# Are Principals Good at Identifying Effective Teachers?
# A Comparison of Teachers' Principal Ratings
# and Residual Gain on Standardized Tests.

James J. Gray

Dissertation submitted to the graduate degree program in Educational Leadership & Policy Studies and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Education (Ed.D).

Dissertation Committee:

Chair: ______________________________
Dr. Argun Saatcioglu

Committee Members: ______________________________
Dr. George Crawford

______________________________
Dr. Howard Ebmeier

______________________________
Dr. Michael Imber

______________________________
Dr. Marc Mahlios

Date Defended: April 9, 2010

The Dissertation Committee for James J. Gray certifies that this is the approved version of the following dissertation:

Are Principals Good at Identifying Effective Teachers?

A Comparison of Teachers’ Principal Ratings and Residual Gain on Standardized Tests.

________________________________

Dr. Argun Saatcioglu

Date Approved: ______________________

Table of Contents

# List of Tables

# Abstract

The purpose of this study was to answer the question: Are principals good at identifying effective teachers? Some studies have suggested they are not, but the evidence is not consistent. It is troubling that research results are inconsistent regarding principals' abilities to identify effective teachers. Why is there a disconnect between principals’ evaluations of teachers and student gain scores - the operational definition of effective teachers? One would think principals would be good at identifying effective teachers, given that the expectation for hiring, developing, and evaluating teachers is a major facet of their responsibilities. This inconsistency suggests there could be a methodological issue. In other words, the method used for determining a principal’s ability to identify effective teachers may be leading to these mixed findings. Perhaps the metric commonly used to measure teacher effectiveness is incompatible with identifying effective teachers. Could using a newer type of standardized test as the metric along with a more focused method of data analysis lead to consistent positive correlations between principal ratings and teacher effectiveness?

This study examined the relation between principals’ identification of effective teachers and the student gain scores from Fall and Spring Measures of Academic Progress (MAP) - computer-adaptive tests for reading, mathematics, and language usage developed by Northwest Evaluation Association (NWEA, 2006).

This study fitted individual level data to a value-added model to estimate teacher effects on students’ 06-07 MAP gain scores and then ran subsequent regression analyses to estimate principals' ranking effects on teachers’ average Spring 07 MAP scores, on teachers’ average 06-07 MAP gain scores, and on teachers’ value-added effects on students’ 06-07 MAP gain scores.

The findings showed that principals can identify their effective math teachers but they can’t identify their effective communications and English teachers. Principals’ rankings of teachers tend to correlate more with math teachers than communications and English teachers regarding student gain scores and teachers’ value-added to student gain score

## Acknowledgements

I have read that digital records are the nearest thing to eternity we have on earth. It is fitting that the support I have received from others would be immortalized.

Thank you Dr. Argun Saatcioglu for being my advisor and guiding me through these last two years of my ten-year journey. I could not have completed my dissertation without your help. Thank you Dr. George Crawford, Dr. Howard Ebmeier, Dr. Michael Imber, Dr. Marc Mahlios, and Dr. Saatcioglu for your enlightening and thoughtful discussions during my defense. Thank you also for making your classes stimulating as I truly enjoyed and looked forward to all my classes. I believe a day is wasted if I don't learn something new and useful. There were no wasted days at Kansas University. Dr. Susan Twombly, Jan Kazar, Denise Brubaker, and Mary Ann Williams thank you for your help.

Thank you to my Shawnee Mission School District family for their support: Dr. Gene Johnson, Dr. Gillian Chapman, Dr. Joy Torgerson, Dan Gruman, Margie Prall, Deborah Pfortmiller, Carla Allen, Dr. Kevin Peters, Susan Ostermeyer, Jim Wink, Dr. Larry King, Dr. Chuck McLean, Debbie Ryan, Dr. Julia Crain, Connie Springfield, Keith Burgat, Erica Warren, Dick Kramer, Bill Taft, Don Perkins, Alice Capson, Helaine Cohen, Peggy Augustine, Chris Wiley, and many others who with their kindness supported my cause. Also, thank you Don Barta for your friendship through these years.

Special thanks to my loving wife, Karen, who endured my reclusive doctoral pursuit. Thank you to my daughters, Samantha and Jamey, and to my son, Dustin, for their continued encouragement. Thank you to my extended family, David, Dean, and Andrea, and my grandchildren, Phoenix, Katie, Breckan, and Josephine for their support. My thoughts are with you, Hattie. Thank you to my brothers and sisters, Richard, Mary Ann, Donald, Jeanne, Arthur, Larry, and Willie and their families.

I am truly blessed to be thankful.

# Chapter 1

# Introduction

## 1.1

## Can Principals Identify Effective Teachers?

Are principals good at identifying effective teachers? Some studies have suggested they are not, but the evidence is not consistent (Jacob, 2006, pp. 60-62; Milanowski, 2004, pp. 49-50; Wilkerson, 2000, pp. 185-187; Medley, 1984, pp. 46-47).

Teacher behaviors are typically used by principals when rating teachers for their effectiveness with all students. One particular study using student gain scores raised some serious questions concerning the relationship of teacher behavior and teacher effectiveness. In 1977, Donald Medley published a correlation study from 14 different teacher effectiveness studies comparing teacher behaviors – what they do – with student gain scores on standardized tests for students disaggregated by low and high socio-economic status (SES). The correlations were often mixed with conflicting positive and negative correlations within most behaviors.  Of the forty teacher behaviors compared, only six had mostly positive correlations for both low and high SES students (Medley, 1977, pp. 23 – 65). This is disconcerting as it suggests principals are not using the right criteria for identifying effective teachers.

Several other studies since then have come to contrasting conclusions about the value of using principals' behavioral ratings of teachers to predict teacher effectiveness. Some studies have concluded there is no relationship whatsoever between the ratings and student learning; others have found some moderate relationship.

In one study, teachers were rated according to a list of competencies selected by teachers from a larger list created by teacher educators believed to be necessary for successful teaching. Scores from these ratings were correlated with student gains. Of the few statistically significant correlations, about half were negative and half were positive. Behaviors considered by teachers and believed by the experts who created the original lists to indicate effective teaching were about as likely to indicate ineffective teaching as effective teaching (Medley, 1984, pp. 46-47).

In another study with negative implications (Wilkerson, 2000), teacher performance was rated by an extended group of participants in the education process, including principals, students, and teachers themselves. David Wilkerson, et al. found principal ratings of teachers were “most disappointing” at predicting student achievement in all assessed areas, failing to meet the predetermined rejection level of 0.05 for predicting student achievement (Wilkerson, 2000, pp. 185-187).

However, in a Cincinnati Public Schools study (2001-2002), Anthony Milanowski found a moderate degree of criterion-related validity between teacher performance evaluation scores and student achievement. This conclusion was made even though the relatively small correlations within the 0.3 and 0.4 range indicated only a small proportion (9% to 16%) of variance in student achievement was potentially due to variation in teacher performance (Milanowski, 2004, pp. 49-50).

Another study (Jacob, 2003) found some correlations between principal ratings and teachers’ effectiveness for the best and worst teachers. Principals identified their best teachers 52 percent of the time in reading and 69 percent of the time in mathematics. There were similar results for identifying the worst teachers. However, the principals were significantly less successful when ranking categorically teachers in the middle 60-80 percent of the ability distribution. Even so, principals’ ratings of all teachers were generally quite high with an average of 8.1 on a 10-point scale, which seems incongruous with the findings (Jacob, 2006, pp. 60-62).

In another study of the relationship between teacher evaluation and student achievement the assumption that assessments of teaching behaviors will reflect measures of student achievement was explored. The results were mixed with respect to the question whether teachers’ evaluation scores were related to the average achievement of those teachers’ students. The estimated relationship of teacher evaluation scores to student achievement was positive but not statistically significant in almost all cases (Kimball, 2004, p 70).

It is troubling that research results are inconsistent regarding principals' abilities to identify effective teachers. Why is there a disconnect between principals’ evaluations of teachers and student gain scores - the operational definition of effective teachers? One would think principals would be good at identifying effective teachers, given that the expectation for hiring, developing, and evaluating teachers is a major facet of their responsibilities. This inconsistency suggests there could be a methodological issue. In other words, the method used for determining a principal’s ability to identify effective teachers may be leading to these mixed findings. Perhaps the metric commonly used to measure teacher effectiveness is incompatible with identifying effective teachers. Could using a newer type of standardized test as the metric along with a more focused method of data analysis lead to consistent positive correlations between principal ratings and teacher effectiveness?

### 1.2 Research Question

This study examines the relation between principals' identification of effective teachers and the student gain scores from Fall and Spring Measures of Academic Progress (MAP) - computer-adaptive tests for reading, mathematics, and language usage developed by Northwest Evaluation Association (NWEA, 2006).

Are principals good at identifying effective teachers?  is the focal question of this study. Specifically, can principals pick up on teachers' value-added performance as opposed to their performance in terms of their students' raw average scores? For example, a teacher's principal may think the teacher is effective because the teacher's students do well, but the teacher may not be adding much value to their performance because the student may already be good at the beginning. Alternatively, a teacher's students may be doing poorly compared to other students but the teacher may be adding much value to student performance because the students were behind in achievement in the first place. Can a principal pick up on such patterns? (Saatcioglu, 2010).

# Chapter 2

# Literature Review

## 2.1 Why MAP?

MAP is a newer type of standardized test used as a metric, in contrast to previous studies using traditional standardized tests when determining gain scores. Each student’s growth is listed on a continuous scale representing the instructional level of the student. This Rausch Item (RIT) scale is an equal-interval growth scale that accurately measures a student's academic growth based on pre- and post-test results (Olson, 2004, p.4). Using students’ RIT scores permits a comparison of individual student progress against grade level norms of large populations based on high-quality items, individualized forms, and accurate large population comparisons (Peterson, 2004, p. 66).

Standardized tests typically measure student achievement in important but low-level objectives of skills. This is because the low-level objectives within goals such as computational skill, knowledge of principles and facts, and the ability to recognize points in a brief passage can be measured more readily with paper-and-pencil tests. The kind of teaching needed for maximizing  achievement  for such simple  objectives differs from the kind of teaching necessary for producing  achievement in more complex,  high-level objectives - the ones typically not measured by these tests. Medley (1984) has suggested that since the available achievement tests do not measure some of the most important outcomes of teaching, poorer teachers often do as well as or better at teaching low-level objectives than the better teachers.

Although MAP may also test low-level objectives, the computer-adaptive testing process potentially provides a means for a more accurate assessment. As a student takes the test, the computer adjusts the difficulty of the questions so that each student takes a unique test. If a question is answered correctly, a more difficult item is displayed. Conversely, a less difficult item is displayed when a question is answered incorrectly. As the items are selected within the test, the estimate of achievement becomes more precise. This iterative process is repeated until the test is completed (Cronin, 2005, p.18).

Another weakness of most standardized tests is the lack of intact cohort student groups from one teacher's tested student group to the next year teacher's tested student group. Typically, standardized tests are given once a year with shifting membership within classes from year-to-year. MAP addresses this weakness with a pre-test and post-test within the same school year. Students are given the MAP at the beginning of the school year as a pre-test and again at the end of the school year as a post-test. Within these cohort groups, individual student gain scores and subsequent teacher effectiveness can be determined.

