeable about the criteria for bonus payments but less knowledgeable about the weight assigned to specific criteria when determining payouts. For example, while most were correct that school-wide student performance and teacher skills and knowledge performance were the two levels by which pay is determined, fewer teachers were correct that half of their incentive pay was based on a teacher's demonstration of skills and knowledge. In Toledo, teachers generally did not know what the criteria were for determining payouts or the weight assigned to specific criteria, with less than one-quarter responding correctly to those items. Furthermore, both in the TAP districts and in Toledo, only half of all teachers were knowledgeable regarding the total incentive amount for which they were eligible. In Cleveland, there were no instances in which more than half of teachers responded correctly to any of the items pertaining to knowledge.

Attitudes about Supplemental Pay

This factor examines various factors that teachers felt should or should not be important in determining supplemental pay. Strictly speaking, teachers' attitudes about supplemental pay do not represent an implementation issue; rather, it represents a preexisting factor that will influence the implementation as well as the outcome of the program.

As Table 3 shows, teachers were somewhat positive about the factors that are included in the scale and this remained consistent over time; overall scores ranged from 50 to 53 points on a scale of 1 to 100, with 100 representing all teachers regarding all of the factors included in the scale as being important to determining supplemental pay.[7] With regard to differences by district, the score for teachers in Cincinnati was significantly higher than in all three of the other districts in all three years of the survey, as shown in Table 3.

[7] This survey item asks only if specific factors should be considered by teachers to be important or not important in determining supplemental pay. The item was not designed to ask teachers directly whether or not they support the idea of supplemental pay.

Table 3. Attitudes about supplemental pay

Attitudes	Overall	Cincinnati	Columbus	Cleveland	Toledo
Status (year 1).....	53.1	63.8*	52.3	52.1	44.7
Status (year 2).....	53.0	66.1*	51.3	55.6	42.9
Change (year 1-year 2)	0.0	2.3	-1.0	3.6	-1.8
Status (year 4).....	50.1	58.2*	53.6	51.3	40.6
Change (year 2-year 4)	-2.9	-7.9	2.4	-4.3	-2.3
Change (year 1-year 4)	-2.9	-5.5	1.3	-0.7	-4.1

* statistically significant difference at the .05 level.

This set of survey items formed the centerpiece of the comparison group survey. No significant differences were found at the factor level between teachers in the treatment schools and those in the comparison schools. Therefore, this study provides no evidence to suggest that participation in this pay-for-performance program causes teachers to view the key factors related to performance—such as teacher evaluation, student performance as measured at the classroom level, and student performance as measured at the school level—as any more or less important in determining pay.

Item-level results from year 4 show that the specific factors that teachers overall felt should be important to supplemental pay include the following:

- Fulfillment of additional roles (e.g., serving as a mentor to other teachers) (70 percent);
- Teaching in "hard-to-staff" schools (69 percent);
- Level of participation in professional development (61 percent); and
- Teacher performance, as determined by principal evaluations, observations, teaching portfolios, etc. (58 percent).

However, other critical factors associated with the OTIF model, such as student performance on standardized tests as measured at the classroom level and student performance as measured at the school level, were considered important by smaller but substantial percentages of classroom teachers (41 and 38 percent, respectively). The fact that fewer than half of all classroom teachers considered student performance at either the school or classroom level as important to supplemental pay is noteworthy, given the importance of student performance in influencing whether or not supplemental pay is received under OTIF.[8]

[8] The item is not intended to measure the degree to which a teacher supports or opposes a particular factor. The response options were only "important" or "not important."

Correlational Relationships (Implementation Factors)

This subsection presents the summary results of multiple regressions between implementation scales and potential covariates including years of experience, grade span taught, subject taught, receipt of a bonus payment, and level of knowledge about the OTIF program.

- Years of experience. In year 2, teachers with more than 20 years of experience were less likely than teachers with 10-19 years of experience to view the factors related to performance that are associated with the OTIF model as important in determining supplemental pay. However, in year 4, neither of the implementation factors was related to experience.

- Grade span taught. Compared to high school teachers, teachers at both the elementary and middle levels were less likely in year 2 to view the factors related to performance that are associated with the OTIF model as important in determining supplemental pay. In year 4, this difference held for elementary teachers but there was no significant difference between middle school teachers and high school teachers.

