Comparing Student Growth and Teacher Observation to Principal Judgments in the Evaluation of Teacher Effectiveness

Derek C. Briggs
Nathan Dadey
Ruhan Circi Kizil

University of Colorado
Center for Assessment, Design, Research and Evaluation

CADRE

October 1, 2014

Acknowledgments: We thank Allison Timberlake, Qi Qin and Melissa Fincher of the Georgia Department of Education for their invaluable assistance with this project.

Introduction

The validity of Georgia's Teacher Keys Effectiveness System (TKES) hinges on the assumption that evidence of student growth and ratings from direct observation of teacher practice can be used to distinguish teachers with respect to their efficacy. For Georgia teachers with students who take state-administered achievement tests, evidence of student growth comes in the form of *student growth percentiles* (Betebenner, 2009). A student growth percentile (SGP) for any specific test subject takes on a range from 1 to 99, and thereby ranks the performance of a student relative to peers across the state who had a similar history of test score performance in previous years[1]. Students with higher SGPs (e.g., above 50) are those who have done better than predicted on the current year/grade relative to their peers; students with lower SGPs (e.g., below 50) are those who have done worse than predicted. The inference to be made is that a student who has performed better/worse than comparable peers has demonstrated more/less academic growth. If the average student in a teacher's class tends to demonstrate performance on subject-specific tests that is above/below that of peers with similar prior academic achievement, it suggests that the quality of teaching the student experienced may have also been above/below average. This is formally quantified for each teacher in the TKES by taking the mean of SGPs (a "MeanGP") across students.

An important feature of SGPs is that they statistically adjust for differences in the prior achievement of students in any teacher's classroom. As a summary indicator of growth, an MeanGP is a fairer basis for comparing teachers than counting up the proportion of students who achieve a certain performance level by the end of the year. However, while an MeanGP does adjust for differences in students' prior achievement, there are many things it does not adjust for, and some of these may confound the interpretation of an MeanGP as an indicator of teacher effectiveness. For example, an MeanGP may not disentangle the efficacy of a teacher from contextual factors such as proportion of students in a classroom who are in poverty, English Language Learners, receiving special education services, are new to the school, etc.

For a primer on the student growth percentile methodology as it has been implemented in Georgia, see http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Georgia-Student-Growth-Model.aspx

Evidence of teacher efficacy related to direct observations of teacher practice come from the Teacher Assessment on Performance Standards (TAPS). The TAPS consist of 10 dimensions: professional knowledge, instructional planning, instructional strategies, differentiated instruction, assessment strategies, assessment uses, positive learning environment, academically challenging environment, professionalism, and communication. Each dimension is scored using a rubric with scores from 0 to 3, where a "0" represents the practices of an ineffective teacher, a "1" represents the practices of a teacher that "needs development," and "2" represent the practices of a teacher that is "proficient," and a "3" represents the practices of a teacher that is "exemplary." A total score is then computed as the sum of these 10 dimensions, and in theory, can range from 0 to 30.

The true efficacy of a teacher is impossible for a secondary analyst to observe directly, it can only be inferred. If it could be directly observed, then one could simply compare each teacher's MeanGP or TAPS score to this truth to determine how often they are in accord. In the absence of this, some external criterion of teacher efficacy is needed against which these indicators can be compared. In this study, we use principal judgment for this external criterion. With the possible exception of department chairs, principals are probably best-suited to provide this criterion because they are often directly responsible for hiring many of their school's teachers, enacting a schoolwide curricular plan, observing teachers in practice, engaging with them during the school year, and for getting feedback on them from parents and students alike.

In one of the only previous empirical studies that has relied on principal judgment to rate teacher efficacy, Jacob & Lefgren (2007) found that principals were quite good at identifying those teachers whose students demonstrated especially low or especially high achievement gains, even though they had much more difficulty distinguishing between teachers in between. In that study, Jacob & Lefgren used estimates of teacher effects from a value-added model as an objective criterion for high or low teacher efficacy and then examined to what extent principals were able to reach similar conclusions on the basis of subjective ratings. In the present study, a similar comparison is possible, but to some extent we turn this approach on its head by viewing principal judgments as the best available criterion for teacher efficacy, at least at the extremes. We then examine to what extent a MeanGP or a mean TAPS score would lead to inferences that are consistent with a principal's judgment.

This study differs from that of Jacob & Lefgren in two other important aspects. First, unlike Jacob & Lefgren we do not specify a value-added model in the sense that as a growth model, the SGP approach was not designed to disentangle or isolate the effect of a teacher from all other factors that could explain variability in test performance. Unlike a teacher effect estimate from a value-added model, an MeanGP is intended to have a correlational interpretation—if a teacher has a high/low MeanGP, this might be attributed to high/low teacher efficacy, but it might also be attributed to other factors. A wider range of evidence would need to be consulted before deciding whether a correlation can support a causal conclusion that could lead to high stakes consequences. To the extent that principals have internalized a wider range of evidence, their judgments are the appropriate check relative to inferences based solely on MeanGPs. Second, the data we gather for this study is an order of magnitude larger than the data that was available to Jacob & Lefgren. Their analysis was based on 220 teachers and 13 principals from one school district in a Western state prior to the implementation of a high stakes teacher evaluation system; our analysis is based upon a stratified probability sample of 12,619 teachers and 1,013 principals in 99 school districts currently in the process of implementing a state mandated teacher evaluation system.

We find evidence that both aggregate growth statistics (in the form of a MeanGP) and TAPS scores tend to be in accord with principal judgments about their most and least effective teachers. When asked to rate teachers according to the level of professional development support the teachers would need to have a strong positive impact on student achievement, principals categorized 39% of teachers as needing minimal support, 20% as needing maximum support, and the remaining 40% as falling somewhere in between. Teachers rated as needing minimal support have an MeanGP that is 6.1 percentile points higher than teachers rated as needing maximal support, and a mean TAPS score that differed by 2.2 points. Principals were also asked to select a single teacher that they thought was most successful at increasing student achievement, and a single teacher that they thought was least successful. Teachers rated as most successful had a mean

MeanGP that was 10.5 percentile points higher than teachers rated as least successful, and a mean TAPS score that was 3.5 points higher. When asked to explain the rationale for their choices, principals typically focused on the ability of teachers to use assessment data to differentiate instruction, establish rapport with their students, and manage their classrooms efficiently. Additional analyses suggest that mean MeanGP differences are largest for those who teach students in the subjects of math, social studies and science, and that these differences are insensitive to differences in a principal's years of experience at a school.

The remainder of this report is structured as follows. In the next section we describe the approach used to design and administer our principal survey, the characteristics of schools that were included in our sample, and the way that MeanGPs were computed. In the third section of the paper we present our overall findings that compare the mean MeanGPs of teachers by the rating category in which they were placed by principals. The fourth section of the paper examines the consistency of our overall findings by disaggregating our results by relevant teacher subgroups. The fifth section compares principal rating against teachers' rubric scores on Georgia's Teacher Assessment on Performance Standards (TAPS). The sixth section includes a summary of the written explanations that principals provided for their ratings. Finally, the last section concludes with the implications of our study for the use of MeanGPs and TAPS scores to evaluate teacher efficacy as part of Georgia's TKES.

Methods

Principal Survey

We developed a survey with items that would elicit principal judgments about differences in teacher efficacy. To this end we created a sequence of items linked to the roster of teachers specific to each principal's school. The first item on the survey prompted principals to rate teachers with respect to the level of support each of their teachers requires in order to have a positive impact on their students' academic achievement. The next two items prompted principals to pick the one teacher who they would regard as the most successful at increasing student achievement, and the one teacher who they would regard as the least successful. Finally, for about 1/4 of principals we also provided open-ended prompts asking them to describe the criteria that they used to pick the teacher who was most/least successful at increasing student achievement. (The text included in the full survey is provided in Appendix A.)