## 2.2 Principals and Teacher Evaluation

Today, principals most commonly use a summative evaluation process. Following steps required by law, principal-conducted evaluations usually involve a pre-observation conference, a checklist-guided observation, and a post-observation conference culminating in a signed and filed official form (Ponticell, 2004, p.52). This process is of questionable value for many reasons. The sample size is too small to be of any use, the sample is not random, the judge is not free of significant social bias from non-classroom relationships with the teacher and the checklists are unsubstantiated (Stufflebeam, 2000, p. 263). Indeed, there is evidence that checklist reporting systems lack accuracy and are influenced by discrimination (Peterson, 2004, p. 71).

Principals also might use classroom walk-throughs for teacher evaluations, brief and unscheduled informal classroom visits lasting three to six minutes. The principal looks for evidence of student learning and implementation of staff development themes. Walk-throughs can be more valuable data sources than formal observations because they sample more reliably with a greater number of observations, are more flexible in focusing on what makes a difference in student learning, and are less intrusive on actual ongoing instruction (Keller, 1998, as cited in Peterson, 2004, p. 62). However, walk-throughs are not intended to be full-fledged observations for individual teacher‘s evaluations. Yet, pieced together from multiple classrooms, walk-throughs provide a good overview of a school’s progress towards its goals rather than individual teacher’s progress (Berube, 2006, p.14).

Principals may also use student achievement data to evaluate teachers. Although student achievement data is the single most compelling single indicator of teacher quality, most teacher evaluations use administrator estimates of student achievement and not the best objective evidence available.  In fact, the most widely used evaluation systems do not feature direct information on student learning – student achievement data – but instead rely on principal reports of teacher performance. Ideally, student gain scores should play an important role in teacher evaluation but in reality defensible student achievement data are difficult to obtain on all teachers (Peterson, 2004, pp. 64-65).

Portfolios, evidence compiled by teachers showing their professional growth, are also sometimes used by principals in evaluations. The most significant perceived advantage of portfolio-based assessment is the reflection piece. The opportunity to reflect often leads to more teacher collaboration and sharing, and encourages changes in teaching practices. Usually teachers report the portfolio process as a richer, more in depth picture of their performance than the typical “snapshot” observation. But both principals and teachers report the time-consuming portfolio process is “one more thing” to do in their busy schedules (Attinello, 2006, p 146).

Other evaluative tools principals use may include student and parent surveys, peer review of materials, documentation of professional activity, teacher test scores, National Board Certification, documented benefits from action research, and school improvement participation. But these sources aren’t easily applied in an appropriate manner. It is critical that procedures for collection and interpretation of data are well designed and conducted correctly. Flexibility is necessary because no single one is valid for every teacher and no individual source is available for each teacher (Peterson, 2004, pp. 63-64).

In their evaluations, what characteristics do principals use to inform them about effective teachers? Research has identified characteristics likely to be observed in classrooms of highly effective teachers. These include: time on task is high and focused on academic content; learning goals are clear; instruction encourages active learning; individual differences between students are acknowledged and accommodated; skills-based instruction is balanced with higher-level instruction; skills are taught in context; and the classroom climate is supportive and collaborative (Prothero, 2002, p. 49).

## 2.3 Effective Teachers and Student Achievement Data

Educational researchers typically define effective teaching by students' residual gain on standardized tests. Even though test scores do not capture all facets of student learning, test scores are widely available, objective and are recognized as important indicators of achievement by educators, policymakers, and the public (Rockoff, 2004).

However, defining effective teaching in terms of student achievement is not the only way effective teaching can be defined. There are other research-based definitions of teacher effectiveness. For example, some research define effective teaching as the display of direct instruction behaviors while other research define effective teaching as the embodiment of constructivist teaching and learning. In each case, effectiveness is a function of how effectiveness is originally defined. Consequently, the indicators for teaching effectiveness are directly linked to how researchers define effectiveness and these indicators are often subjective and not independent from each other (Sergiovanni, 2001).

Effective teaching may be defined in terms of process-product teaching or maybe in the somewhat different terms of subject-matter teaching. Process-product teaching identifies teaching behaviors and patterns of teacher-student interactions related with student achievement gains. Teachers made a difference by exposing students to academic content and the opportunity to learn and spending a great deal of time instructing their students. On the other hand, subject-matter teaching or teaching for understanding identifies teachers as stimulators of student learning. Teachers make a difference by inducing conceptual change through in depth study of fewer topics and creating a learning community where dialogue promotes understanding (Brophy, 1992).

Obviously, each of these models of teaching by definition would have different indicators for effective teaching adding further to the complexity of identifying effective teachers. When one adds to the mix still other models of teaching ,such as; teaching for authentic learning, multiple forms of learning, or learning as a social act, the multiple indicators of effective teaching become even more varied (Leinhardt, 1992). What then would be a common indicator for determining effective teaching?

For better or for worse, student test scores have increasingly become a commonly accepted standard for teacher effectiveness. Therefore, this study will focus on the definition of effective teaching in terms of student achievement. There exists a clear and undeniable link between teacher effectiveness and student achievement. Effective teachers foster achievement gains beyond that expected from students' past achievements. This is confirmed through the overall finding from value-added studies that effective teachers are essential for student success (Stronge, 2008, p 170). Subsequently, principals' being able to identify these effective teachers is essential for student success.

Students' residual gain on standardized tests as a metric for teacher effectiveness is used within widely-cited value-added studies such as Tennessee Value-Added Assessment System study, TVAAS (Sanders, 1996, p. 1), Dallas Public Schools (Mendro, 1998), Project STAR studies (Nye, 2004, p. 244), The Hamilton Project (Gordon, 2006, p.6), Virginia Study (Stronge, 2008), and the Washoe County Study (Kimball, 2004).

TVAAS Study (1991 – 1995):

The Tennessee Value-Added Assessment System (TVAAS) was designed to determine individual teacher's influence on the rate of academic growth for student populations. This system is contingent on three key components: a testing process with scales that are strongly curriculum aligned and produce measurement that extends above and below grade level; an ongoing expanding longitudinal data base; and a statistical process that produces unbiased and efficient estimates of variable effects through a multivariate, longitudinal analysis.

During this study (1991-1995), the TVAAS database included nearly three million records for Tennessee's entire grade two through eight student population providing individual student's TCAP achievement scores in mathematics, reading, language arts, science, and social studies from 1990 through 1996. Using this data, the TVAAS study followed a sample cohort group of second grade students through fifth grade using mathematics achievement test scores from two of Tennessee's larger metropolitan school systems. The purpose of this study was to determine value-added effects of teacher sequence over students' three-year movement from teacher to teacher.

The study's methodology basically followed a two-stage model:

Stage One: Value-Added Teacher Effects

Shrinkage estimates of teacher effects were estimated from a longitudinal analysis fitting the data to the following value-added mixed model:

Current score = a + b*(previous math score) + t(i) + error

Where
a = constant to be estimated from the data
b = regression coefficient
t(i) = shrinkage estimate of the teacher effects

The teacher effects from the above regression analysis determined the arbitrary placement of teachers within quintile levels of effectiveness. Teachers demonstrated the lowest degree of effectiveness in the first quintile and the greatest degree of effectiveness in the fifth quintile. This process was repeated independently for grades three, four, and five. By encoding individual student records with the teacher effectiveness quintiles, the student progress was traceable through identified sequences of teacher effectiveness.

Stage Two: Quintile-Level Teacher Effects

Each cohort group's analysis spanned three years of student TCAP achievement scores. The effects of quintile levels of teachers from previous grades on students' current year scores was determined by:

Fifth grade score = a + b*(second grade score) + tq3(i) + tq4(i) + tq5(i) + error

Where

a = constant to be estimated from the data
b = regression coefficient
tq3(i) = Quintile level of the third grade teacher
tq4(i) = Quintile level of the fourth grade teacher
tq5(i) = Quintile level of the fifth grade teacher

Second grade scores were included so that the subsequent teacher quintiles would not be biased due to any disproportionate assignment of students to various teacher sequences. The results for both school systems were highly significant for all three grade level transitions giving credence to the effects of prior teachers' quintile level and the individual student's three-year sequence of teachers' quintile level on student achievement. As an extreme example, the average difference of the student

TCAP percentile level between a low-low-low year-to-year sequence of teacher levels and a high-high-high sequence in grade five was 52 to 54 percentile points. Other comparisons such as low-low-high teacher and average-average-high sequences of the 125 possible combinations of teacher level sequences were analyzed and demonstrated that teacher effects were cumulative and additive with very little compensatory effects. An effective teacher receiving students from relatively ineffective teachers from prior years can facilitate academic gain during the school year; however, the residual effects of relatively ineffective teachers do continue through subsequent student achievement scores. The findings of this study confirm that the year-to-year teacher sequence can have a potentially dramatic effect on student achievement scores either for the better or for the worse and the learning lost through ineffective teaching cannot be fully recovered (Sanders, 1996). For this reason, it is very important for principals to be able to identify their effective and ineffective teachers.

TVAAS Study (1995-1996):

The TVAAS Study (1995-1996) used a subset of the TVAAS data from the 1995 and 1996 TCAP scores for five subjects (math total, reading total, language total, social studies, and science) and three grades (third, fourth, and fifth). Each of the fifteen subject-grade combinations was analyzed separately for two different sets of Tennessee school systems. One set consisted of thirty East Tennessee school systems and the other of twenty-four Middle Tennessee systems. Using these data, this study attempted to measure the magnitude of teacher effects while considering the influences of intra-classroom heterogeneity, student achievement level, and class size on academic growth. Among these influences, intra-classroom heterogeneity was of special interest due to the prevailing practice of ability grouping classrooms.

The data were fitted to the following value-added mixed model:

$$Y = M + S + H + C + H*C + T(S*H*C) + A + A*S + A*H + A*C + A*H*C + A*T(S*H*C)+ E$$

Where

Y = Student’s gain score
M = Overall mean gain
S = School system
H = Heterogeneity-In-Achievement (3 groups used)
C = Class size (2 groups used)
H*C = Heterogeneity-by-class-size interaction
T(S*H*C) = Teacher – each one nested within a particular combination of system, heterogeneity groups, and class size group
A = Achievement level (4 groups used)
A*S = Achievement-by-system interaction
A*H = Achievement-by-heterogeneity interaction
A*C = Achievement-by-class-size interaction
A*H*C = Achievement-by-heterogeneity-by-class-size interaction
A*T(S*H*C) = Achievement-by-teacher interaction
E = Random error term

When the z-values for the above variables were compared by grade level it clearly showed that the two most important factors affecting student gain were teacher effects and the achievement level of the student. Teacher effect size was highly significant in every analysis and had a larger effect size than any other factor in twenty of the thirty analyses. Student academic level was significantly related to academic progress although not to the degree of teachers’ effects. Interestingly, class size was not significant in most cases with only three of thirty analyses being significant. The effect of class heterogeneity seemed to be minor relative to student achievement. Teachers do make a difference in student achievement regardless of the  homogeneity and heterogeneity of student ability levels within their classes. Disconcerting is the evidence that low-achieving students were more likely to be placed with less effective teachers. Again, there is a clear link of teacher effectiveness and student achievement using student achievement data (Wright, 1997). And, principals need to be able to identify their effective teachers.