- Subject taught. In years 2 and 4, neither of the implementation factors was related to subject taught.

- Receipt of a bonus payment. In year 2, teachers who had received a bonus payment were more likely than those who had not to view the factors related to performance that are associated with the OTIF model as important in determining supplemental pay. However, in year 4, neither of the implementation factors was related to whether or not individual teachers had received a bonus payment.

- Level of knowledge about OTIF. In years 2 and 4, neither of the implementation factors was related to individual teachers' level of knowledge about the OTIF program.

Findings on Outcomes

Support for OTIF

This factor addresses the results of items on teachers' support for the program in general. As Table 5 shows, overall, teachers' support for the program was very consistent across the three years, with scores within the 62 to 64 point range on a scale of 1 to 100, with 100 representing all teachers selecting the most supportive response on all of the items included in the scale. With regard to differences by district, support among teachers in Cincinnati was significantly stronger than in the other three districts in year 2, and support among teachers in Columbus was significantly lower than in the other three districts in year 4, which is also noted in Table 5. However, neither difference persisted from year to year.

Table 5. Support for OTIF

Support	Overall	Cincinnati	Columbus	Cleveland	Toledo
Status (year 1).....	62.7	64.4	59.3	62.1	65.4
Status (year 2).....	63.2	68.3*	59.1	60.0	64.5
Change (year 1-year 2)	0.5	3.8	-0.2	-2.1	-0.9
Status (year 4).....	64.4	67.5	59.7*	62.4	66.4
Change (year 2-year 4)	1.2	-0.7	0.6	2.4	1.9
Change (year 1-year 4)	1.7	3.1	0.4	0.3	1.0

* statistically significant difference at the .05 level.

At the item level, the mean responses of teachers overall in year 4 also indicate that they were supportive of the program in general; for example, 77 percent of teachers agreed or strongly agreed with the statement, "I support implementing the program at my school." However, the results suggest that teachers did not necessarily agree that motivation was affected by the program, with more than half of teachers overall disagreeing or strongly disagreeing with the following items:

- The program will encourage teachers to work harder than in prior years to get more pay (57 percent); and
- The amount of the incentive will be large enough to motivate me to examine my teaching practices more closely (54 percent).

Collaboration

This factor addresses perceptions of the sense of collaboration and support among teachers. As Table 6 shows, teachers overall had a strong sense of collaboration across the three years, with scores within the 72 to 74 points range on a scale of 1 to 100, with 100 representing all teachers selecting the highest possible response on all items included in the scale. With regard to differences by district, the collaboration score for Cleveland teachers was significantly lower than for the other districts in year 1, and the score for Columbus teachers was significantly lower than for the other districts in year 2, as noted in Table 6. However, neither of these differences persisted from year to year.

Table 6. Collaboration

Collaboration	Overall	Cincinnati	Columbus	Cleveland	Toledo
Status (year 1).....	73.3	74.6	73.2	67.7*	77.7
Status (year 2).....	72.8	73.9	68.2*	74.2	73.9
Change (year 1-year 2)	-0.5	-0.7	0.5	1.0	-3.8
Status (year 4).....	74.6	74.1	69.6	75.2	77.3
Change (year 2-year 4)	1.8	0.2	1.4	1.0	3.4
Change (year 1-year 4)	1.3	-0.5	1.9	2.0	-0.4

* statistically significant difference at the .05 level.

Item-level results in year 4 indicate that although teachers had positive views of how teachers in their school collaborated, their responses were mixed on the specific effects of bonus pay on collaboration; for example, less than half (46 percent) agreed or strongly agreed with the statement, "The prospect that teachers at my school can earn a bonus encourages staff to work together."

Perceived Changes Resulting From OTIF

This factor addresses the extent of change teachers perceived as a result of the implementation of the program. It covers multiple components, including positive changes among teachers (e.g., higher levels of satisfaction), negative changes among teachers (e.g., higher levels of resentment or competition), as well as changes in working with students (e.g., expectations on students).