The idea for the level of support variable came from a conversation the lead author had with an elementary school principal at a Denver public school. In this conversation the lead author asked the principal about the sort of language that would be most likely to get principals to feel comfortable making distinctions about the ability of their teachers to improve student achievement. The principal suggested that a positive way to frame the question would be in terms of the level of professional development support each teacher would need. In the survey we define professional development by example:

> "All teachers can benefit from support in the form of professional development (PD) that helps them become better at their job. Examples of these kinds of PD supports might include:
>
> - Workshops offered at the district or school level
> - Presentations offered by professional speakers from outside the school
> - Periodic meetings in teacher teams during the school year
> - One-on-one coaching and feedback on teaching from a mentor or mentors
> - Taking coursework at an institution of higher education "

The level of support variable invokes some key ideas from the literature on formative assessment (c.f., Black & Wiliam, 1998), since one purpose of providing teachers with professional development is for this to help them improve the quality of their instructional practices. We hypothesized that the level of support teachers are provided through professional development can be characterized with respect to two dimensions: the frequency with which they are encouraged to participate in such activities, and the quality of the feedback that they receive from these activities. We define higher quality feedback as being individualized and targeted; lower quality feedback as being more general and targeted to the needs of a group. We combined these two dimensions into a single four point rating scale as follows

1. *The Teacher needs Infrequent Support with little or no Individualized Feedback*. Example: Teacher participates in PD opportunity offered once a year by district or school to all teachers regardless of grade/content specialization.
2. *The Teacher needs Frequent Support with little or no Individualized Feedback*. Example: Teacher participates in PD opportunities offered up to once a month by district or school, targeted to specific group of teachers by grade/content specialization.
3. *The Teacher needs Infrequent Support with Significant Individualized Feedback.* Example. Teacher participates in PD opportunities offered up to once a month by district or school, targeted to specific group of teachers by grade/content specialization, and includes individualized feedback and/or peer mentoring.
4. *The Teacher needs Frequent Support with Significant Individualized Feedback.* Example: Teacher participates in PD opportunities offered by district or school that are ongoing (multiple times a month), or takes coursework at an institution of higher education. The PD is targeted to the teacher's grade/content specialization and includes individualized feedback and/or peer mentoring. PD may also include meetings with school leadership.

We regard this as a quasi-ordinal scale in the sense that a rating of a 4 is meant to indicate a teacher whom a principal believes to require more support than a teacher with a rating of 1, 2 or 3, and a rating of a 1 indicates a teacher whom a principal believes to require less support than a teacher with a rating of a 2, 3 or 4, but it is not clear that a rating of 3 necessarily indicates a greater level of support than a rating of a 2. Of greatest interest are the MeanGP comparisons that would be subsequently made between teachers who were rated by principals as requiring the most (4) and the least (1) amount of professional support to be successful in increasing student achievement.

The level of support variable represents an indirect approach to get principals to reveal their perceptions about a teacher's ability to improve student achievement. We also took a direct approach by asking principals to pick one teacher that they regarded as most successful at improving student achievement, and one that they regarded as least successful. In a previous attempt to get principals to provide us with this information when conducting a similar study in a large urban district in another state, we had found the principals were often unwilling to name any teachers as "least successful." If a similar pattern held in Georgia, then we hoped to use the levels of support variable as an alternative. As it turned out, principals in Georgia were very forthcoming in their judgments about which of their teachers they perceived as least/most successful at improving student achievement. Nonetheless, the levels of support rating remained quite valuable because it allowed us to make MeanGP comparisons with the full roster of teachers at each principal's school, in contrast with the least/most variable which by definition restricted the sample to two teachers per school.

Beyond two short questions about principal experience, we asked no additional questions of principals beyond the ones described above to make it more likely that principals would be able to complete the survey within about 15 minutes. Before providing teacher ratings after logging into the Qualtrics survey environment, principals were informed about the purpose of the survey and that their responses would not be used to evaluate or make any high stakes decisions about teachers and schools. They were also informed that their participation was voluntary, and that this study was being conducted by researchers at the behest of the Georgia Department of Education, but not by staff affiliated with the Georgia Department of Education (GADOE).

Sample Characteristics

Although our principal units of analysis in this study are teachers, the primary sampling units for our survey were the 181 school districts in the state of Georgia. We initially divided these districts in quartiles as a function of their total student enrollment and then selected a stratified random sample of 90 school districts. Among these 90 were 17 districts that represented early "pilot" adopters of the TKES when Georgia was first awarded its RT3 grant. Another nine of these early adopter districts was not part of our initial stratified sample, but after consultation with GADOE staff we were asked to include these additional districts in the sample with certainty. A main reason for including these districts was that because they were further along in the TKES implementation process. (Later we check to see whether our results are sensitive to length of TKES implementation by restricting our sample to the pilot schools.) In total, our final sample included 99 school districts. All unique public schools within these districts were sent an invitation to participate in our survey.

There were a total of 1,394 schools in our sampled districts. The principals from 1,013 out of these 1,394 school completed the survey over a time period from May 13, 2014 through June 27, 2014, for a response rate of 73%. The high response rate can be attributed in large part to a highly coordinated and systematic follow-up process between CADRE staff, staff from the GADOE, and district-level coordinators. The median response time for principals was 6.7 minutes, and the mean was 10.8 minutes, indicative of a response time distribution skewed by a relatively small proportion of principals who appear to have taken more than 25 minutes to complete the survey.

Table 1 compares several relevant school-level characteristics for the sample of principals that responded to the survey relative to three other groups of schools: the full population of Georgia schools, schools that were not included in our stratified random sample, and schools that were included in the sample but had principals that did not complete the survey. In the average public school in Georgia, 51% of students are male, 62% of students are eligible for free or reduced price lunch services, 10% of students receive special education services, and 6% of students have limited English proficiency. Georgia's schools have a great deal of racial/ethnic diversity, with overall proportion of Black students (40%) roughly equal to the overall proportion of White (42%) students. There are also a significant proportion of Hispanic students (11%). The principals who completed our survey came from schools that were generally representative of schools in the state. Because the sample was stratified by district size, this led to an oversampling of large school districts, and so the schools in our sample have slightly higher proportions of Black students (43%), slightly lower proportions of White (39%) and Asian (2%) students, and slightly higher proportions of students eligible for free and reduced price lunch services (65%). With respect to the demographic characteristics of students in their schools, the principals who did not respond to our survey came from slightly less advantaged schools relative to principals who did respond—the average proportion of students eligible for free or reduced price lunches in non-responding schools was 68%.

Table 1. Characteristics of Schools in Sample Relative to Full Population of Schools

					Schools Included in Sample (N= 1394)			
	All Georgia Public Schools (N= 2195)		Schools Not Included in Sample (N=801)		Responded to Survey (N=1013)		Did Not Respond to Survey (N=381)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Male	51	5.3	51	5.9	51	4.7	51	5.5
Free/Reduced Lunch	62	26	56	27	65	25	68	24
Special Education	10	6	11	6	10	6.3	10	8
Limited English	6	10	5	9	7	11	6	11
Race/Ethnicity								
Asian	4	5	4	6	2	5	3	5
Black	40	31	30	28	43	31	49	34
Hispanic	11	14	12	13	12	14	11	15
Two or More Races	3	2	3	2	3	2	3	2
White	42	30	52	28	39	29	34	30

Note: Numbers in cells are all expressed in percents

Teacher Rosters

The 1,013 principals who responded to the survey were collectively provided with a total of 16,478 teachers to rate on our level of support variable. On average, this amounted to about 16 teachers per principal, but the median number of teachers rated was 12. Teachers were included on a school's roster if they were listed as the teacher of record for at least 15 students for whom SGPs were available collectively across one or more tested subjects. As a result, teachers in grades K-2 and those teaching students in specialized subjects for which no state tests were available were not part of the roster available for principals to rate. To avoid burdening principals in large schools, we truncated the number of teachers given to any principal to rate at 30. In total, there were 215 schools in which teacher rosters were truncated in this manner. At these schools, the mean and median number of teachers eligible for inclusion was 56 and 49 respectively. For any school with more than 30 teachers, we took a random sample from the full pool of eligible teachers.