Dallas Public Schools:

Dallas Public Schools use norm-referenced tests for the initial identification of potentially outstanding teachers and schools and potentially ineffective teachers and schools. However, using student achievement data as part of teacher and school accountability has been an ongoing debate because of the potential for bias. This debate centers on the need for multiple outcome variables and a need for controlling for exogenous influences.  Clearly, outcome variables need to be related to important educational goals and resources should be allocated to maintain the extensive databases and take the multiple measures required to measure student achievement.

The Dallas Public Schools attempt to avoid this potential bias through the Accountability Task Force. The teachers and administrators on this task force jointly decide what variables are to be used to determine teacher and school effectiveness and also assign weighted values to each of these variables based on the variable’s relative value to the student achievement.

Dallas Public Schools’ research clearly show that for students with average prior achievement levels, groups of students can lose as much as twenty percentile points in a year. Over three or four years, students with ineffective teachers can be fifty percentile points lower than students with effective teachers. Therefore, identification of effective and ineffective teachers is critical for student success. Norm-referenced tests have been found to be sufficient for the initial identification of potentially outstanding teachers and schools and potentially ineffective teachers and schools (Mendro, 1998).

Project Star:

The Tennessee Class Size Experiments or Project STAR study followed a randomly assigned cohort group of kindergarten students through their third grade using reading and mathematics achievement data to determine variations in teacher effectiveness.  The study involved 79 elementary schools in 42 Tennessee school districts using three different treatment conditions: small classes (13 -17 students), larger classes (22 – 26 students), or larger class with a full-time aide. Teachers were also assigned randomly to the different types of classes. These class assignments were maintained through the third grade. Since the classes were initially equivalent by random assignment, any differences in student achievement among classes would be due to either the treatment condition or differences in teacher effectiveness. Therefore within a school, any systematic variance in achievement between classrooms with the same treatment would be due to teacher effectiveness.

Stanford Achievement Test (SAT) reading and mathematics test scores for kindergarten through third grade were used to measure student achievement. The analyses used a three-level hierarchical linear model (HLM) where level one examined teacher effects on achievement gains and then examined teacher effects on achievement status. The level-two model examined the variation of coefficients between classes within schools. And the level-three model examined the intercept for the level-two coefficients of the kth school.

Level One Model: Teacher effects on achievement gains

$$Y_{ijk} = B_{0jk} + B_{1jk}\, PRETEST_{ijk} + B_{2jk}\, FEMALE_{ijk} + B_{3jk}\, SES_{ijk} + B_{4jk}\, MINORITY_{ijk} + \varepsilon_{ijk}$$

Where:

$PRETEST_{ijk}$ = Achievement test in previous year for $Y_{ijk}$ (Gain Score)

$FEMALE_{ijk}$ = Gender dummy variable

$SES_{ijk}$ = Free or reduced lunch dummy variable

$MINORITY_{ijk}$ = Minority dummy variable (Black, Hispanic, or Asian)

$\varepsilon_{ijk}$ = Student-specific variable

Teacher effects on achievement status:

$$Y_{ijk} = B_{0jk} + B_{2jk}\, FEMALE_{ijk} + B_{3jk}\, SES_{ijk} + B_{4jk}\, MINORITY_{ijk} + \varepsilon_{ijk}$$

Where:

$FEMALE_{ijk}$ = Gender dummy variable

$SES_{ijk}$ = Free or reduced lunch dummy variable

$MINORITY_{ijk}$ = Minority dummy variable (Black, Hispanic, or Asian)

$\varepsilon_{ijk}$ = Student-specific variable

Level-Two Model:

$$B_{0jk} = \pi_{00k} + \pi_{01k}\, SMALL_{jk} + \pi_{02k}\, AIDE_{jk} + \xi_{0jk}$$

Where:

$B_{0jk}$ = Intercept for level-one model for jth class of the kth school

$\pi_{00k}$ = the school-specific intercept for school k

$SMALL_{jk}$ = Indicator for small class size

$\pi_{01k}$ = School-specific slope for SMALL in school k

$AIDE_{jk}$ = Indicator for having a full time classroom aide - regular sized classes

$\pi_{02k}$ = School-specific slope for AIDE in school k

$\xi_{0jk}$ = Classroom-specific random effect

The variance of $\xi_{0jk}$ described the variance of the average achievement gains across classes due to the effects of student gender, SES, minority group status, and treatment assignment. All other coefficients were constrained to be constant within schools,

$B_{1jk}$ = $\pi_{10k}$ , $B_{2jk}$ = $\pi_{20k}$, $B_{3jk}$ = $\pi_{30k}$, and $B_{4jk}$ = $\pi_{40k}$

Level-Three Model:

Variation across schools of each of the school-specific regression coefficients are modeled as random and free to vary. The level-three model for the intercept at level-two coefficient of the kth' school is therefore:

$\pi_{00k}$ = $\gamma_{000}$ + $n_{00k}$

$\pi_{01k}$ = $\gamma_{010}$

$\pi_{02k}$ = $\gamma_{020}$ where m = 0, …, 2, the $\gamma_{0m0}$ are fixed effects and $n_{00k}$ is a school-specific random effect. Similarly, the level-three models for the other level-two coefficients are:

$\pi_{10k}$ = $\gamma_{100}$

$\pi_{20k}$ = $\gamma_{200}$

$\pi_{30k}$ = $\gamma_{300}$

$\pi_{40k}$ = $\gamma_{400}$ where $\gamma_{m00}$ are fixed effects and m = 1, …,4.

The results of this study indicated that teacher effects are real and are consistent in magnitude of estimates as shown by previous studies. There are substantial differences among teachers in their ability to produce achievement gains in their students. The difference in having a not so effective 25th percentile teacher and an effective 75th percentile teacher is over one third of a standard deviation (0.35) in reading and almost half a standard deviation (0.48) in mathematics. Similar differences were evident between an average 50th percentile teacher and a very effective 90th percentile teacher: 0.33 in reading and 0.46 in mathematics.

It would be tempting to consider the intervention of replacing a teacher estimated to be at the 25th percentile with a teacher estimated to be at the 75th percentile. However, this study's calculations probably overstate the effect of such an intervention since the estimates are based on the potential effects of interventions if a perfect predictor of teacher effectiveness was available. There is no such perfect predictor. Even the direct empirical estimates of teacher effects regressed from value-added models would have substantial statistical estimation error and therefore be imperfectly correlated with true teacher effectiveness. Nonetheless, as with other studies, these differences suggest that interventions to improve teacher effectiveness and predict the effectiveness of teachers through student achievement would be promising strategies for improving student achievement (Nye, 2004). This points out again the importance of principals being able to identify effective teachers.

Hamilton Project:

The Hamilton Project, a study of teachers' impact on student achievement, used mathematics, reading, and language arts scores from the Stanford 9 test for 2000 through 2002 and scores from California Achievement Test for 2003 as metrics for identifying effective teachers. A value-added model (VAM) was used to determine a teacher's effect on student average gain in performance:

$$
\begin{aligned}
S_{it} = {} & \beta_{1gr,yr}\, Math_{it-1} + \beta_{2gr,yr}\, \mathrm{Re}ad_{it-1} + \beta_{3gr,yr}\, LangArt_{it-1} + \\
& \lambda_{1gr,yr}\, Race/Eth_{i} + \lambda_{2gr,yr}\, ELD_{it} + \lambda_{3gr,yr}\, FreeLnch + \lambda_{4gr,yr}\, Male_{i} + \\
& \lambda_{5gr,yr}\, GATE_{it} + \lambda_{6gr,yr}\, \mathrm{Re}\,peat_{it} + \delta_{teacher,rear} + \varepsilon_{it}
\end{aligned}
$$

Where:

$S_{it}$ = Math score for person i in year t
For 2000 – 2002, the math score is used from Stanford 9 test.
For 2003, the math score is used from the California Achievement Test (CAT).
$Race/Eth_{i}$ = Vector of six racial/ethnic categories
$ELD_{it}$ = Vector of five categories of English language development level
$\mathrm{Re}\,peat_{it}$ = Dummy indicating whether a person is currently repeating a grade
Previous spring math, reading, and language arts scores are also included.

The study used the academic achievement data of about 150,000 Los Angeles Unified School District students from 9,400 classrooms to examine the effectiveness of their teachers and to determine how long it would take to make a reliable distinction between more and less effective teachers. To test how well a district could predict teacher effectiveness, this study focused on teachers who were in their first, second, and third year of teaching during 2000 through 2003. Their students' achievement was measured during each of the three years, controlling for students' previous test scores and demographics.  Based on their estimated impact on their students' achievement during their first two years of teaching, teachers were ranked by quartiles. These quartile teacher rankings provided a lot of information about a teacher's impact. By the third year, average students assigned to a bottom quartile teacher during the teacher's first two years lost an average of five percentile points relative to similar students with similar baseline scores and demographics. In contrast, average students assigned to a top quartile teacher gained five percentile points with an average third year difference of ten percentile points between being assigned a top-quartile or a bottom-quartile teacher.

Value-added or the average gain in performance for each teacher's students was determined to be such a significant component of measuring teacher effectiveness that the Hamilton Project recommended funding for federal grants to help states link student performance with the effectiveness of individual teachers. It was also recommended that the bottom-quartile of first- year teachers should be non-renewed each year so that over a three-year period the lower 10% of all teachers would be replaced with better performing teachers (Gordon, 2006) – pointing again to the importance of principals being able to determine their teachers' effectiveness.

Virginia Study:

The Virginia Study used data for 1936 third grade students and 85 teachers in a moderately sized Virginia urban school district with about 23,000 total students. The study examined the relationship between teacher effectiveness and student achievement. Effective teachers were defined as teachers who foster achievement gains beyond that expected from the student's past achievement. One-stage Ordinary Least Squares, OLS, two-stage OLS, and two-stage, two-level HLM econometric models were fitted to the data in this study. It was found that Two-stage OLS regression models provided an adequate fit. Literature reviewed had also recommended the use of two-level HLM regression models but also found OLS solutions to be highly correlated and relatively free of bias.

The data were fitted to an OLS model to estimate the achievement expectations for each student. Then, actual achievement was compared to the expected achievement estimates. Positive differences indicated student achievement beyond expectation, zero differences indicated achievement commensurate with expectation, and negative differences indicated achievement below expectation. The difference scores for students were standardized, aggregated, and averaged to make a composite for each teacher. Analysis of the distribution of these teacher composites identified the least and most effective teachers within the bottom and top quartiles respectively.

Since ineffective and effective teacher samples were small, it was decided to do a set of case studies using the qualitative approach of exploratory cross-case analysis. Within this analysis, teachers' classroom characteristics were observed for both effective and ineffective teachers and compared for five effectiveness categories and a total of twenty descriptors within the categories. In all categories, teachers identified through the OLS regression model as highly effective, top-quartile, teachers were superior in their observed classroom characteristics to those teachers identified as least effective, bottom-quartile teachers.

Based on this clear link between teacher effectiveness and student learning, this study recommended using student achievement data fitted to regression models for predicting effective teachers. Seemingly at odds within a teacher evaluation system, both accountability and professional growth can both be addressed by examining teacher effects on student achievement and by identifying effective teachers as well as ineffective teachers. Subsequently, the classroom characteristics and behaviors of teachers with higher than expected student achievement can be incorporated within a school’s professional development efforts (Stronge, 2007).