As Table 7 shows, the responses from classroom teachers regarding changes were largely positive and consistent throughout the three years, with scores within the 65 to 66 point range on a scale of 1 to 100 representing all teachers selecting the highest possible response on the items concerning positive changes and all teachers selecting the lowest possible response on the items concerning negative changes (Table 3-11). At the district level, the scores for teachers in Columbus were significantly lower than in the other three districts in all three years.

Table 7. Perceived changes resulting from OTIF

Perceived changes	Overall	Cincinnati	Columbus	Cleveland	Toledo
Status (year 1).....	65.3	67.8	60.0*	65.0	68.2
Status (year 2).....	66.1	68.7	59.9*	66.9	67.5
Change (year 1-year 2)	0.8	0.9	-0.1	1.9	-0.7
Status (year 4).....	66.8	66.6	62.2*	68.8	68.1
Change (year 2-year 4)	0.8	-2.1	2.2	1.8	0.6
Change (year 1-year 4)	1.6	-1.2	2.2	3.7	-0.1

* statistically significant difference at the .05 level.

At the item level, teachers' responses overall in year 4 were consistent across the individual items included in this scale; for example, relatively few teachers agreed or strongly agreed with the following items, which suggests that adverse impacts of the program were limited:

- Teachers seem more competitive than cooperative (13 percent);
- I think about transferring to another school/district more often (24 percent); and
- I have noticed increased resentment among teachers (33 percent).

It is worth noting that two-thirds of teachers (67 percent) agreed or strongly agreed with the statement, "I have become a more effective teacher" as a result of the OTIF program.

Correlational Relationships (Outcomes Factors)

This subsection presents the summary results of multiple regressions between outcomes scales and potential covariates including years of experience, grade span taught, subject taught, receipt of a bonus payout, and level of knowledge about the OTIF program.

- Years of experience. Compared to teachers with 10 to 19 years of experience, teachers who have taught more than 20 years reported a lower level of support for the program in year 4. In addition, in year 4, teachers with more than 20 years of experience were also less likely to report that positive changes had occurred in their school as a result of the program. Although neither of these differences was found in year 2, teachers with more than 20 years of experience did report a lower sense of collaboration in year 2 than those with 10 to 19 years of experience.

- Grade span taught. In year 2, middle school teachers were less likely than high school teachers to report that positive changes had occurred in their school as a result of the program. However, in year 4, none of the outcomes factors were related to grade span.

- Subject taught. In year 4, reading teachers reported a higher sense of collaboration than teachers of math and science. Reading teachers were also more likely to report that positive changes had occurred in their school as a result of the program. However, in year 2, none of the outcomes factors were related to subject taught.

- Receipt of a bonus payout. In years 2 and 4, none of the outcomes factors were related to whether or not individual teachers had received a bonus payout.

- Level of knowledge about OTIF. In year 4, teachers who were more knowledgeable about the OTIF program were more likely to be supportive of its implementation and were more likely than teachers with less knowledge to report that positive changes had occurred in their school. In year 2, none of the outcomes factors were related to individual teachers' level of knowledge.

Conclusions

Examining teachers' responses to pay-for-performance policies is especially important given the rapidly increasing number of states and school districts that have begun adopting these types of policies. Despite numerous experiments with pay-for-performance policies in recent years, it is not fully clear to what extent teachers are supportive of these policies or in what ways, if any, teachers are affected, or under what circumstances. This paper has resulted in several findings that add some clarification in these areas; in sum, we found the following:

Teachers' support for the pay-for-performance policies was high, with one exception being that fewer teachers support rewarding teachers based on student performance on tests as measured at the classroom-level and school-level. While less than half of teachers indicated that student performance should be important in determining supplemental pay, factors such as teacher evaluations, participation in professional development, fulfillment of additional roles, and teaching in "hard-to-staff" schools were each cited by a majority of teachers as important in determining supplemental pay. Teachers were more supportive of the program as a whole than the bonus pay component specifically, with over half of teachers indicating, for example, that "the extra money is not really a major part" of the program. This suggests that teachers may have placed more value in the professional development they received, the sense of professional recognition they felt (i.e., by received a bonus), or the enhanced sense of collaboration they reported as a result of setting school-wide performance goals.