Out of the original total of 16,478 teachers included on school rosters, principals indicated that 3,358 were either no longer at the school or that they had had no interaction with the teacher. For another 501 teachers, no rating was provided, and two others were removed because one had less than 15 students and the other was selected as both least and most successful by the same principal. This left us with a final sample of 12,617 teachers for whom a levels of support rating was available. From this sample, 94.8% of principals identified a single teacher as "most successful at increasing achievement," and 91.9% identified a single teacher as "least successful at increasing student achievement." This left us with a sample of 1,891 teachers flagged on the basis of our most/least successful variable. There were very few principals unable or unwilling to pick a teacher who was in their view "least successful" at increasing student achievement.

Computing Mean Student Growth Percentiles

A teacher's MeanGP was computed by taking the mean of all 2012-13 SGPs linked to that teacher across tested subjects. As of the 2013-14 school year during which the data for this study was gathered, students in Georgia public schools took state-administered Criterion Referenced Competency Tests (CRCTs) in the subject areas of mathematics, reading, English Language Arts (ELA), science and social studies from grades 3 through 8. As of grade 9 (i.e., high school), students shift from taking CRCTs that are specific to grade levels to taking End of Course Tests (EOCTs) that are specific to courses. The courses for which EOCTs exist are Mathematics I, Mathematics II, Coordinate Algebra, Georgia Performance Standards (GPS) Algebra, Analytic Geometry, GPS Geometry, United States History, Economics, Biology, Physical Science, Ninth Grade Literature and Composition, American Literature and Composition. Among the 12,617 teachers who were rated by principals with respect to the level of support variable, a MeanGP was computed on the basis of SGPs for students who may have taken tests in as many as five different content areas. In other words, if an elementary school teacher was listed as the teacher of record for students with SGPs available in math, science, social studies, ELA and reading, the teacher's MeanGP would be computed by taking the average of all these SGPs. The mean and SD of MeanGPs across all teachers in Georgia (including teachers not included on our survey rosters) was 49.7 and 10.6 respectively[2]. For the sample of teachers included in our study, this mean and SD was 49.4 and 10.3. Table 2 summarizes the frequency distribution of teachers who were linked to students with SGPs in unique content areas.

Table 2. SGPs from Unique Content Areas Used to Compute Teacher MeanGPs

Linked to Students with Unique SGPs in	Number of Teachers
1 Content Area	5,187
2 Content Areas	3,249
3 Content Areas	1,519
4 Content Areas	527
5 or More Content Areas[1]	2,135
TOTAL	12,617

Note: [1]There were a total of 40 teachers in high school grades linked to students with more than 5 unique SGPs. These teachers were intervention specialists.

In the analysis that follows, we compare principal ratings as a function of a teacher's overall MeanGP computed across all content areas, and then examine to what extent differences observed vary by MeanGPs in specific content areas.

All MeanGPs were computed using data on student growth from the 2012-13 school year. We apply three different approaches for computing these MeanGPs. In the first, a teacher's MeanGP is based on SGPs that have not been adjusted for measurement error. In the second, a teacher's MeanGP is based on SGPs that have been adjusted for measurement error using the SIMEX approach (Shang, Betebenner, van Iwaarden, in press). Finally, we also introduce a third method for computing MeanGPs that makes adjustments for contextual differences between teachers at the classroom level. In this approach, we use each teacher's observed MeanGP as an outcome variable and regress this on a series of teacher level variables as follows:

[2] The mean MeanGP is not exactly 50 because Georgia makes use of "baseline" referenced SGPs for many (but not yet all) of its test subjects. When available, we used baseline referenced SGPs in our computation of MeanGPs. These were not available for six EOCTs (GPS Algebra, GPS Geometry, Mathematics I, Mathematics II, Coordinate Algebra, Analytic Geometry). For these tests we use cohort referenced MeanGPs.

$$Y = b_0 + b_1 FRL\% + b_2 ELL\% + b_3 SWD\% + b_4 ACHIEVE + e.$$

In the regression equation above, *Y* represents a teacher's combined MeanGP, *FRL%* indicates the percentage of students associated with a teacher who are eligible for free or reduced price lunch services, *ELL%* indicates the percentage of students that are English language learners, *SWD%* indicates the percentage of students with disabilities (students with an individualized education place), and *ACHIEVE* represents students' mean prior grade achievement (computed after first standardizing all prior year subject-specific test scores to have mean of 0 and a standard deviation of 1). The last term in the model is an error term that is assumed to be independent of the included covariates and independent across teachers[3]. All four of the covariates included in the model above are examples of how classroom composition might differ in ways that could lead two equally strong teachers to face different challenges when it comes to increasing student achievement. These variables are meant to be illustrative rather than exhaustive; examples of other variables that could have been included would be racial/ethnic composition, attendance rates, student "churn" (students that enter and exit the classroom throughout the year), proportion of students in gifted and talented program, etc. Indeed, one challenge with this approach is that it can be unclear where one should stop in adding factors that need to be controlled.

After estimating the regression coefficients above, we then compute for each teacher a residualized MeanGP as $resMGP = Y - \hat{Y}$. This represents the amount by which a teacher's observed MeanGP is above or below the amount that would be predicted given the characteristics of the students in their classrooms. Note that by construction, *resMGP* will have a mean of 0 and will be uncorrelated with *FRL%*, *ELL%*, *SWD%* and *ACHIEVE.* To the extent that some teachers are more effective or ineffective with the various classroom compositions defined by these variables, and to the extent that they are purposefully sorted to these sorts of classrooms, this approach is likely to overadjust (c.f., Ballou, Sanders, & Wright, 2004), removing variability in

[3] This assumption is surely violated by the clustering of teachers within schools and school districts. However, because our sample sizes are so large in this regression context, involving the full population of teachers in the state, producing cluster-adjusted standard errors would have no impact on conventional tests of statistical significance.

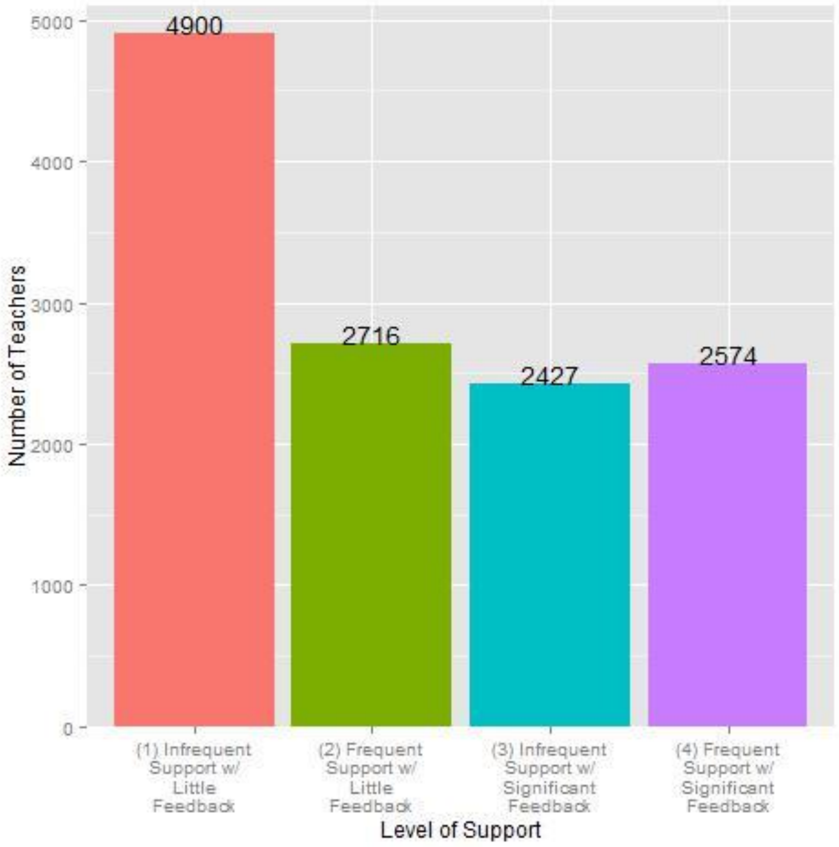
observed MeanGPs that might reasonably be attributed to heterogeneity in teaching quality.