Washoe County:

Washoe County School District is the second largest district in Nevada with over 58,000 students and 84 schools. The Washoe County study examined the relationship between teacher evaluation scores and student achievement. The data used in this study were the 2000–01 and 2001-02 student achievement scores from norm-referenced tests (District CRT and Terra Nova) for third-, fourth, and fifth-grade students, and 2001-02 teachers’ performance composite evaluation scores. Scores for the individual teacher’s four performance composites were averaged for a single indicator for teacher quality. Demographic data also used were student gender, race, special education status, and free or reduced lunch status and teacher education, experience, and year-round schedule.

The study fitted the data to a two-level HLM econometric model:

Level One:

$$\begin{aligned}&\text{PostTest} = \beta_0 + \beta_1 \text{ PreTest} + \beta_2 \text{ Female} + \\ &\beta_3 \text{ Non-White} + \beta_4 \text{ SpecialEd} + \beta_5 \text{ FRL} + R_i\end{aligned}$$

Level Two:

$$\begin{aligned}&\beta_0 = \gamma_{oo} + \gamma_{01} \text{ EvaluationScore} + \gamma_{02} \text{ Education/Experience} + \\ &\qquad \gamma_{03} \text{ Year-RoundSchedule} + \upsilon_0 \\ &\qquad \beta_1 = \gamma_{10} \quad \beta_2 = \gamma_{2o} \quad \beta_3 = \gamma_{3o} \quad \beta_4 = \gamma_{4o} \quad \beta_5 = \gamma_{5o}\end{aligned}$$

The results were mixed as to whether teacher evaluation scores are good predictors of student achievement with only four of the nine student achievement variables being significant at the 0.05 level. Perhaps a reason for the mixed results could be related to the context of teacher evaluation in the district. By their nature, teacher evaluations generally are relatively low-stakes where evaluators were less focused on differentiating teacher performance than they were on improving teacher morale through positive feedback and helping teachers identify areas of growth. It is also possible the evaluation standards were not specific enough to comprehensively assess teacher performance. This evaluation system is generic with respect to instructional content – the same evaluation form for all teachers regardless of what is taught or grade level. This may explain why all three of the fifth grade student achievement variables were highly significant with p-values around 0.01 and the third grade's were not significant with p-values around 0.30 (Kimball, 2004).

. In summary, the literature reviewed affirms the value of using econometric value-added models and regression analyses to determine teacher effectiveness and examine the relation between teacher effectiveness and student achievement levels. Even though principal evaluations and ratings have continued to have mixed results in their relation with student achievement, student achievement measured by standard assessments has been demonstrated to be a well-accepted metric for estimating teacher effectiveness. The literature also sends a clear message that principals need to be able to identify their effective as well as ineffective teachers because their ability to do so is critical for students' success.

## 2.4 Principal Ratings and Effective Teachers

### $360^{0}$ Feedback ® Evaluation System Study

How do principal ratings relate with effective teaching? A validation study of student, principal, and self-ratings for teachers was conducted in 1996 about the Wyoming Lincoln County School District's $360^{0}$ feedback ® evaluation system. The $360^{0}$ feedback ® evaluation system is based on using a full circle of appraisers within the evaluation process. This is similar to the business community's multiple feedback evaluation system where executives and managers are evaluated by their superiors, subordinates, peers, and customers. The system's goal is to improve individual's evaluations and ultimately improve the product for total customer satisfaction (Smith, 1993, as cited in Wilkerson, 2000, p. 181). Just as customer feedback is essential to the business feedback evaluation system, such team evaluations may not be enough in the field of education without including student achievement within the mix. For this reason, students were included in the Lincoln $360^{0}$ feedback ® evaluation system.

This study examined the relation of K-12 student achievement to teacher performance as measured by principal ratings, student ratings, and teacher self-ratings. Gain scores from criterion-referenced pretests and posttests were used along with the teacher performance rating scores. The criterion tests were developed over a three-year curriculum alignment process facilitated by the School Improvement Model at Iowa State University. Pretests were given in the fall of 1995 and posttests in the spring of 1996 over the subjects of mathematics, language arts, and reading.

Three parallel questionnaires designed to reflect the same aspects of teacher competence were used for teacher ratings from students, self-rating teachers, and principals. The student rating instrument was a questionnaire with twenty positive descriptors of teacher and student behavior using a three-point Likert-type scale for grades K–2 and a five-point scale for grades 3-12. The teacher self-feedback instrument elicited teachers' self-perceptions regarding the quality of their performance. Likewise the principal feedback instrument elicited principals' perceptions about the quality of the teachers' performance. All three instruments were tabulated the same way where an "Almost Always" rating for each of the twenty items would result in a total score of 80.

In addition, the principal completed the district's teacher summative evaluation, a fifteen item four-scale instrument where exceeds professionally competent = 4, professionally competent = 3, competent = 2, and unsatisfactory = 1, for a possible top score of 60. During this study, the principal did not see students' ratings of teachers or student achievement scores until after submitting the summative evaluations.

Regressing student gain scores onto student ratings, teacher self-ratings, principal ratings, and teachers' summative evaluations had almost a completely one-sided result. Student ratings were the best predictor of student achievement for all three subject areas and were highly significant at 0.05 or lower. In contrast, the principal rating and principal summative evaluations were not significant failing to meet the 0.05 rejection level in all subject areas. Mathematics teacher self-ratings were the only significant teacher-level predictor with language arts and reading teacher self ratings not significant. The important finding in this study is that students can discriminate teacher performance in relation to their own learning while nothing could be said about the relation between principal's teacher ratings and student achievement (Wilkerson, 2000).

### Jacob Study: Principal Ratings and Teacher Value-Added Effects

Another study (Jacob, 2003) was spurred by the question whether principals are capable of determining which teachers should be rewarded in a merit pay system. This study examined the relation of principal ratings of teachers to student achievement. Thirteen elementary school principals from a mid-sized western school district were asked to rate their teachers on ten characteristics using a scale of 1 to 10 (1= inadequate and 10 = exceptional).  These characteristics included dedication and hard work ethic, classroom management, parent satisfaction, positive relationship with administrators, and the ability to improve math and reading achievement. These ratings were mean-centered and standardized by characteristic for each school. Longitudinal student achievement gain scores from 1998 through 2003 on the district's criterion-referenced tests were used to estimate the value added by each teacher.

By comparing principal ratings with estimated teacher value-added effects positive correlations (0.32 in reading and 0.36 in mathematics) were found between principal ratings and teacher effectiveness for the best and worst teachers. Principals were very good at identifying teachers who produced the largest or smallest achievement gains in their schools. Principals identified their best teachers 52 percent of the time in reading and 69 percent of the time in mathematics. There were similar results for identifying the worst teachers. However, the principals were less successful when ranking categorically teachers in the middle 60-80 percent of the ability distribution (Jacob, 2006, pp. 60-63).

### Cincinnati Public Schools study (2001-2002)

The Cincinnati Public Schools is a large urban district with 48,000 students and 3,000 teachers in more than 70 schools and programs. In response to state-level changes in teacher licensing requirements, the district designed a new teacher evaluation system for use from 2000 through 2003. There were sixteen performance standard with accompanying rating rubrics describing levels of performance as unsatisfactory, basic, proficient, and distinguished. There were four performance areas measured: domain 1 – planning and preparation, domain 2 – creating an environment for learning, domain 3 – teaching for learning, and domain 4 – professionalism. Teachers were evaluated based on two evaluative sources: six classroom observations and teacher compiled portfolio. Principals conducted two of the observations while teachers hired specifically as evaluators for a three-year term conducted the other four observations. The teacher evaluators made a summative rating based on the summaries of the six observations on each of the standards in domains 2 and 3. Principals rated teachers on the standards in domains 1 and 4 using primarily information from the teachers' portfolios. The evaluation system was initially designed to be used as part of a part of the pay system and became a very high-stakes issue for many teachers. Predictably, this link between the evaluation system and the pay system was rejected in a special election. However, the district has continued to use the evaluation system for new teachers and some veterans. In either case poor evaluations could lead to termination.

The Cincinnati Public Schools study examined the relation between the district's newly implemented standards-based evaluation system with teacher effectiveness.

They wanted to be justified in inferring that teachers with high evaluation performance scores are better performers and produced more student learning. Specifically, the study focused on the relation between teacher evaluation scores and value-added measures of student achievement. If no empirical relationship were found then the new evaluation system should be discontinued or revised.

The study used teacher evaluation scores for 270 teachers evaluated in 2000-01 and 335 teachers evaluated in 2001-02. Student achievement was used from the district's annual March criterion-referenced, standard, and state proficiency tests' scores for grades three through eight. Student demographic variables included race, gender, receipt of free or reduced lunch, special education status, and days enrolled in school. The analysis used a value-added model where student achievement was defined as the residual from a regression of the 2001-02 test score in a subject on the prior year's score and other student-level variables thought to potentially affect student test performance. The first step involved producing an average achievement level for each teacher's students – controlling for prior achievement in the subject and student characteristics thought to influence test scores. A two-level hierarchical linear model was estimated to do this:

Level 1 Model:

$$\text{Posttest} = \beta_0 + \beta_1\ \text{Pretest} + \beta_2\ \text{Female} + \beta_3\ \text{Free/Reduced Lunch} + \beta_4\ \text{Non-White} + \beta_5\ \text{Special Ed} + \beta_6\ \text{Days enrolled in school} + \text{R}$$

Where: $\beta_0 \ldots\ldots \beta_6$ = Within classroom regression coefficients

$\text{R}$ = Level 1 error on individual student residual.

All level 1 predictors were grand-mean centered.

Level 2 Model:

$$\beta_{0j} = \gamma_{00} + \upsilon_{0j}$$

Where: $\beta_{0j}$ = Intercept in classroom j

$\gamma_{00}$ = Average intercept across classrooms

$\upsilon_{0j}$ = Teacher specific differences from the average of the classroom intercepts.

The slopes for all level 1 variables were considered fixed. From the level 1 model, the empirical Bayes (EB) intercept residuals were determined as the measure of the average student performance relevant to each teacher. Given the grand mean-centering, the EB intercept residuals were the difference for the "average" student: average in prior year test score and other characteristics at Level 1. The EB intercept residuals were then correlated with teacher evaluation scores. Partial correlations between the evaluation scores and the EB intercept residuals, controlling for teacher experience and the year in which the teacher was evaluated, were all positive except for seventh grade science. In reading and mathematics there were moderate positive relationships between teacher evaluation scores and student achievement but a weak relationship in science.

The study demonstrated that the Cincinnati Public Schools teacher evaluation scores had a moderate degree of criterion-related validity. The evaluation system was able to identify which teachers had higher than expected levels of achievement, as measured by test scores, to a degree greater than chance. This conclusion was made even though the relatively small correlations within the 0.3 and 0.4 range indicated only a small proportion (9% to 16%) of variance in student achievement was potentially due to variation in teacher performance. However, very high correlations between teacher evaluation scores and student achievement measures are unlikely to happen (Milanowski, 2004).

The finding from studies examining the relation of principal ratings and effective teachers were mixed with some support for using principal ratings to identify the best and worst teachers only. However, using ratings to differentiate the effectiveness of teachers other than the best and the worst was not supported.