Teachers' knowledge of the program was limited, with some exceptions, throughout the five year period of implementation. Teachers had considerable difficulty when asked to indicate in a survey if statements regarding the payment portion of the program were accurate or not. A substantial increase in knowledge among teachers occurred in the second year of implementation, but such increases in knowledge had largely eroded by the final year of the program. Knowledge was especially low in Cleveland throughout the three years. Limited understanding of the policies among teachers raises serious questions with regard to the policies' ability to leverage changes in teacher behavior.

Teachers' participation in the program generally had no effect on their views on compensation and related issues. Survey responses of teachers indicate that their views on issues such as which types of factors should determine pay and perceptions of their working conditions (e.g., the importance of teacher collaboration) remained largely unchanged as a result of their participation in the program; survey data also show that participation had little to no effect on teachers' views about compensation and their perceptions of their working conditions as compared to teachers in schools that did not participate in the program. Based on the lack of significant differences between the responses of teachers in treatment schools and those in comparison schools, there is no evidence to suggest that participation in a pay-for-performance program leads teachers to change their views on compensation.

Teachers expressed mixed opinions regarding changes that have occurred as a result of the implementation of the program. Some positive changes were cited, including higher rates of collaboration among teachers; however, there were some indications that many teachers felt

increased resentment among their colleagues. There is some evidence that these variations were driven by district, with teachers in Columbus less likely than those in the other three districts to report positive changes as a result of the program in all three years. Although statistically significant, such differences cannot explain why, for example, one-third of all teachers reported increased resentment among their colleagues. Thus, it is likely that there were school-level factors (i.e., other than grade level, which was accounted for in this study) that influenced teachers' perceptions of changes that occurred from the program.

Some of the variations in teachers' survey responses can be explained by district. In addition to differences in levels of knowledge, some differences were found by district in nearly all of the five factors. Teachers in Cincinnati had more positive attitudes about performance-based pay throughout the three years, while teachers in Columbus responded less favorably in all three years than those in other districts concerning changes that had resulted as a result of the program. Given that Cincinnati and Columbus both implemented the TAP model (with only minor accommodations to the model to fit the context of each district) but teachers in those two districts differed in their perceptions of the program, there is no evidence to suggest that the type of pay-for-performance model played a role in influencing teachers' responses. It should be noted that Columbus was the only district in which student performance at the classroom level was part of the criteria for determining teachers' pay; however, the survey was not designed to examine whether that could explain the less favorable perceptions of the program among teachers in that district.

Variations in teachers' survey responses cannot be explained by their individual characteristics, such as their years of experience, the grades and subjects they taught, whether they had received a bonus payout, and their level of knowledge about the program. A substantial body of prior research suggests that the potential exists for variations in responses to school reform policies based on teachers' individual characteristics. Nevertheless, this study found very few significant relationships between the aforementioned characteristics and how teachers responded. None of the relationships that were found persisted from year to year. Perhaps most notably, the receipt of a bonus payout had no significant effects on individual teachers' experiences in and perceptions of the program.

Lastly, although this paper helps address some of the gaps in understanding of pay-for-performance policies, there are a couple of factors to keep in mind when interpreting its findings. First, the survey results indicate that teachers' knowledge of the program was very limited. Although teachers expressed high levels of support for the program overall, it is unclear which specific aspects of the program they had in mind when responding to the survey items. It is likely that the teachers unions

played an important role in generating support for and favorable perceptions of the program; OTIF is distinct from many other recently adopted pay-for-performance programs in the sense that it received substantial support from the local teachers unions and that the unions played a key role in the program's design phase. There are some indications that when teachers are provided opportunities to help shape pay-for-performance programs, they respond more favorably.[9] Although this study found that, in year 4, teachers with more knowledge of the program were more likely to support its implementation, it remains unclear to what extent the support expressed reflects teachers' careful consideration of the policy versus the unions' endorsement of it.

Another factor to consider when interpreting the findings pertains to the relatively small amounts of the bonus payments offered in the OTIF program. The four districts each decided to keep the dollar values relatively small, in the range of $1,000 to $3,000 on average; relative to other pay-for-performance initiatives, these incentive amounts are modest at best (Prince et al., 2010). These decisions were based on the need for long-term financial sustainability of the initiative, a desire to distribute incentives to as many educators as possible, and reluctance to avoid differentiating teachers too greatly. The combination of small incentive amounts and teachers' limited knowledge of the program begs the question of whether larger bonus amounts would have garnered increased attention among teachers, thereby resulting in more knowledge about the specific features of the pay-for-performance program, and, ultimately, led to different outcomes.