Is Student Growth Evidence Consistent with Principal Judgments?

Levels of Support Ratings

Principals made some clear distinctions among their teachers with respect to the level of support variable, as shown in Figure 1. Principals categorized 39% of teachers as those that would need the lowest level of support in order to have a strong positive impact on their students' achievement, 20% that would need the highest level, and 41% that would fall somewhere in between.

Figure 1. Frequency Distribution of Principal Ratings on Level of Support Variable



We find only a small difference between the teachers placed into these level of support categories with respect to years of experience. As Table 3 indicates, the mean years of experience for teachers thought to require the least amount of support was 13.6, while the mean for teachers thought to require the most support was 12.7.

Table 3. Level of Support Rating by Years of Teaching Experience

Level of Support Rating	Years of Experience				
	N	Mean	SD	Min	Max
1 (Infrequent, General Feedback)	4901	13.6	7.8	0	44
2 (Frequent, General Feedback)	2716	12.8	8.2	0	42
3 (Infrequent, Individualized Feedback)	2427	12.8	8.0	0	46
4 (Frequent, Individualized Feedback)	2575	12.7	8.3	0	43

Principals subsequently selected 960 and 931 of these 12,617 teachers as most and least successful at increasing student achievement. Among the 960 teacher selected as most successful, 69% had been rated as requiring minimal support, 25% had been rated as requiring an intermediate level of support, and 6% had been rated as requiring maximal support. Among the 931 selected as least successful, 68% had been rated as requiring maximal support, 27% had been rated as requiring an intermediate level of support, and 5% had been rated as requiring minimal support. This indicates a fairly strong association between the level of support variable a least/most successful rating.

Comparing Level of Support Ratings by MeanGPs

Figure 2 and Table 4 compare the MeanGPs of 12,617 teachers combined across subjects by each level of support category. There is a notable and statistically significant difference of 6.1 percentile points between the mean MeanGPs of the teachers in lowest and highest support categories. Expressed relative to an SD of the MeanGP distribution, this represents an effect size of 0.59 SDs. There are also statistically significant differences between the mean MeanGPs of teachers needing the least and most support relative to teachers in the middle categories, but these differences are smaller. In summary, those teachers who principals rate as needing more professional development support also tend to be teachers with students demonstrating relatively low growth in academic achievement.

Figure 2. Boxplots of MeanGP Distributions by Level of Support Rating

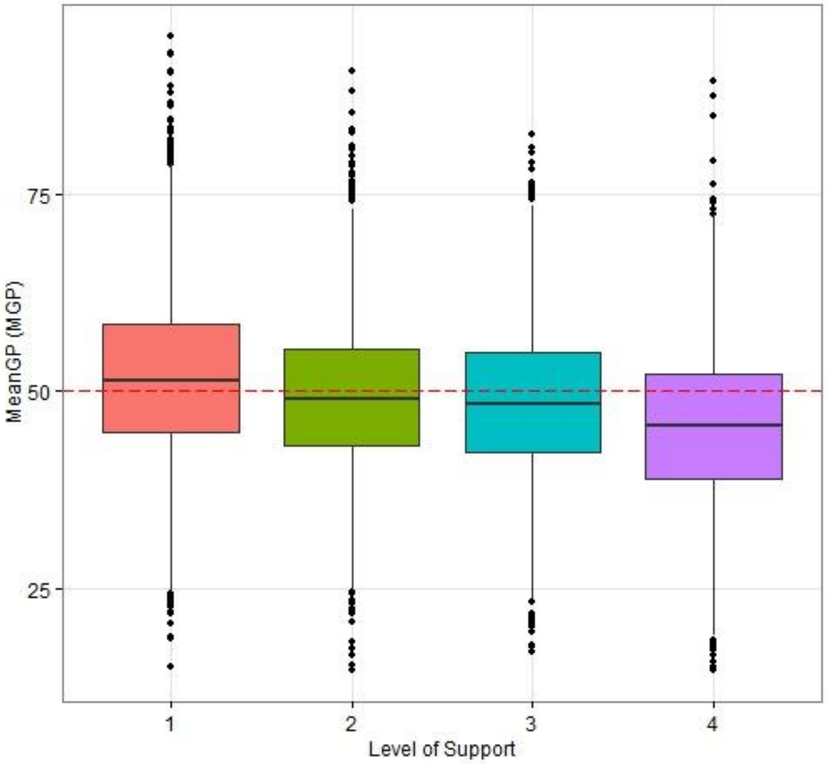


Table 4. Mean MeanGP by Level of Support Rating

Level of Support Needed	Teacher 2012-13 MeanGP		
	Mean	SD	SE
1 (Low)	51.8	10.2	0.15
2	49.4	10.0	0.19
3	48.6	9.6	0.20
4 (High)	45.7	10.1	0.20

Comparing MeanGPs for Teachers Rated as Least/Most Successful at Increasing Student Achievement

Now we focus on the subset of 1,891 teachers whom principals rated as either the least or most successful at increasing student achievement. Figure 3 and Table 5 summarize the main results, which indicate a large and statistically significant difference of 10.5 percentile points between the average MeanGPs of teachers rated by their principals as most vs. least successful. This difference is equivalent to a full SD of the MeanGP distribution. Figure 3 shows the smoothed MeanGP distributions of the two groups. There is a clear rightward shift for the teachers selected by their principals as most effective relative to those selected as least effective. However, notice that there are also numerous examples of teachers rated by their teachers as most successful who have an MeanGP that is relatively low, and vice-versa. In other words, while the MeanGP of the average teacher in this sample tends to be in accord with a principal's judgment of efficacy, there are still exceptions to the rule.