A legitimate concern can be raised about how teacher effects on student achievement are estimated within regression models used within the reviewed studies. In most of the reviewed studies, rather than estimating unobserved teacher effects separately by using individual teacher effect dummies, the unobserved teacher effects are instead culled from the residual errors or they are combined with observable school resources. As Dan Goldhaber, et al (Goldhaber, 1997), point out in their review of econometric models used for determining teacher effect on student achievement, the inherent unobservable characteristics, such as, teacher skill, behavior, and motivation and classroom student peer effects are omitted from the models. These are usually left languishing within the residual error as unobservable characteristics or combined with observable characteristics.

Goldhaber et al examined the typical regression model: $Y_{ij} = \beta X_{ij} + \gamma S_j + \varepsilon_{ij}$

Where: $Y_{ij}$ = achievement for student i at school j
$X_{ij}$ = individual and family background variables
$S_j$ = vector of schooling resources which do not vary across students
$\varepsilon_{ij}$ = random error term

Suppose, $S_j$, the vector of schooling resources could be expressed in two parts: observable characteristics $Z_1$, such as class size, teacher experience, teacher degree level and unobservable characteristics $Z_2$, such as teacher skill, behavior, and motivation, and student peer effects.

The true model would then be: $Y_{ij} = \beta X_{ij} + \gamma_1 \ Z_{1j} + \gamma_2 \ Z_{2j} + \varepsilon_{ij}$

Where: $Y_{ij}$ = achievement for student i at school j

$X_{ij}$ = individual and family background variables

$Z_1$ = Observable school characteristics

$Z_2$ = Unobservable school characteristics

$\varepsilon_{ij}$ = random error term

By including $Z_2$ within the model, the unobservable characteristics of teacher effects and student peer effects could now be used to explain their portion of the variance in student achievement (Goldhaber, 1997). For this reason, separate teacher and student dummies should be used within regression models to pick up the effect each teacher and each student have on student achievement. In this study, individual teacher dummies for 134 teachers will be used to pick up each teacher’s effect on student achievement. Likewise, individual student dummies for 2872 students will be used as a control variables to pick up each student’s effect on student achievement.

# Chapter 3

# Methodology

## 3.1 Student, Teacher and Principal Data

This study used student, teacher, and principal data for seven middle schools from a large Midwest school district. Table A shows the middle schools' demographics:

Table A 2006-2007 Middle School Demographics

| School | Attendance | Enrolled | SES | SES % | White | White % | 7th Grade | 8th Grade | Teachers |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 95.3% | 556 | 87 | 15.6% | 441 | 79.3% | 282 | 274 | 42 |
| 2 | 95.7% | 501 | 41 | 8.2% | 437 | 87.2% | 230 | 271 | 44 |
| 3 | 94.7% | 494 | 91 | 18.4% | 357 | 72.3% | 250 | 244 | 39 |
| 4 | 96.4% | 584 | 45 | 7.7% | 492 | 84.2% | 283 | 303 | 45 |
| 5 | 95.6% | 675 | 53 | 7.9% | 610 | 90.4% | 338 | 337 | 50 |
| 6 | 95.9% | 575 | 10 | 1.7% | 415 | 72.2% | 281 | 294 | 47 |
| 7 | 95.1% | 913 | 147 | 16.1% | 648 | 71.0% | 435 | 478 | 71 |

Note: Principal survey for School 6 was not returned.

(KSDE, 2010)

Since student achievement data was available for all 7th and 8th students but not for other grade levels this study was limited to the school district's seven middle schools. This achievement data included MAP RIT scores for Fall, 2006, and Spring and Fall, 2007.

The 7th and 8th grade students' MAP RIT scores from the Fall, 2006, and Spring, 2007, and Fall, 2007 MAP assessments in the subject areas of communications, English, and mathematics were used to examine the relation of student achievement and principals' ranking of teachers by their effectiveness. Confidentiality was maintained by assigning codes to teachers and students only for the purpose of data analysis and not for identifying individual students or teachers. Students' achievement scores were linked with the teacher codes as a tracking method throughout this study.

The school district provided data from 8,153 total MAP scores for 2,557 communications students, 2724 English students, and 2872 math students and 134 teachers (46 communications, 40 English, and 48 math teachers) were used for the analyses. Student demographic data include gender, race, free or reduced lunch eligibility, and gifted status. Teacher demographic data include gender, race, number of students, years' experience, and education level.

The school district's director of testing services distributed human subjects' informed consent forms and principal surveys to the middle school principals. The director also acted as the principal's contact person. Principals completed the human subjects' informed consent form and returned it in a provided pre-addressed envelope. Principals completed the survey (Appendix A), cut off the teachers' names, and returned the survey using a separate provided pre-addressed envelope. Six principal surveys were returned as one principal declined the survey.

The middle school principals completed a survey asking them to select teachers from a list of teachers they expect to have higher than normal gain scores from their students. (Appendix A). They also ranked in order (1, 2, 3…..) the teachers from the most effective teacher to the least effective teacher within each subject area. Since the number of teachers within subject areas varied among the schools, the principals' teacher rankings were converted to quintiles. For example, a number one principal's teacher ranking for the top teacher was converted to a five as a quintile measure.

Principals' teacher ranking lists were converted to quintiles using the following formula:

Given k number of teachers in a school's subject area and $k/5 = z$

Then, the quintiles would be:

Quintile 1: $k \geq \text{rank} > k - z$
Quintile 2: $k - z \geq \text{rank} > k - 2z$
Quintile 3: $k - 2z \geq \text{rank} > k - 3z$
Quintile 4: $k - 3z \geq \text{rank} > k - 4z$
Quintile 5: $k - 4z \geq \text{rank} > k - 5z$

Example:

Eight teachers in a department

$k = 8$ and $k/5 = 1.6$

Quintiles would be:

Quintile 1……. $8 \geq \text{rank} > 8 - 1.6$
$8 \geq \text{rank} > 6.4$
Quintile 2 ……. $6.4 \geq \text{rank} > 4.8$
Quintile 3 ……. $4.8 \geq \text{rank} > 3.2$
Quintile 4 ……. $3.2 \geq \text{rank} > 1.6$
Quintile 5 ……. $1.6 \geq \text{rank} > 0$

| Principal Rank | Quintile |
|---|---|
| 1 | 5 |
| 2 | 4 |
| 3 | 4 |
| 4 | 3 |
| 5 | 2 |
| 6 | 2 |
| 7 | 1 |
| 8 | 1 |

### 3.2 Method of Analysis

Based on the literature reviewed, the econometric value-added model with subsequent regression analyses best fit this study's data and purpose. This study fitted individual level data to a value-added model to estimate teacher effects on students' 06-07 MAP gain scores and then ran subsequent regression analyses to estimate principal's ranking effects on teachers' average Spring 07 MAP scores, on teachers' average 06-07 MAP gain scores, and on teachers' value-added effects on students' 06-07 MAP gain scores.

To estimate principal's ranking effect on student Spring 07 MAP score:

$$Y_t = \beta_0 + X_t \beta + \varepsilon_t$$

Where: $Y_t$ = Teacher's average Spring 07 MAP score, $X_t$ is a vector of predictors that include number of students, gender, years experience, masters degree, and principal's ranking. This is done by aggregate and individual subject areas.

To estimate principal's ranking effect on teacher's average MAP 06-07 gain score:

$$Y_t = \beta_0 + X_t \beta + \varepsilon_t$$

Where: $Y_t$ = Teacher's average MAP gain score 06-07, $X_t$ is a vector of predictors that include number of students, gender, years experience, masters degree, and principal's ranking. This is done by aggregate and individual subject areas.

To estimate principal's ranking effect on teacher's value-added to student gain scores:

$$\gamma_{\mathrm{t}} = \beta_0 + \mathrm{X}_t \beta + \varepsilon_t$$

Where: $\gamma_{\mathrm{t}}$ = Teacher's value-added effects on MAP gain scores 06-07, $\mathrm{X}_t$ is a vector of predictors that include number of students, gender, years experience, masters degree, and principal's ranking. This is done by aggregate and individual subject areas.

The above teachers' value-added effects on MAP gain scores 06-07 are obtained by using the following regression model fitted on individual level data:

$$\mathrm{Y}_{\mathrm{i}} = \alpha + \beta K_i + \sum_{1}^{l} \lambda_i S_i + \sum_{1}^{t} \gamma_i T_i + \varepsilon_i$$

Where: $i$ = student, $l$ = number of students, $t$ = number of teachers, $\mathrm{Y}$ = MAP gain score (06-07), $K$ = Fall 06 MAP score, $S$ = Dummy for student ID, $\lambda$ = Coefficient for the dummy for student ID, $T$ = Dummy for teacher ID, and $\gamma$ = Coefficient for the dummy for teacher ID, $\varepsilon_i$ = residual error

(Saatcioglu, 2010)

The acceptable level of significance for this study is $p < 0.05$.

# Chapter 4

# Results

## 4.1 Research Question: Are principals good at identifying effective teachers?

Table 1 shows the data and descriptors used in this study:

Table 1: Data Variables and Descriptors

| | |
|---|---|
| | Student Level Data |
| Student ID | Student identification number used for student dummies |
| Male | Student gender varaible, male = 1, female = 0 |
| Race | Individual student race variables for white, black, asian, hispanic, multi-race, native american |
| LowSES | Student free or reduced lunch eligibility |
| Gifted | Student gifted status |
| Fall 06 MAP Score | Fall 2006 MAP 7th and 8th grade reading and math scores |
| Spring 07 MAP Score | Spring 2007 MAP 7th and 8th grade reading and math scores |
| Fall 07 MAP Score | Fall 2007 MAP 7th and 8th grade reading and math scores |
| | Teacher Level Data |
| Teacher ID | Teacher identification number used for teacher dummies |
| Students Enrolled | Teacher's number of students enrolled in their classes |
| Male Teacher | Teacher gender variable, male = 1, female = 0 |
| Teacher Race | Individual teacher race variables for white, black, asian, hispanic, multi-race, native american |
| Years Experience | Teacher's years experience |
| Masters Degree | Teacher's education level, 1 = Masters, 0 = Bachelors |
| Principal's Rank | Quintile principal ranking of teacher(5= Top Teacher) |
| Mean Spring 07 MAP | Teacher's Average Spring 07 MAP Score |
| Mean MAP Gain 06-07 | Teacher's Average MAP Gain Score 06-07 |
| Teacher Value-Added | Teacher's value-added to MAP 06-07 gain scores |

The following chart is a side-by-side comparison of teachers' rank order for teacher's average Spring 2007 MAP score, teacher's average student gain score, and teacher value-added to student gain scores:

Table 2:Teachers' Rank Order Comparison by Dependent Variable

| Rank Order | Teacher's Average Spring 07 MAP Score | Teacher's Average MAP Gain Score 06-07 | Teacher's Value-Added to Gain Scores |
|---|---|---|---|
| 1 | 361 | 322 | 344 |
| 2 | 347 | 367 | 343 |
| 3 | 345 | 362 | 326 |
| 4 | 321 | 312 | 333 |
| 5 | 336 | 141 | 322 |
| 6 | 322 | 341 | 341 |
| 7 | 335 | 113 | 367 |
| 8 | 333 | 311 | 335 |
| 9 | 343 | 326 | 346 |
| 10 | 354 | 366 | 354 |
| 11 | 263 | 354 | 362 |
| 12 | 326 | 364 | 342 |
| 13 | 342 | 211 | 325 |
| 14 | 344 | 167 | 324 |
| 15 | 325 | 351 | 364 |
| 16 | 367 | 346 | 355 |
| 17 | 324 | 251 | 361 |
| 18 | 346 | 365 | 351 |
| 19 | 223 | 152 | 316 |
| 20 | 221 | 122 | 345 |
| 21 | 341 | 352 | 323 |
| 22 | 366 | 333 | 366 |
| 23 | 316 | 128 | 365 |
| 24 | 226 | 344 | 331 |
| 25 | 351 | 214 | 312 |
| 26 | 331 | 225 | 321 |
| 27 | 126 | 343 | 314 |
| 28 | 243 | 111 | 313 |
| 29 | 144 | 222 | 347 |
| 30 | 128 | 325 | 128 |
| * | * | * | * |
| * | * | * | * |
| * | * | * | * |
| * | * | * | * |
| 61 | 135 | 115 | 213 |
| 62 | 212 | 363 | 252 |
| 63 | 365 | 244 | 112 |
| 64 | 334 | 134 | 266 |
| 65 | 235 | 331 | 251 |
| 66 | 116 | 224 | 263 |
| 67 | 312 | 361 | 221 |
| 68 | 113 | 263 | 212 |
| 69 | 162 | 223 | 115 |
| 70 | 311 | 144 | 253 |
| * | * | * | * |
| * | * | * | * |
| * | * | * | * |
| * | * | * | * |
| 103 | 151 | 114 | 145 |
| 104 | 167 | 138 | 262 |
| 105 | 145 | 315 | 138 |
| 106 | 262 | 262 | 136 |
| 107 | 165 | 231 | 151 |
| 108 | 315 | 151 | 165 |

Table 2 tells an interesting story. Please note the highlighted teachers and how they rank within each rank ordered column. Teacher 361 is the top teacher in column 1 with the highest average Spring 07 MAP score, but dropped to 67th for average gain score, and is 17th for value-added to gain scores. Teacher 322 is consistently ranked in the top of the three columns – 6th for average Spring MAP score, 1st for average gain score, and 5th for value-added to gain score. Teacher 344 is 14th for average Spring MAP score, 24th for average gain score, and 1st for value-added to gain scores. Other teacher rankings tell similar stories. The teachers' rank orders don't seem to be correlated very well among the ranked categories. This clearly shows the dilemma principals face when rating their teachers. What do they look at to determine their effective teachers? If value-added to gain scores is what principal's should be using to determine effective teachers, are they able to discern their teachers' value-added effect on student achievement in conjunction with their teachers' average Spring 07 MAP score and average MAP gain score?

Table 3 gives some insight into a principal's dilemma:

Table 3: Correlations for Dependent Variables

| | Teacher's Average Spring 07 MAP Score | Teacher's Average 06-07 MAP Gain Score | Teacher's Value Added to MAP Gain Score |
|---|---|---|---|
| Teacher's Average Spring 07 MAP Score | 1.000 | | |
| Teacher's Average 06-07 MAP Gain Score | 0.188<br>(0.029) | 1.000 | |
| Teacher's Value Added to MAP Gain Score | 0.669<br>(0.000) | 0.463<br>(0.000) | 1.000 |

Note: Significance value in parenthesis

Table 3 further illustrates a principal's dilemma with its pair-wise correlations of the dependent variables.  The weak correlation between a teacher's average Spring 07 MAP score and a teacher's average 06-07 MAP gain score is expected as a raw score is different from a gain score by definition. The somewhat strong correlation between teacher's value-added to MAP gain score and teacher's average 06-07 MAP gain score is not surprising as teacher value added effects are generated by a value-added model using gain scores as the dependent variable with the value-added teacher dummy's effect coefficient. The strong correlation between a teacher's value-added to MAP 06-07 gain score and a teacher's average Spring 07 MAP score is also not surprising as the gain score is a raw score difference score and the Spring 07 MAP score is the raw score used in this differencing.

Based on these correlations, teacher's average Spring 07 MAP score rank and teacher's value-added to MAP gain score rank  would more likely be similar than teacher's average 06-07 MAP gain score rank and teacher's value-added to gain score rank. That is, it is more likely for a teacher to have a high average Spring 07 MAP score rank and a high value-added to gain score rank than it is likely for a teacher to have a high average gain score rank and a high value-added to gain score rank. Then, if principals are using student achievement data to identify effective teachers, using teachers' average Spring 07 MAP score rank would correlate more with teacher value-added to student gain score rank than using teachers' average 06-07 MAP gain score rank.

The least likely to be similar would be a teacher's average 06-07 MAP gain score rank and a teacher's average Spring 07 MAP score rank. That is, for teachers having a high average 06-07 MAP gain scores, it is not likely these same teachers would have a high average 07 Spring MAP scores. This makes sense as students with low Fall MAP scores potentially could have higher gain scores after taking the Spring MAP because they have larger range of potential gain. On the other hand, students with high Fall Map scores may not have higher gain scores than student with low Fall MAP scores since their scores are already high in the first place.

So, are principals good at identifying their effective teachers? Table 4 shows the multiple regression models' coefficient results for the dependent variables teacher's average Spring 07 MAP score, teacher's average MAP 06-07 gain score, and teacher's value-added to the MAP 06-07 gain score regressed onto the independent variables. Multiple regression models coefficient results with standard errors are in Appendix B. Examining the table's significant predictor effects reveal some interesting findings:

Table 4: Multiple Regression Models

| | All Subjects | | | Communications | | | English | | | Mathematics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean Spring 07 MAP | Mean MAP Gain 06-07 | Teacher Value-Added | Mean Spring 07 MAP | Mean MAP Gain 06-07 | Teacher Value-Added | Mean Spring 07 MAP | Mean MAP Gain 06-07 | Teacher Value-Added | Mean Spring 07 MAP | Mean MAP Gain 06-07 | Teacher Value-Added |
| Students Enrolled | 0.054 | 0.594 * | 0.329 | 0.011 | 0.343 | -0.157 | 0.087 * | 0.331 | -0.112 | 0.102 | 1.018 * | 0.037 |
| Male Teacher | 1.129 | -0.682 | 3.131 | 6.629 | 1.817 | 4.432 | 8.094 | -0.569 | -1.378 | -11.378 | -2.332 * | -0.471 |
| Years Exp. | 0.093 | 0.034 | 0.004 | 0.177 | 0.019 | -0.021 | 0.164 | 0.037 | -0.019 | 0.063 | 0.044 | 0.121 |
| Masters Degree | -1.717 | -0.461 | 0.441 | -3.996 | -1.271 | 1.320 | -0.572 | -0.283 | -2.641 | -0.525 | 0.322 | 3.239 * |
| Principal's Rank | -0.340 | 0.325 * | 0.547 | 1.263 * | 0.203 | 0.150 | 0.318 | 0.032 | 0.367 | -3.246 * | 0.534 * | 0.207 * |
| Constant | 224.945 * | 0.478 | -9.500 * | 218.782 * | 2.133 | -8.788 * | 215.225 * | 1.459 | -6.416 * | 239.764 * | -1.310 | -6.334 |
| R-Squared | 0.03 | 0.12 | 0.07 | 0.12 | 0.07 | 0.10 | 0.20 | 0.08 | 0.12 | 0.13 | 0.42 | 0.34 |

Note: * $p < 0.05$ (two-tailed)

The number of students taking the MAP per teacher does matter – the more students a teacher has taking the MAP, the more the effect on student gain scores overall and in particular for mathematics. The number of students taking the MAP per math teachers has a strong carry-over effect on gain scores contributing to the aggregate effect of the number of students for all subjects' gain scores.

For English, the more students a teacher has taking the Spring 07 MAP, the more the effect on these scores. But this effect does not have as strong of a carry-over effect on all subjects' Spring 07 MAP scores as what seemed to be true for math teachers' carry-over effect contribution to aggregate effect on all subjects' gain scores.

It is interesting that math teachers with masters degrees have a very strong value-added effect on student gain scores while nothing could be said about communications and English teachers with masters degrees.

Principals were able to identify communications teachers having high average Spring 07 MAP scores. But, principals were not able to identify communications teachers who had high average MAP gain scores or high value-added to gain scores.

Based on teachers' value-added to student gain scores, principals were able to identify their effective math teachers but not able to identify their effective communications and English teachers. They did rank high their math teachers who have high average gain scores and high value-added to these gain scores. However, principal's rank effect on teachers' value-added to gain score is half that of the principal's rank effect on teacher's average gain scores. So, principals were able to identify their high value-adding teachers only half as well as they were able to identify teachers with high average gain scores. It is interesting to note, the high negative coefficient for principal's rank related to average Spring 07 MAP scores may indicate that principals are adjusting for bias caused by the nature of students' math enrollment. Students are usually enrolled in math classes by ability levels, so some teachers will have higher ability level students than others and these students will usually have high MAP scores. The negative coefficient may be due to principals ranking high those teachers who are working with lower ability level students who may not have high Spring 07 MAP scores. Principals may be using teacher's average MAP gain and sensing their teacher's value-added to gain scores to identify such teachers – as indicated by the positive principal's rank effects for teachers' average MAP gain and value-added to gain scores. On the other hand, this negative coefficient could also mean that principals are dead wrong about math teachers' Spring 07 MAP scores.

Considering the positive student enrollment effect, the negative male teacher effect on math gain scores, and the relatively high r-squared value, the principal's rank effect for math teachers is very strong. The overall principal's rank effect on all three mathematics dependent variables indicates that principals can determine the effectiveness of their math teachers.

In conclusion, four things stand out from the findings:

1) Principals can identify their effective math teachers. However, principals can't identify their effective communications and English teachers. Principals' rankings of teachers tend to correlate more with math teachers than communications and English teachers regarding student gain scores and teachers' value-added to student gain scores. This disparity between being able to identify effective math teachers as opposed to not being able to identify effective communications teachers may be due to the objective, test-

oriented, nature of mathematics as opposed to the more subjective, interpretive, nature of communications and English. It may be easier for principals to know when effective teaching is happening in mathematics than it is to know when effective teaching is happening in communications and English.

2) Principal's rank effect on teachers' value-added to gain scores is half that of their effect on teachers' average gain scores. Principal's only do half as well identifying effective value-adding teachers than they do identifying teachers with high average gain scores.

3) Principal's rank has a negative effect on teachers' average Math Spring 07 scores. This may indicate that principals are adjusting teachers' ranking because of perceived teachers' value-added to gain scores.

4) What are principals using when they rank their teachers? It is not clear from the findings what principals are using. Principals' rankings of teachers don't seem to follow a clear pattern or rationale across the subject areas. Most noticeable, principals seem to be clueless when determining communications and English teachers' value-added effects on student gain scores. So, wouldn't it make sense for a school district to let their principals know their teachers' value-added effects on students' gain scores along with their teachers' average raw and gain scores?