[9] For example, Chicago's Recognizing Excellence in Academic Leadership (REAL) program, a Teacher Incentive Fund recipient in 2007, was initially opposed by the Chicago Teachers Union, but after members were given an opportunity to help shape the plan, the union gave its endorsement (Rossi, 2007; Dell'Angela, 2007).

References

Azordegan, J., Greenman, J., & Coulter, T. (2005). *Diversifying teacher compensation*. Denver, CO: Education Commission of the States.

Ballou, D. & Podgursky, M. (1993). "Teachers' Attitudes toward Merit Pay: Examining Conventional Wisdom." *Industrial and Labor Relations Review*, 47(1), 50-61.

Ballou, D., Sanders, W., & Wright, P. (2004). "Controlling for student background in value-added assessment of teachers." *Journal of Educational and Behavioral Statistics,* 29(1), 37-66.

Berends, M., Bodilly, S., & Kirby, S. (2002). "Looking Back over a Decade of Whole-School Reform: The Experience of New American Schools." *Phi Delta Kappan,* 84(2), 168-75.

Berends, M. Chun, J. Schuyler, G. Stockly, S. (2001). *Challenges of Conflicting School Reforms: Effects of American Schools in New American Schools in a High-Poverty Districts*. Santa Monica, CA: RAND.

Berman, P. & McLaughlin, M. W. (1978). *Federal programs supporting educational change: Vol. 8. Implementing and sustaining innovations*. Santa Monica, CA: RAND.

Clotfelter, C., Ladd, H. F., & Vigdor, J. (2007). *How and why do teacher credentials matter for student achievement?* (NBER Working Paper No. 12828). Cambridge, MA: National Bureau for Economic Research.

Cooper, C. L., Dyck, B., & Frohlich, N. (1992). "Improving the effectiveness of gainsharing: The role of fairness and participation." *Administrative Science Quarterly*, 37(3), 471-90.

Cornett, L., & Gaines, G. (1994). *Reflecting on Ten Years of Incentive Programs: The 1993 SREB Career Ladder Clearinghouse Survey*. Atlanta, GA: Southern Regional Education Board.

Darling-Hammond, L. (2004). "Standards, Accountability, and School Reform." *Teachers College Record,* 106(6), 1047-1085.

Dell'Angela, T. (2007, March 26). Top schools get a freer hand. *Chicago Tribune.*

Farkas, S., Johnson, J., & Foleno, T. (2003). *Stand by me: What teachers really think about unions, merit pay, and other professional matters*. New York: Public Agenda.

Figlio, D., & Kenny, L. (2007). "Individual teacher incentives and student performance." *Journal of Public Economics*,91(5-6), 901-14.

Firestone, W. A. (1989). "Educational Policy as an Ecology of Games." *Educational Researcher,* 18(7), 18-24.

Fuhrman, S., Clune, W., & Elmore, R. (1991). "Research on Education Reform: Lessons on the Implementation of Policy." In Odden, A. (Ed.), *Education Policy Implementation*. Albany, NY: State University of New York Press.

Goldhaber, D. (2008*). The Politics of Teacher Pay Reform*. Nashville, TN: National Center on Performance Incentives.

Goldhaber, D., DeArmond, M., & DeBurgomaster, S. (2007). *Teacher Attitudes About Compensation Reform: Implications for Reform Implementation*. Seattle, WA: University of Washington, Center on Reinventing Public Education.

Goldhaber, D. (2002). The mystery of good teaching. *Education Next, 2*(1), 50-55.

Gross, S. E., & Bacher, J. P. (1993). "The new variable pay programs: How some succeed, why some don't." *Compensation and Benefits Review*, 25(1), 51-56.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). "Why Public Schools Lose Teachers." *Journal of Human Resources*, 39(2), 326-54.

Heneman, H. & Milanowski, A. (1999). "Teachers' Attitudes about Teacher Bonuses under School-Based Performance Award Programs." *Journal of Personnel Evaluation in Education,* 12(4), 327-41.