Figure 3. MeanGP Distributions by Effectiveness Rating

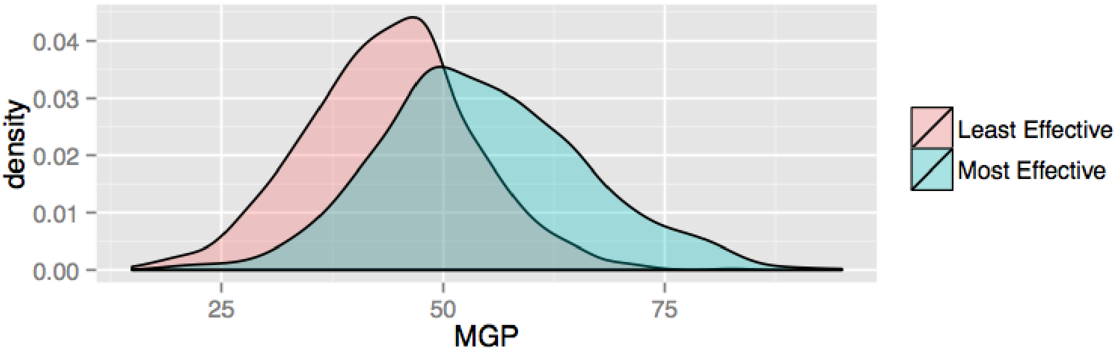
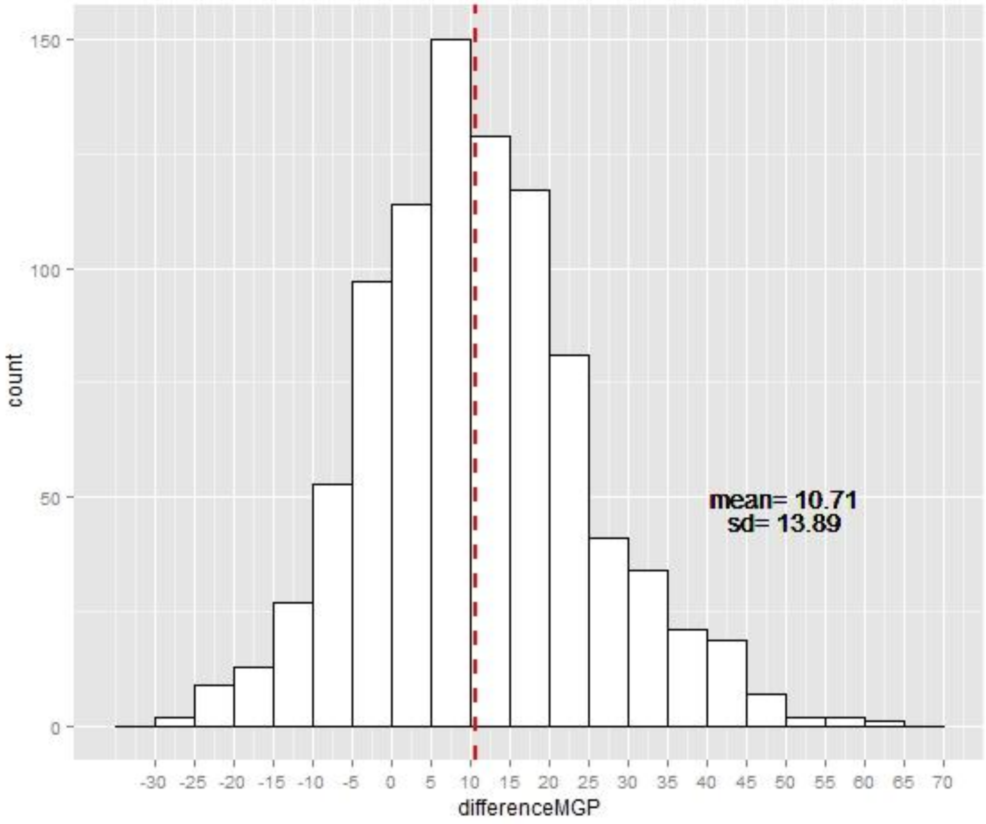


Table 5. Summary Statistics of MeanGP Distributions by Effectiveness Rating

Principal Rates Teacher as	N	2012-13 MeanGP Mean	SD
Most Effective	960	54.4	11.7
Least Effective	931	43.9	9.5
Difference		10.5	

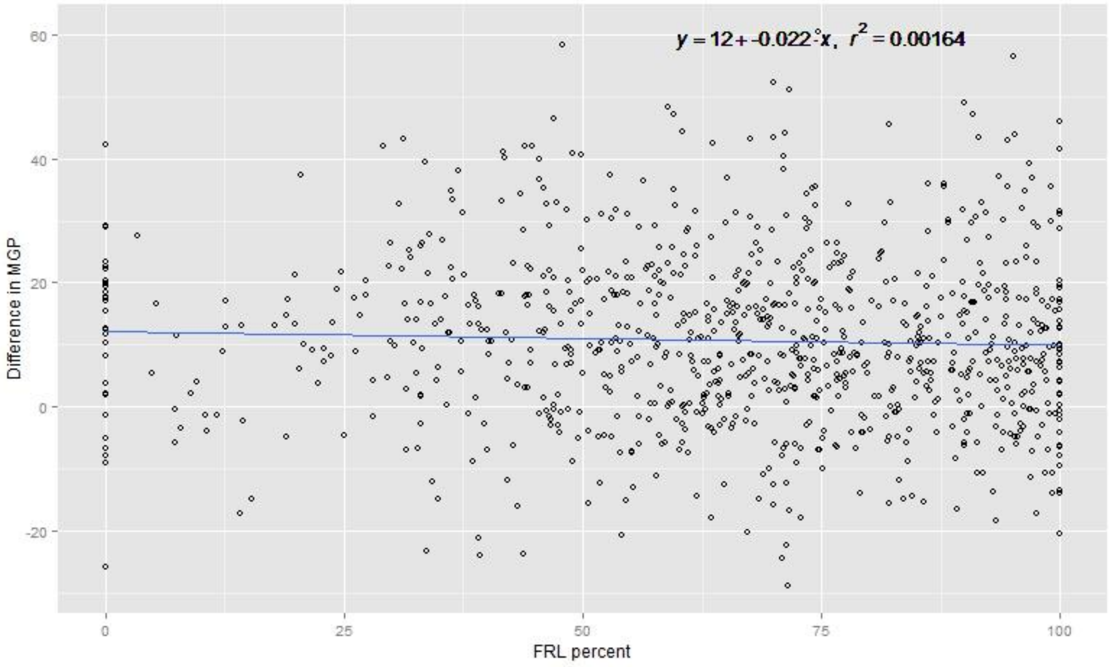
To what extent is their variability in the difference between the meanGPs of teachers rated as least/most successful *by the same principal*? To answer this question we restrict our sample of principals to the 919 who provided a rating for both a least and most successful teacher, and the compute the difference in meanGP for each principal. Figure 4 below shows the histogram of the meanGP difference; the dashed red line indicates the location of the mean (10.7). Here we can see there is in fact considerable variability in these meanGP differences (SD = 13.9). More than 75% of the time, the meanGPs of teachers identified as most successful at raising student achievement were greater than those of teachers identified as least successful, often substantially greater (for 25% of principals, the difference was greater than 18.7 percentile points. However, for about 20% of principals, the difference was negative—the teacher identified as least successful had an MGP that was higher than the teacher identified as most successful.

Figure 4. Histogram of the Distribution of MeanGP Differences from Principals who Rated Teachers as Least and Most Successful at Increasing Student Achievement



One threat to the interpretation of these results is the possibility that at some high performing schools, even the least successful teacher might have a meanGP that is bigger than the meanGP of the most successful teacher at a low performing school, and vice-versa. To examine this possibility more closely using schools as the units of analysis, we regressed the most/least successful teacher difference in meanGPs on the percent of students in the school eligible for free and reduced price lunches (FRL%). If high-performing schools also tend to have wealthier students in attendance (on average), and if higher quality teachers are attracted to these schools, then the difference in meanGPs should have an association with FRL%. The regression line superimposed on the scatterplot shown in Figure 5 shows no sign of this association. For a small subset of schools with fewer than 25% of students eligible for FRL services, there is some indication that the spread in MeanGP differences is a bit smaller than that observed in schools with greater than 25% of students eligible. However, overall we see little evidence of any functional relationship between MeanGP differences for schools that appear to differ in terms of the wealth/resources of their students' households.