# Chapter 5

# Discussion

## 5.1 Overview

Researchers typically define effective teaching by students' residual gain on standardized tests even though there are a number of other ways to define effective teaching. The complex and often subjective nature of teaching and learning is seemingly at odds with the limits of such an objective and empirical definition of effective teaching. Arguably, effective teaching has no single definition but is continually being defined within the context of teaching and student learning. Granted, test scores do not capture all the facets of student learning, test scores are widely available, objective and are recognized as important indicators of achievement by educators, policymakers, and the public (Rockoff, 2004). Again, for better or for worse, test scores have increasingly become a commonly accepted standard for teacher effectiveness. Because of this common acceptance, this study defined teacher effectiveness in terms of student achievement.

Initially, the purpose of this study was to affirm principals' ability to identify effective teachers. The question posed, "Are principals good at identifying effective teachers?" had a mixed answer of yes for math teachers, but no for communications and English teachers. Apparently, math teachers can be more readily identified as being effective than teachers from other subjects. Principals seem to be able to pick up on the value-added to student achievement gain for math teachers than they do for the other teachers. This is evident from the negative effect principal's ranking had on teachers' average Spring 07 MAP scores and the opposite positive effects principal's ranking had on teachers' average gain scores and teachers' value-added to gain scores.

This study affirmed other studies' use of teachers' value-added effects on student achievement as an indicator of teachers' effectiveness (Stronge, 2007, Gordon, 2006, Kimball, 2004, Nye, 2004, Milanowski, 2004, Mendro, 1998, Wright, 1997, Sanders, 1996). Using teachers' value-added effects as an indicator of teachers' effectiveness can be a valuable addition to the ways principals identify their effective teachers. Even though research supports using teachers' value-added effects on student achievement, the evidence from this study indicates that principals are not picking up on this value-added effect when assessing the overall effectiveness of their teachers. So, if principals are not sensing their teachers' value-added effects on student achievement with what they are currently doing, then letting them know what these effects are would be very helpful.

Currently, most principals use student achievement data in the form of raw data and gain scores derived from this raw data. However, this does not give a complete picture of their teachers' effectiveness. They are missing a key indicator – teachers' value-added to student gain scores. This was clearly illustrated by table 2's 3-way comparison of the teacher's rank for raw score, gain score, and value-added to student gain scores. Principals need to know about the impact this 3-way comparison has on teacher ranking when determining a teacher's effectiveness. Indeed, knowing a teacher's value-added effect relative to other teachers would be a valuable indicator and addition to the tools principals use in determining teachers' effectiveness.

## 5.2 Implications:

A teacher evaluation system should include measures of teachers' value-added to student achievement as one of the system's indicators of teacher effectiveness. The often ego-stroking nature of current evaluation systems where nearly every teacher receives a satisfactory evaluation is incompatible with the principal's task of determining a teacher's effectiveness. However, a system driven solely by test-based measures of value-added would never be accepted as fully legitimate. Value-added measures of effectiveness should be a part of a viable evaluation system that includes alternative ways of discerning among teachers beyond simply test scores (Gordon, 2006).

Value-added measures should be used in conjunction with the formal and informal evaluation processes currently being used by principals. Informed discussions between principals and teachers and among teachers about value-added effects' link with other indicators of teacher effectiveness should enhance the evaluative process. But, care must be taken if using NCLB standards-based tests in determining teacher value-added. There may be a ceiling effect built into the test where students could perform better than the standards' limits allow and their test scores would not truly reflect their performance levels (Koedel, 2009). Care must also be taken to weigh a teacher's single year's value-added effect against a teacher's series of years' value added effects. Bear in mind, however, a one-year's negative effect from an ineffective teacher can last through three years' of highly effective teachers (Mendro, 1998). Because of this, principals should make retention decisions before ineffective teachers become tenured (Gordon, 2006, Mendro, 1998, Sanders, 1996).

Principals should be evaluated on their ability to identify effective and ineffective teachers. This would put the onus on principals to identify effective and non-effective teachers early. A data system linking student performance with the effectiveness of individual teachers over time needs to be in place to assist and support principals' decisions concerning teacher effectiveness. Thereby, students' performance could be tracked from year-to-year and linked with their teachers to inform principals of teachers' effectiveness. Technical support along with interpretive training needs to be provided to principals (Gordon, 2006). Even so, the value-added effects derived through such a data base should be linked with alternative indicators of teacher effectiveness (McCaffrey, 2004).

Ideally, value-added effects should become an essential part of a school's collaborative dialogue. Teacher and principal evaluations and professional development should weave effectiveness data throughout the evaluative process and professional growth opportunities. Both purposes of accountability and professional growth can be met by teachers and principals' collaborative examining of teacher effects on student achievement and sharing what effective teachers are doing when they experience higher than expected student gains. By including teacher value-added effects within a teacher's evaluation, a critical empirical perspective is provided to the multifaceted process of teacher evaluation. This inclusion would be further strengthened when such teacher effectiveness data is associated with professional development opportunities involving the behaviors and characteristics of effective teachers. Ultimately, by doing so, student achievement will benefit (Stronge, 2007).

In reality, teachers and principals would far less likely use measures of student achievement when judging teachers' effectiveness. They would more likely use other measures, such as, student behavior and affect in class, student feedback on courses, student success in college, or student success after college. Since most teachers look at test scores but feel that tests do not tell the whole story, they would more likely view teacher value-added effects in the same manner. Instead, teachers would continue to rely on data gathered anecdotally rather then systematically and to rely on intuition and experience rather than using empirical data.  Unfortunately, data-driven feedback such as value-added effects usually do not change the way teachers think about their effectiveness (Ingram, 2004). Hopefully, teachers and principals can make a paradigm shift and include teacher value-added effects within their reflecting on teachers' effectiveness.

### 5.3 Recommendations for Future Research

This study leads to possible areas for inclusion in a larger scale study:

1) Follow-up interviews with principals to find out exactly what they used to determine teacher rankings.

2) Follow-up interviews with randomly selected teachers to find out how they define effective teaching what they do to be an effective teacher .

3) Include teachers' class period dummies in this study's value-added model to pick up the effects of an individual teacher's class periods on student gain scores. The unique effect of each one of a teacher's class period could have an effect on the overall teachers' value-added effects.

4) Include multiple years to examine teachers' longitudinal value-added effects.

5) Include multiple years to examine the relation principals' value-added training and use of value-added data with a principals' ability to identify effective teachers.

6) Examine the relationship of teacher characteristics and behaviors with highly effective teachers.

### 5.4 Summary

The purpose of this study was to find out if principals could identify their effective teachers. Along with this potential affirmation, there would have been an important tacit assurance of principals' performance responsibilities assumed true – their competency in hiring, developing, and evaluating teachers. However, the findings only affirmed the ability of principals to identify their effective math teachers. Principals do have the ability to perceive the value-added effects of their math teachers, but principals were not able to do the same for their communications and English teachers.

This mixed finding points to a definite need to find a way to help principals identify their effective teachers. As recommended by multiple studies (Gordon, 2006, Mendro, 1998, Sanders, 1996), principals should use a teacher's value-added effect on student achievement gain as one indicator of the teacher's effectiveness. Doing so requires a school district's commitment to maintain the required database, provide statistical software to generate the value-added effects, and give the technical assistance necessary for principals to use and interpret the value-added effects.

Knowing a teacher's value-added effect on student achievement gain would enhance a principal's perspective when assessing a teacher's effectiveness through their formal and informal evaluation processes. After considering all the indicators of a teacher's effectiveness, principals can then make an informed decision rather than rely on perception to determine a teacher's effectiveness.

## Appendix A

**Principal Survey Instructions:**

- **Higher Gain Scores** - Check the box for all teachers you expect higher than normal gain scores.
- **Effective Rank** - Rank, 1, 2, 3, …, teachers for each subject area. No ties please.

**Delivery Instructions:**

- Cut off teacher names along the dotted line.
- Return this survey to Jim Gray in large return envelope.
- Sign and return the blue consent form to Jim Gray in small return envelope.
- Questions - Call XXXXXXX (XXX-XXXX).

Please cut off teacher names when survey is complete

| Teacher Name | Subject Area | Higher Gain Scores (✓) | Effective Rank (1, 2, 3 …) | Teacher Code |
|---|---|---|---|---|
| | Communications | | | |
| | Communications | | | |
| | Communications | | | |
| | Communications | | | |
| | Communications | | | |
| | Communications | | | |
| | Communications | | | |
| | Communications | | | |
| | | | | |
| | English | | | |
| | English | | | |
| | English | | | |
| | English | | | |
| | English | | | |
| | English | | | |
| | English | | | |
| | | | | |
| | Mathematics | | | |
| | Mathematics | | | |
| | Mathematics | | | |
| | Mathematics | | | |
| | Mathematics | | | |
| | Mathematics | | | |
| | Mathematics | | | |

# Appendix B

Table 1: Multiple Regression Models

Dependent Variable: Teacher's Average Spring 07 MAP Score

| | All Subjects | | | Communications | | | English | | | Mathematics | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Independent Variable | | | | | | | | | | | | |
| Students Enrolled | 0.054 | | (0.033) | 0.011 | | (0.039) | 0.087 | * | (0.037) | 0.102 | | (0.086) |
| Male Teacher | 1.129 | | (5.468) | 6.629 | | (9.345) | 8.094 | | (9.299) | -11.378 | | (8.889) |
| Years Experience | 0.093 | | (0.125) | 0.177 | | (0.158) | 0.164 | | (0.152) | 0.063 | | (0.274) |
| Masters Degree | -1.717 | | (3.283) | -3.996 | | (3.800) | -0.572 | | (4.094) | -0.525 | | (7.278) |
| Principal's Rank | -0.340 | | (0.903) | 1.263 | * | (0.626) | 0.318 | | (1.088) | -3.246 | * | (1.618) |
| Constant | 224.945 | * | (4.217) | 218.782 | * | (5.210) | 215.225 | * | (5.095) | 239.764 | * | (9.165) |
| R-Squared | 0.03 | | | 0.12 | | | 0.20 | | | 0.13 | | |

Notes: Standard errors are shown in parenthesis. ***** p < 0.05 (two-tailed)

Table 2: Multiple Regression Models

Dependent Variable: Teacher's Average Gain Score

| | All Subjects | | | Communications | | English | | Mathematics | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Independent Variable | | | | | | | | | | |
| Students Enrolled | 0.594 | * | (0.256) | 0.343 | (0.460) | 0.331 | (0.399) | 1.018 | * | (0.499) |
| Male Teacher | -0.682 | | (1.019) | 1.817 | (2.669) | -0.569 | (2.441) | -2.332 | * | (1.173) |
| Years Experience | 0.034 | | (0.024) | 0.019 | (0.048) | 0.037 | (0.040) | 0.044 | | (0.038) |
| Masters Degree | -0.461 | | (0.620) | -1.271 | (1.116) | -0.283 | (1.058) | 0.322 | | (0.990) |
| Principal's Rank | 0.325 | * | (0.163) | 0.203 | (0.336) | 0.032 | (0.290) | 0.534 | * | (0.263) |
| Constant | 0.478 | | (1.185) | 2.133 | (2.299) | 1.459 | (1.911) | -1.310 | | (2.162) |
| R-Squared | 0.12 | | | 0.07 | | 0.08 | | 0.42 | | |

Notes: Standard errors are shown in parenthesis. * p < 0.05 (two-tailed)