Ingersoll, R. (2001). "Teacher turnover and teacher shortages: An organizational analysis." *American Educational Research Journal,* 38(3), 499-534.

Ingersoll, R. (2004). *Why do high-poverty schools have difficulty staffing their classrooms with qualified teachers?* Washington, DC: Center for American Progress.

Jacob, B. & Springer, M. G. (2007). *Teacher Attitudes on Pay for Performance: A Pilot Study*. Nashville, TN: National Center for Performance Incentives.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City*. National Bureau for Economic Research, Working Paper 12155. Cambridge: NBER.

Kelly, C., Odden, A., Milanowski, A. & Heneman, H. (2000). *The Motivational Effects of School-Based Performance Awards*. Philadelphia, PA: Consortium for Policy Research in Education.

Ladd, H. (1999). "The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes." *Economics of Education Review*, 18(1), 1-16.

Lavy, V. (2002). "Evaluating the effect of teachers' group performance incentives on pupil achievement." *Journal of Political Economy,* 110(6), 1286-1317.

McLaughlin, M. W. (1989). *The RAND Change Agent Study Ten Years Later: Macro Perspectives and Micro Realities*. Santa Monica, CA: Center for Research on the Context of Secondary School Teaching.

McLaughlin, M. W., & Marsh, D. (1978). "Staff Development and School Change." *Teachers College Record,* 80(1), p. 69-94.

Milanowski, A. (2006). *Performance Pay System Preferences of Students Preparing to Be Teachers.* Madison, WI: Consortium for Policy Research in Education.

Nelson, S. (2008). *Performance-Based Pay in the Federal Government*. Nashville, TN: National Center on Performance Incentives.

Odden, A. (1991). *Education Policy Implementation*. Albany, NY: State University of New York Press.

Odden, A., & Marsh, D. (1988). "How Comprehensive Reform Legislation Can Improve Secondary Schools." *Phi Delta Kappan,* 69(8), 593-98.

Podgursky, M. J. & Springer, M. G. (2006). *Teacher Performance Pay: A Review*. Nashville, TN: National Center on Performance Incentives.

Prince, C. D., Koppich, J., Azar, T. M., and Witham, P. J. (2010). *Research Synthesis: Questions Specific to Performance Pay: How Large Do Incentives Need to Be in Order to Be Effective?* Center for Educator Compensation Reform. U.S. Department of Education, Office of Elementary and Secondary Education, Washington, DC.

Ritter, G. W., Holley, M. J., Jensen, N. C., Riffel, B. E., Winters, M. A., Barnett, J. H. & Greene, J. (2008). *Year Two Evaluation of the Achievement Challenge Pilot Project in the Little Rock Public School District.* Fayetteville, AR: University of Arkansas, Department of Education Reform.

Rossi, R. (2007, September 4). New incentive for teachers. *Chicago Sun-Times.*

Schacter, J., Schiff, T. Thum, Y. M., Fagnano, C., Bendotti, M., Solomon, L., Firetag, K., & Milken, L. (2004). *The Impact of the Teacher Advancement Program on Student Achievement, Teacher Attitudes, and Job Satisfaction.* Santa Monica, CA: Milken Family Foundation.

Smith, T. M., & Ingersoll, R. M. (2004). "What are effects of induction and mentoring on beginning teacher turnover?" *American Educational Research Journal,* 41(3), 681-715.

Springer, M. G., Podgursky, M. J., Lewis, J. L., Ehlert, M. W., Gronberg, T. J., Hamilton, L. S., Jansen, D. W., Lopez, O. S., Peng, A., Stecher, B. M., & Taylor, L. L. (2009). *Texas Educator Excellence Grant (TEEG) Program: Year Two Evaluation Report.* Nashville, TN: National Center on Performance Incentives.

Vigdor, J. L. (2008). *Teacher Salary Bonuses in North Carolina*. Nashville, TN: National Center on Performance Incentives.

Welbourne, T. M., & Gomez-Mejia, L. R. (1995). "Gainsharing: A critical review and a future research agenda." *Journal of Management,* 21(3), 559-609.