Figure 5. The Relationship Between MeanGP Differences for Teachers Rated as Least/Most Successful by School Level FRL%



A Closer Examination by Subgroups

Up to this point we have presented results for teachers irrespective of grade level with MeanGPs that were combined across multiple test subjects when available. In what follows we examine whether the differences in mean MeanGPs by principal ratings are bigger or smaller for the subset of pilot districts who had had more experience implementing the TKES, for certain grade levels, and for certain test subjects. We also explore whether these differences vary by a principals' years of experience at a given school. Finally, we compare differences after using two methods for computing a teacher's MeanGP that attempt to account for factors that might bias these values up or down. The results from this full set of new comparisons are summarized in Table 7.

Comparisons Disaggregated by RT3 Pilot Districts

The RT3 pilot districts consist of 853 schools. Out of these schools, 598 (70%) had principals that responded to our requests to participate in the survey (recall that the overall response rate for all sampled schools was 73%). The principals at these schools rated 7,461 teachers on the level of support variable and 1,139 on the least/most successful variables (N= 578 for most successful, N= 561 for least successful). The differences in mean MeanGPs were very similar to that found for the full sample of teachers. The mean difference on the level of support variable was 5.5, and on the most vs. least successful variable the difference was 10.7. The provides additional support to the notion that principals were not simply rating teachers on the basis of their knowledge of teachers' MeanGPs. Because these principals had the most experience interpreting the elements of the TKES, they would have been most likely to have internalized MeanGPs as a key factor in evaluating teacher efficacy, and would have also had more experience rating teachers on the basis of the TAPS rubrics. Yet the mean MeanGP differences associated with these principal ratings are not significantly bigger than those from principals with less experience implementing the TKES.

Comparisons Disaggregated by Test Subject

There were 17 unique subject-specific tests that could serve as the basis for a teacher’s MeanGP. Here we focus on the five subject-specific tests that are given to all Georgia students in grades 3 through 8 to see whether there is evidence that mean MeanGP differences vary significantly by test subject. A first noteworthy finding is that the correspondence between principal ratings and teacher MeanGP is considerably stronger in the test subjects of math, science and social studies relative to the subjects of ELA and reading. For teachers associated with students who took tests in math, science or social studies, the differences in mean MeanGPs for teachers rated low vs. high on the level of support variable was 7.2 (science), 8.1 (math) and 8.3 (social studies); for teachers rated most vs. least successful the difference was 9.8 (science), 11.8 (math), and 14.2 (social studies). In contrast, for teachers associated with students who took tests in ELA and reading, the difference in mean MeanGPs for teachers rated low and high on the level of support variable was 3.7 (ELA) and 3.6 (reading), and for teachers rated most vs. least successful the difference was 5.1 (ELA) and 4.3 (reading).

There are at least two possible interpretations of these results. One is that when principals are thinking about the level of support their teachers require, or when asked to pick teachers who are least/most effective, that they are not typically doing so with a teacher’s expertise in ELA and reading in mind. The other is that MeanGPs are less useful as a way to distinguish the efficacy of teachers in these subject areas. Both explanations may be correct, but without asking principals to rate teachers by distinct content areas, they cannot be disentangled on the basis of the data collected for this study. However, this finding is consistent with evidence in the value-added modeling literature that there is much less variability in the distribution of teacher effects in reading relative to mathematics, making the distinctions to be made even at the extremes of the distribution rather small (Jacob & Lefgren, 2007; Briggs & Weeks, 2011). A somewhat surprising result is that a teacher’s MeanGP appears to provide the greatest degree of discrimination for teachers who teach students in social studies.

Table 7. Consistency of Mean MeanGP Differences by Principal Ratings

	Level of Support Rating		Most vs. Least Successful Rating	
	Teachers Compared	Max vs. Min Support	Teachers Compared	Most vs. Least Successful
Overall	12619	6.1	1891	10.5
RT3 Pilot Districts	7461	5.5	1139	10.7
By Test Subject				
English Language Arts	3366	3.7	998	5.5
Reading	3367	3.6	999	4.3
Math	2969	8.1	982	11.9
Science	2605	7.2	859	10.8
Social Studies	2788	8.3	933	13.8
By Number of Subjects to Compute MeanGP				
One	3005	7.4	591	14.4
Two	1937	4.8	431	8.6
Three	938	4.3	258	8.7
Four	327	3.7	93	8.7
Five	1236	7.0	513	8.7
By Grade Level				
Elementary School (3-5)	3138	5.6	1154	8.4
Middle School (6-8)	2582	7.4	417	14.5
High School (9-12)	1782	5.4	324	12.6
By Principal Experience at School				
3 years or less	3488	5.5	936	9.8
Between 4 and 7 years	2426	7.1	567	11.3
More than 8 years	1104	6.2	264	11.1
By Method of Computing MeanGPs				
MeanGP	7474	6.1	1891	10.5
sMeanGP	7474	5.4	1891	10.1
resMGP	7474	4.8	1891	9.4

Note: sMeanGP indicates a MeanGP that has been adjusted for measurement error using the SIMEX approach; resMPG indicates a MeanGP that has been adjusted for differences in a teachers' classroom context.

Although a majority of teachers (60%) were associated with students who had taken tests in two or more test subjects, the rest (40%) were associated with students with test scores in just one subject[4]. For this latter subset of teachers, differences in mean MeanGPs by level of support ratings and most vs. least successful ratings were 7.4 and

[4] To be clear, this does not mean that these students only took a test in one subject, only that the teacher was only linked to the test scores in one subject as the teacher of record.

14.4, differences that are greater than those found for teacher associated with students taking multiple test subjects. This is especially dramatic for the most vs. least successful rating, where the mean difference for teachers with multiple subject areas is about 8.7, almost 6 percentile points lower than the mean MeanGP difference for teachers associated with a single test subject. This further suggests that principals seem to be picking least or most successful teachers with a single subject area in mind, and that when a combined MeanGP is computed, this may depress the MeanGP difference that best captures the differences in teacher efficacy that correspond to principal ratings.

Comparisons Disaggregated by Grade Level

We also examine mean MeanGP differences for teachers by the grade level of their students. We group grades into elementary (grades 4-5), middle (grades 6-8) and high (grades 9-12) bands. The most notable result here is a strong interaction with the middle school band. Among the 2,582 teachers rated by principals in middle school grades, the differences in mean MeanGPs for teachers rated low and high on the level of support variable was 7.4, and for the subset of 417 teachers rated least vs. most successful the difference was 14.8. The differences for high school teachers was similar, but differences for teachers in elementary grades was significantly smaller, 5.6 on the levels of support variable, 8.6 on the least vs. most successful variable. Again, this is likely an artifact of the subject specialization—teachers in elementary school grades are typically associated with students taking test in multiple subjects relative to middle and high school teachers who are associated with students taking a single test subject.

Comparisons Disaggregated by Principal Experience at School

It would be reasonable to speculate that principals with more years of experience at a school would be better able to make distinctions among their teachers. More experienced principals have had more opportunities to observe teachers and gather information relevant to the evaluation of their efficacy. To explore this, we divided principals into three groups: those with three years of experience or less (N=488), those with between 4 and 7 years of experience (N=292) and those with 8 years or more (N=144). We do see some evidence to support the notion of a small interaction with principal experience: differences in average MeanGPs for teachers with respect to our two rating variables was 0.5 to 1 percentile points lower. However, these differences are rather minor. Average MeanGP difference are typically in accord with ratings whether a principal is relatively inexperienced, moderately experienced or very experienced.

Comparisons Disaggregated by Method of Computing MeanGPs

Finally, we compare mean MeanGP differences as a function of different methods that could be taken to computer a teacher's MeanGP. In particular, we focus on two alternatives to the baseline referenced MeanGPs we have used in all ratings comparisons up to this point: an MeanGP corrected for measurement error using the SIMEX method which we refer to as an sMeanGP, and an MeanGP adjusted for classroom context covariates which we refer to as resMeanGP.