Table 3: Multiple Regression Models

| | Dependent Variable: Teacher's Value-Added | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All Subjects | | | Communications | | | English | | | Mathematics | | |
| Independent Variable | | | | | | | | | | | | |
| Students Enrolled | 0.329 | | (0.522) | -0.157 | | (0.630) | -0.112 | | (0.615) | 0.037 | | (0.919) |
| Male Teacher | 3.131 | | (1.908) | 4.432 | | (3.442) | -1.378 | | (3.777) | -0.471 | | (1.885) |
| Years Experience | 0.004 | | (0.048) | -0.021 | | (0.061) | -0.019 | | (0.061) | 0.121 | | (0.065) |
| Masters Degree | 0.441 | | (1.258) | 1.320 | | (1.586) | -2.641 | | (1.623) | 3.239 | * | (1.606) |
| Principal's Rank | 0.547 | | (0.331) | 0.150 | | (0.427) | 0.367 | | (0.439) | 0.207 | * | (0.100) |
| Constant | -9.500 | * | (2.429) | -8.788 | * | (3.010) | -6.416 | * | (2.984) | -6.334 | | (3.709) |
| R-Squared | 0.07 | | | 0.10 | | | 0.12 | | | 0.34 | | |

Notes: Standard errors are shown in parenthesis. ***** $p < 0.05$ (two-tailed)

# References

Attinello, J. R., Lare, D., & Waters, F. (2006, June). The Value of Teacher Portfolios for Evaluation and Professional Growth. NASSP Bulletin, Vol. 90 No. 2, 132-152. Retrieved January 27, 2008, from Wilson OmniFile, http://vnweb.hwwilsonweb .com.www2.lib.ku.edu:2048/hww/results/getResults.jhtml?_DARGS=/hww/results /results_common.jhtml.7#record_4

Berube, B, & Dexter, R. (2006, Summer), Supervision, Evaluation and NCLB: Maintaining a Highly Qualified Staff. *Catalyst for Change,* Vol. 34, No. 2, 11-17. Retrieved January 26, 2008, from Wilson OmniFile, http://vnweb.hwwilsonweb.com.www2.lib.ku.edu:2048/hww/results/getResults.jhtml?_DARGS=/hww/results/results_common.jhtml.11

Brophy, J. (April, 1992), Probing the Subtleties of Subject-Matter Teaching, *Educational Leadership,* Vol. 49, No. 7, 4-8. Retrieved, April 10, 2010, from EbscoHost, http://web.ebscohost.com.www2.lib.ku.edu

Cronin, J., Kingsbury, G., McCall, M., & Bowe, B. (2005, April). The Impact of the No Child Left Behind Act on Student Achievement and Growth. A technical report from the NWEA Growth Research Database. Retrieved October 16, 2007 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/00 00019b/80/1b/c8/98.pdf

Goldhaber, D. D. & Brewer, D. J. (1997, Summer). Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity. *The Journal of Human Resources*, Vol. 32, No. 3, 505-523. Retrieved April 2, 2006, from JSTOR, http://www.jstor.org.www2.lib.ku.edu.

Gordon, R., Kane, T. J., Staiger, D. O., (2006, April). The Hamilton Project: Identifying Effective Teachers Using Performance on the Job, The Brookings Institution. Retrieved September 16, 2007, from CSA Illumina, http://www.brookings.edu/views/papers/200604hamilton_1.pdf

Ingram, D., Louis, K. S., Schroeder, R. G., (June, 2004). Accountability Policies and Teacher Decision Making: Barriers to the Use of Data to Improve Practice. *Teachers College Record,* Vol. 106, No. 6, 1258-1287. Retrieved, April 10, 2010, from EbscoHost, http://web.ebscohost.com.www2.lib.ku.edu

Jacob, B. & Lefgren, L. (2006, Spring). When Principals Rate Teachers. *Education Next,* Vol. 6, No. 2, 58-64. Retrieved December 24, 2006, from Wilson OmniFile, http://vnweb.hwwilsonweb.comwww2.lib.ku.edu.

Kimball, S. M., White, B., & Milanowski, A.T. (2004). Examining the Relationships Between Teacher Evaluation and Student Assessment Results in Washoe County, *Peabody Journal of Education*, 79(4), 54-78. Retrieved February 7, 2010, from Wilson OmniFile, http://vnweb.hwwilsonweb.com.www2.lib.ku.edu:2048

Koedel, C., Betts, J. (2009, March). *Value-Added to What? How a Ceiling in the Test Instrument Influences Value-Added Estimation,* NBER Working Paper Series, No. 14778. Retrieved, January 30, 2010, from http://www.nber.org.www2.lib.ku.edu:2048/papers/w14778

KSDE (2010). Kansas State Department of Education. K-12 School Statistics. Retrieved, April 10, 2010, from http://www3.ksde.org

Leinhardt, G., (April, 1992). What Research on Learning Tells Us About Teaching, *Educational Leadership,* Vol. 49, No. 7, 20-25. Retrieved, April 10, 2010, from EbscoHost, http://web.ebscohost.com.www2.lib.ku.edu

McCaffrey, D. F., Lockwood, J. R., Koretz, D. Louis, T. A. & Hamilton, L. (Spring, 2004). Models for Value-Added Modeling of Teacher Effects, *Journal of Educational and Behavioral Statistics,* Vol. 29, No. 1, 67-101. Retrieved December 22, 2006, from Wilson OmniFile, http://vnweb.hwwilsonweb.com.www2.lib.ku.edu:2048

McCaffrey, D. F., Lockwood, J. R., Koretz, D. Louis, T. A. & Hamilton, L. (Spring, 2004). Let's See More Empirical Studies on Value-Added Modeling of Teacher Effects: A Reply to Raudenbush, Rubin, Stuart and Zanutto, and Reckase, *Journal of Educational and Behavioral Statistics,* Vol. 29, No. 1, 139-143. Retrieved February 7, 2010, from Wilson OmniFile, http://vnweb.hwwilsonweb.com.www2.lib.ku.edu:2048

Mendro, R. (1998). Student Achievement and School and Teacher Accountability, *Journal of Personnel Evaluation in Education,* Vol. 12, No. 3, 257-267. Retrieved February 18, 2010 from SpringerLink, http://www.springerlink.com.www2.lib.ku.edu:2048

Medley, D. M., (1977), *Teacher Competence and Teacher Effectiveness: A Review of Process-Product Research,* (Tables 6, 9, 16, 18, 31, and 34), Washington D. C.: American Association of Colleges for Teacher Education.

Medley, D. M., Coker, H., & Soar, R. S. (1984). *Measurement-Based Evaluation of Teacher Performance An Empirical Approach, 39-40.* New York: Longman, Inc.

Milanowski, A. (2004). The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati. *Peabody Journal of Education,* Vol. 79, No. 4, 33-53. Retrieved April 9, 2007, from http://vnweb.hwwilsonweb.com.www2.lib.ku.edu:2048/hww/share/shared_main.jhtml?_requestid=130796.

NWEA - Northwest Evaluation Association. Retrieved December 17, 2006, from http://www.nwea.org/assessments/.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004, Fall). How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis,* Vol. 26, No. 3, 237-257. Retrieved December 29, 2006, from Wilson OmniFile, http://vnweb.hwwilsonweb.comwww2.lib.ku.edu.

Olson, A., (2001, May/June). Data-based Change: Using Assessment Data to Improve Education. *Multimedia Schools, Vol. 8, No. 3, 39-43.* Retrieved October 16, 2007, from http://vnweb.hwwilsonweb.com.www2.lib.ku.edu:2048/hww/results/external_link_maincontentframe.jhtml?_DARGS=/hww/results/results_common.jhtml.9

Peterson, K., (2004, June). Research on Teacher Evaluation. *NASSP Bulletin,* Vol. 88, 60-79. Retrieved January 26, 2008, from Wilson OmniFile, http://vnweb.hwwilsonweb.com.www2.lib.ku.edu:2048/hww/results/getResults.jhtml?_DARGS=/hww/results/results_common.jhtml.12

Ponticell, J. A. & Zepeda, S. J.(2004, June). Confronting Well-Learned Lessons in Supervision and Evaluation. *NASSP Bulletin, Vol. 88, 43-59.* Retrieved January 26, 2008, from Wilson OmniFile. http://vnweb.hwwilsonweb.com.www2.lib.ku.edu:2048/hww/results/external_link_maincontentframe.jhtml;hwwilsonid=R3A5H2LLKC0YZQA3DIMSFF4ADUNGIIV0

Prothero, N., (2002, September/October), Improving Instruction Through Teacher Observation, *Principal*, (Reston, VA.) Vol. 82, No. 1, 48-51. Retrieved January 27, 2008 from Wilson OmniFile, http://vnweb.hwwilsonweb.com.www2.lib.ku.edu:2048/hww/results/getResults.jhtml?_DARGS=/hww/results/resultscommon.jhtml.12

Rockoff, J. E., (2004, May). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *The American Economic Review*, Vol. 94, No. 2, pp. 247-252. Papers and Proceedings of the One Hundred Sixteenth Annual Meeting of the American Economic Association San Diego, CA, January 3-5, 2004. Retrieved September 19, 2007, from JSTOR, http://www.jstor.org.www2.lib.ku.edu.

Saatcioglu, Argun, (2010). Notes from doctoral advisory meetings. Assistant Professor of Educational Leadership and Policy Studies, University of Kansas.

Sanders, W. L. & Rivers, J. C. (1996, November). *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement.* University of Tennessee Value-Added Research and Assessment Center. Retrieved March 30, 2006, from http://www.heartland.org/pdf/21803a.pdf.

Sergiovanni, T. J., (2001). *The Principalship: A Reflective Practice Perspective,* (4th Ed.), 221-242. Boston: Allyn and Bacon.

Stronge, J. H., Ward, R. P., Tucker, P. D., & Hindman, J. L., (2007). What is the Relationship Between Teacher Quality and Student Achievement? An Exploratory Study. *Journal of Personnel Evaluation in Education,* Vol. 20, 165-184. Retrieved February 18, 2010 from SpringerLink, http://www.springerlink.com.www2.lib.ku.edu:2048/content/w2j686n2mqu04v31/fulltext.pdf

Stufflebeam, D. L., Madaus, G. F., & Kellaghan, T., (2000). *Evaluation Models: Viewpoints on Educational and Human Services Evaluation,* (2nd Ed.), 263. Boston: Kluwer Academic Publishers.

Wilkerson, D. J., Manatt, R. P., Rogers, M. A., & Maughan, R. (2000). Validation of Student, Principal, and Self-Ratings in 360° Feedback® for Teacher Evaluation. *Journal of personnel evaluation in education,* Vol. 14, No. 2, 179-192. Retrieved April 11, 2007, from http://www.springerlink.com.www2.lib.ku.edu:2048/ content/j454684226128200/fulltext.pdf.

Woolridge, J. M., (2009). *Introductory Econometrics A Modern Approach,* (4th Ed.), Canada: South-Western Cengage Learning.

Wright, P., Horn, S., & Sanders, W., (1997). Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of personnel evaluation in education,* Vol. 11, 57-67. Retrieve February 20, 2010 from SpringerLink http://beta.springerlink.com.www2.lib.ku.edu:2048/content/l7q2242qnj2125w0/fulltext.pdf