Wenglinsky, H. (2000). *How Teaching Matters: Bringing the Classroom Back Into Discussion of Teacher Quality.* Princeton, NJ: Educational Testing Service.

Wiley, E. W., Spindler, E. R., & Seubert, A. N. (2010). *Denver ProComp: An Outcomes Evaluation of Denver's Alternative Teacher Compensation System.* Boulder, CO: University of Colorado.

Winters, M., Greene, J., Ritter, G. & Marsh, R. (2008). *The Effect of Performance-Pay in Little Rock, Arkansas, on Student Achievement*. Little Rock, AR: University of Arkansas.

Appendix

This appendix provides additional information about the procedures used to conduct the factor analysis.

For the teacher survey, multiple questions were developed around key areas to obtain data about the implementation efforts (e.g., professional development) and teacher outcomes (e.g., support of OTIF) of the program. While the descriptive statistics at the item level were informative, it was difficult to describe the findings in those key areas holistically. Therefore, in addition to analyzing survey data at the item level, we introduced a statistical technique, factor analysis, to examine the data.[1]

Factor analysis is a useful technique to explain variability among observed variables in terms of fewer unobserved factors. As a result, factor analysis can be used not only to verify the relationships among items and with factors, but also to calculate weighted scales driven by both theories and empirical data. By using this technique, we can assess the program's implementation efforts and its impact on teachers systematically. We used LISREL for this analysis.

The development for the scales involved the following steps:

- Exploring the item-factor relationship with both theory-driven and data-driven analyses;
- Confirming the link between the mapping and the data, using the confirmatory factor analysis (CFA) technique;
- Ensuring the factors are measuring the same traits between the groups (i.e., districts), using the multi-sample analyses (MSA) in the context of CFA; and
- Establishing the final group-invariant model and calculating weighted scales.

Step 1. Exploring the item-factor relationship

As the first step, the hypothesized factor model was developed based on both theory-driven and data-driven analyses to explore the item-factor relationship. The theory-driven analysis mapped survey items to program implementation and teacher outcome features. At the same time, we

[1] One factor, Knowledge about OTIF, was not developed using factor analysis. Rather, it is an index of the percentage of correct answers.

examined the relationships among items using a series of correlations and the exploratory factor analysis (EFA) technique.

Within each model, items that were not significantly correlated with the factors were dropped from the model. In addition, the EFA results informed how items should be organized within a factor. This step produced four factors that portray different aspects of implementation and outcomes from the viewpoint of the teachers.

- Attitude about factors important to determining supplemental pay. This factor examines how important various factors viewed by teachers in determining a teacher's supplemental pay. It comprises 10 items from the teacher survey.

- Support of OTIF. This factor addresses the extent of teacher's opinions toward the program in general. It is formed by 11 survey items.

- Collaboration. This factor covers the questions about the collaboration and support among teachers. It includes six items.

- Perceived changes resulting from OTIF. This factor assesses the extent of changes teachers indicated as the result of the program implementation and covers multiple components of the changes—the positive changes among teachers, the negative changes among teachers, and the changes in working with students. This factor consists of nine items.

Step 2. Confirming the link between the mapping and the data

As the second step, we examined the overall link between the mapping and the data, using the CFA technique. Based on the analysis, the factor loadings for all items showed substantial and statistically significant relation to the intended factors. This overall model had $X^2_{(54)} = 121.409$ with GFI of 0.949, CFI of 0.978, and RMSEA of 0.05.[2] Based on these fit statistics, we concluded the overall model fits the data well.

[2] GFI, goodness-of-fit index, is one of the absolute fit indexes. This estimates the proportion of variability in the sample covariance matrix explained by the model. It is analogous to R^2 as a multiple regression. The rule of thumb is GFI>0.9 indicates good fit, and values close to zero indicate very poor fit. CFI, comparative fit index, is one of the incremental fit indexes. This index assesses the relative improvement in fit of empirical model compared with the null (or theoretical) model. The rule of thumb is that values greater than 0.90 indicate reasonably good fit. RMSEA, root mean square error of approximation, is one of the patrimony-adjusted indexes. RMSEA estimates the amount of error of approximation per model degree of freedom and takes sample size into account. The rule of thumb is that RMSEA≤0.05 indicates close approximate fit; values between 0.05 and 0.08 suggest reasonable error of approximation and ≥0.10 suggests poor fit (Kline, 2005).