The computation of the resMeanGP variable is based upon a teacher level regression using all 39,148 teachers in Georgia with an MeanGP based on at least 15 students. Table 8 presents the coefficient estimates for each of the classroom context covariates that were included in the teacher-level regression. These coefficients have been transformed so that they can be interpreted in column 2 as the change in MeanGP that would be predicted for a 1 SD increase in the independent variable, holding constant the values of the other independent variables. In column 3, this increase is expressed relative to a 1 SD increase in the overall MeanGP.

Table 8. Teacher Level MeanGP Regressions

Independent Variable	Change in MeanGP per 1 SD	As proportion SD(MeanGP)
FRL%	-1.51	-0.14
ELL%	0.57	0.05
SWD%	-1.10	-0.10
ACHIEVE	1.99	0.19
R^2	11.5%	
N	39,148	

Note: Dependent Variable = MeanGP, SD = 10.6.
All regression coefficients statistically significant at $p < .01$. The actual SDs of each independent variable: SD of FRL% = 28, SD of ELL% = 12, SD of SWD% = 24, SD of ACHIEVE = 1. The SD of the dependent variable, MeanGP = 10.3

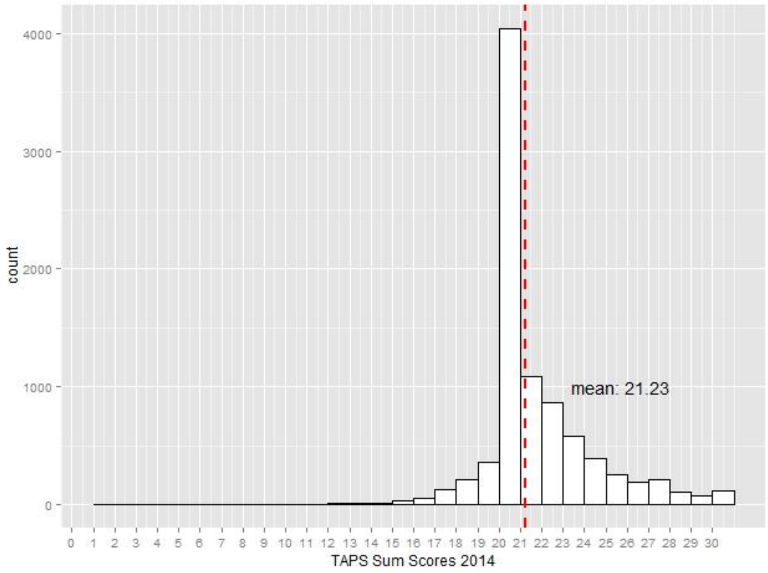
The regression coefficients indicate that differences in the mean prior achievement of a teacher's students have the biggest impact on teacher MeanGPs, followed by the percentage of students eligible for free and reduced price lunches, the percentage of students with disabilities, and percentage of English Language Learners. A 1 SD increase in ACHIEVE is predicted to increase a teacher's MeanGP by about 2 percentile points. In contrast, a 1 SD increase in FRL% (equivalent to 28 percentage points) or SWD% (equivalent to 24 percentage points) is predicted to *decrease* a teacher's MeanGP by about 1.5 and 1 percentiles. Interestingly, a 1 SD increase in ELL% (equivalent to 12 percentage points) is predicted to *increase* a teacher's MeanGP, though by a very small amount (0.6 percentiles).

The use of either sMeanGP or resMeanGP in place of MeanGP has an interesting impact on comparisons of teachers who were rated on opposite ends of the level of support variable or the least/must successful variable. Namely, the use of sMeanGP reduces the average MeanGP difference on the level of support variable from 6.1 to 5.4, and reduces the difference on the least/most successful variable from 10.5 to 10.1. This indicates that a small portion of the observed difference in average MeanGPs can be explained by bias due to measurement error. The use of resMeanGP reduces the average MeanGP difference on the level of support variable from 6.1 to 4.8, and reduces the difference on the least/most successful variable from 10.5 to 9.4. This indicates that a portion of the differences principals appear to perceive about their teachers may well be biased by differences in their classroom contexts. In other words, principals are either slightly more likely to view teachers in classroom with advantaged and/or high-achieving classrooms as effective, or they are somewhat more likely to assign effective teachers to advantaged or high-achieving classrooms. When MeanGPs are adjusted for these contexts, it reduces the average MeanGP difference on the level of support variable by about 21%, and on the least/most successful variable by about 10%.

Are TAPS Scores Consistent with Principal Judgments?

For this analysis we use the 2013-14 TAPS scores and ratings as a point of comparison for our principal survey. Here the key question is to what extent principal ratings of teachers on the level of support variable and as least/most successful at increasing student achievement have some association with the ways these same teachers where scored relative to the TAPS observation rubrics. Naturally, one would expect a teacher rated by a principal as requiring maximal support or being least successful at increasing student achievement to also have TAPS scores that are below average. We restrict attention to a teacher's total TAPS score as well as their classification into four performance categories on the basis of this total score.

Figure 6. Distribution of TAPS Scores for Teachers Rated in Principal Survey



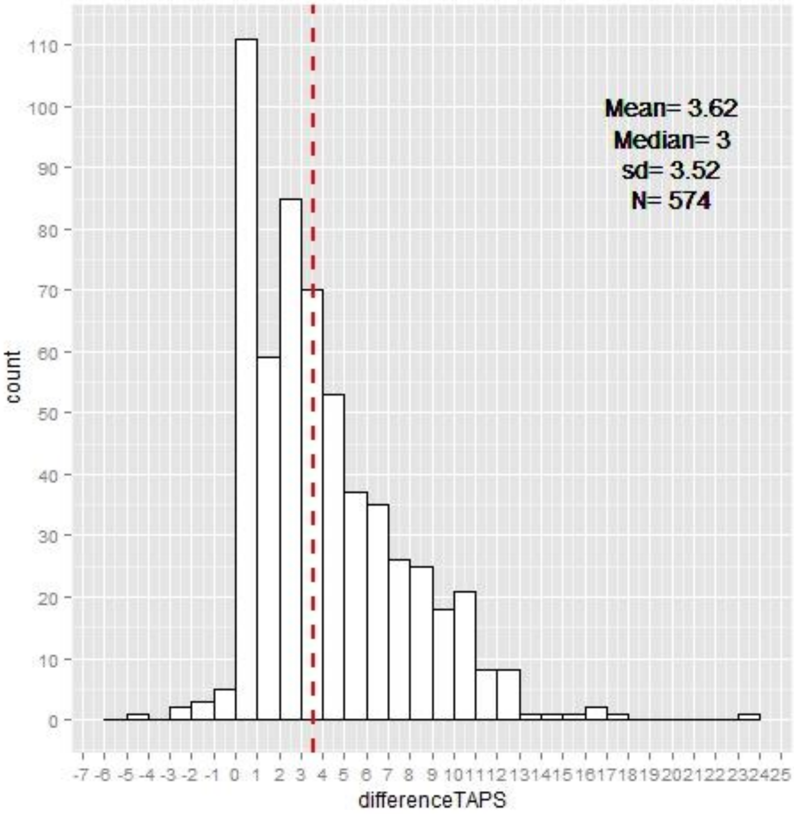
There were a total of 8,761 teachers for whom we had both TAPS scores and a principal rating on the level of support variable. The histogram in Figure 6 shows the distribution of these TAPS scores. The mean and modal score is 21 and a full 50% of all teachers have scores between 20 and 22. About 90% of teachers have TAPS scores between 17 and 26. The SD of TAPS scores is 2.6 points. Table 9 compares the distributions of TAPS scores by a teacher ratings on the level of support variable. Teachers rated as requiring minimal vs. maximal support had a mean TAPS score of 22.1 and 19.9 respectively, which teachers requiring some support sitting in between at about 21. Table 9 also shows the mean TAPS scores for the subset of 618 and 742 teachers who were rated by principals as least and most successful respectively. Those teachers rated as least successful had a mean TAPS score of 19.4, while those rated as most successful had a mean of 22.9, for a difference of 3.5 points on the TAPS raw score scale. Although this difference is rather small with respect to the full range of the possible score scale from 0 to 30, relative to the SD of 2.6, this represents an effect size of 1.3.