Step 3. Ensuring the factors are measuring the same traits between the groups

For the third step, we conducted MSA to ensure that the factors were measuring the same traits between the two groups of districts that implemented performance-based plans that use the different compensation models (i.e., TAP and non-TAP models). In addition, we ensured those factors measure the same traits across years so that we were able to compare and to produce meaningful results (Byrne, 1991; 1998). We tested mainly whether 1) the factor structures (i.e., number of factors) are group invariant, and 2) the patterns of factor loadings are group invariant.

We examined the factor that measures the aspects that impact the implementation for the two groups separately prior to testing the group invariance.[3] Overall fit of the model for the TAP districts was $X^2_{(54)} = 76.892$, with GFI of 0.93, CFI of 0.99, and RMSEA of 0.05; for the non-TAP districts, it was $X^2_{(54)} = 149.15$, with GFI of 0.87, CFI of 0.94, and RMSEA of 0.10. Based on these calculations, we found that the baseline models fit somewhat better among TAP districts than non-TAP districts, although we did not know whether the model was significantly different or invariant between two groups at this point.

The basic idea of MSA analysis to test the group invariance involves first specifying a model in which certain parameters (i.e., factor patterns) would constrain to be equal across groups, and then comparing that model with a less restrictive model in which these parameters would be free to take any value. Examining the overall fit of the first model would indicate whether the factor structures are group invariant. Comparison of an X^2 difference between the two models provides a basis for determining whether the patterns of factor loadings are group invariant. A significant X^2 difference would indicate the non-invariance between the groups (Byrne, 1991; 1998).

According to our analysis, the overall fit of the first model in which all the factor structures were constrained was $X^2_{(145)} = 327.509$, with CFI of 0.94 and RMSEA of 0.08. The factor loadings for all items were substantial and statistically significant. We concluded that our hypothesized four-factor model to measure the factors pertaining to implementation and outcomes fit well for both groups.

Overall fit of a second model in which all the factor structures were free of constraints between the groups was $X^2_{(131)} = 278.844$, with CFI of 0.95 and RMSEA of 0.07. The factor loadings for all items were substantial and statistically significant. The X^2 difference, $\Delta X^2_{(14)}$, is 48.665, which indicates a significant difference between two models. These findings indicate that although the factor

[3] This examination was applied for each group separately. This is not required step, but it is customary (Byrne, 1998; Jöreskog and Sörbom, 1996).

structures are group invariant, the patterns of factor loadings are different between groups and some items are functioning differently.

Step 4. Establishing the final group-invariant model and calculating weighted scales

The above findings led to the final step, testing equality of constraints to establish the group-invariant model. There are a few ways to test the equality of constraints. One is to evaluate the modification indices (MI) of the constraint model. Since the MI provides an approximate amount of X^2 decrease when a particular constraint is released, one can use the MI to identify the constraint(s) that has the large MI values and make re-parameterization of the model (Jöreskog and Sörbom, 1996). Although this way is the simplest, the actual amount of X^2 change can be larger than the predicted amount. Another way is to release all constraints sequentially, each time assessing the statistical significance of the X^2 change in fit (Byrne, 1991). In our study, we used both methods.

The examination of the MI did not suggest a need to release any items. After re-specifying the group-invariant model, similar steps were taken to test the year invariance model. We calculated individual teacher scores by multiplying the item response with the factor loading. Therefore, the scales were weighted by the proportion of items the scale contributed to the factor. For presentation, we standardized the scale scores on a range of 0 to 100, with 100 representing the highest possible score.

References

Byrne, B.M. (1991). The Maslach burnout inventory: Validating factorial structure and invariance across intermediate, secondary, and university educators. *Multivariate Behavioral Research* 26(4): 583-605.

Byrne, B.M. (1998). *Structural Equation Modeling With LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kline, R.B. (2005). *Principles and Practice of Structural Equation Modeling.* 2nd Ed. New York: The Guilford Press.

Jöreskog, K., and Sörbom, D. (1996). *LISREL 8: User's Reference Guide.* Chicago: Scientific Software International, Inc.