Table 9. Principal Ratings by TAPS Scores

Principal Rating		Teachers' Total TAPS Score			
	N	Mean	SD	Min	Max
Level of Support Variable					
1 (Infrequent, General Feedback)	3415	22.1	2.6	16	30
2 (Frequent, General Feedback)	1940	21.1	2.2	11	30
3 (Infrequent, Individualized Feedback)	1645	20.8	2.3	12	30
4 (Frequent, Individualized Feedback)	1761	19.9	2.5	2	30
Least/Most Successful Teacher					
Least Successful	618	19.4	?	4	28
Most Successful	742	22.9	?	16	30

To examine this more carefully, we restrict attention to the 574 principals who selected two teachers as those who were least/most successful at increasing student achievement. We then compute for each case the difference in the total TAPS scores observed for the two teachers within the same school. The resulting distribution is shown in Figure 7, with the mean of 3.6 indicated by the dashed red line. In general then, TAPS scores do offer some ability to discriminate between teachers who appear to differ in their efficacy, but the difference relative to the full TAPS score scale is relatively small.

Figure 7. Difference in TAPS Score for Teachers rated as Least vs. Most Successful at Raising Achievement in the Same School.

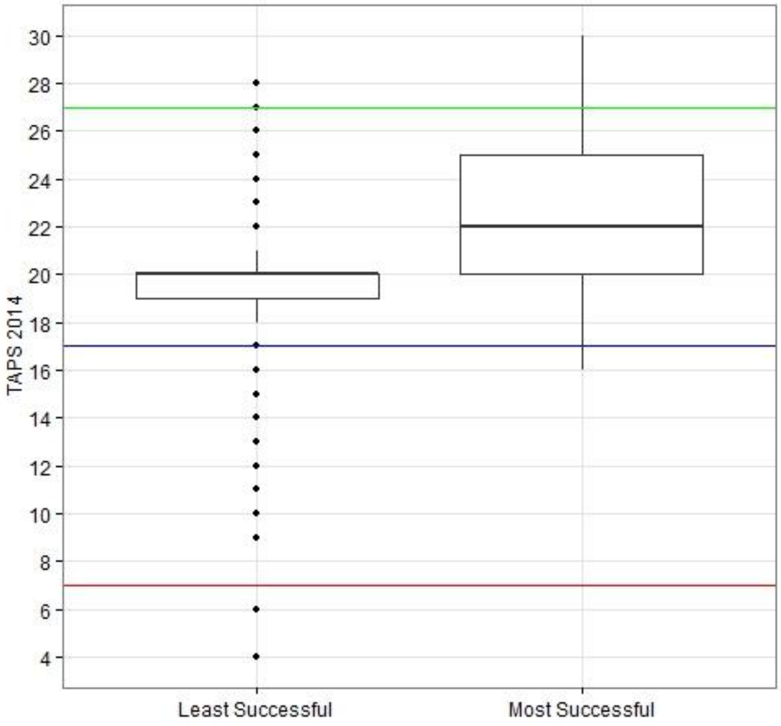


One possible concern about the TAPS scores is that the key demarcation between teachers rated by principals as needing maximal vs. minimal support and those rated as least vs. most successful seems to be at the TAPS total score threshold of 20. Yet teachers are placed into performance levels based on their TAPS score as follows:

- 0-6 = “Ineffective”
- 7-16 = “Needs Development”
- 17-26 = “Proficient”
- 27-30 = “Exemplary”

Importantly, the score band from 17-26 encompasses almost the full population of teacher ratings, and would place into the same category of "Proficient" teachers that principals typically rate as both least and most successful at increasing student achievement. It may well be the case that the thresholds chosen above were made on the basis of criterion-referenced expectations. However, the evidence here suggests some disconnect between these thresholds and principal judgments. To help place this into stark relief, the boxplots in Figure Z compare the TAPS score distributions for teachers rated as least/most successful at increasing student achievement. Horizontal lines have been superimposed at the thresholds of 7, 17 and 27. Again, notice that this would encompass almost all of teachers rated as least successful by their principals, and more than 75% of teachers rated as most successful. This serves to obscure any differentiation of teachers on the basis of their TAPS scores.

Figure 8. Boxplots of TAPS Score by Teachers Rated as Least or Most Successful at Increasing Student Achievement.



Note: The dots in the least successful boxplot indicate outliers. Horizontal lines represent demarcations between performance levels.

What Rationales Do Principals Provide for their Ratings?

So far these results suggest that on average, information about a teacher's MeanGP or TAPS scores will be in accord with the judgments of principals if they were asked to name the teachers in their schools who (1) needed the different levels of PD support, and (2) are the least and most successful at increasing student achievement. Because teacher MeanGPs are based on 2012-13 data, and TAPS scores are based on observations during the 2013-14 school year, this implies that principal ratings could incorporate new information about the teacher's job performance between the release of MeanGPs in the fall of 2013, the observations of principals during the 2013-14 school year, and the administration of our survey in the spring of 2014. A threat to the validity of this finding is that when asked to rate teachers, some principals could have directly referenced each teacher's subject-specific MeanGPs and/or TAPS scores. If this were frequently the case, then the findings here would be largely tautological: the most successful teachers would have higher MeanGPs and TAPS scores than the least successful teachers because this was the basis for many least/most successful distinctions in the first place.

To evaluate this possibility, for about 1/4 of our surveyed principals we inserted open-ended prompts that asked them, following their least/most successful rating, "Why did you identify this teacher as least/most successful in increasing student academic achievement?" A total of 262 and 258 principals provided us with these written rationales. We created four dichotomous variables to categorize these rationales. First, we look for rationales in which principals explain their choice with respect to their perceptions of a teacher's content knowledge, classroom practice, rapport with colleagues at the school, and other characteristics that represent factors that could explain *why* a teacher would be more or less effective at increasing student achievement. These are all rationales that are consistent with the criteria that are the basis for TAPS ratings. Second, we look for rationales in which principals explain their choice with respect to evidence they have gathered over time about student test outcomes. Within this latter category, we further distinguish between principals who explicitly reference evidence about student growth within a school year, and principals who explicitly mention SGPs, or who describe the concept of a student growth percentile even if they don't reference the actual term.

A premise of this study was that when asked to rate teachers with respect to those who are least/most effective at increasing student achievement, principals would do so on the basis of multiple sources of evidence. In particular, we expected principals to make holistic judgments on the basis of both "inputs" (characteristics of the teacher that the principal views as a prerequisite for having an effect on student achievement) and "outputs" (historical trends in the achievement of students assigned to the teacher). If, in contrast, principals only focus on outputs, and further, if they focus primarily on evidence of student growth made available to them through Georgia's growth model, then the results of our overall analysis presented above could be much more equivocal.

Table 10 summarizes the rationales they provided for their choices. Although it was certainly the case that a sizable proportion of principals reference student test performance as a basis for selecting teachers as most or least successful at increasing student achievement (46% and 32% respectfully), it was relatively rare for principals to make a direct link between student achievement and growth—at least in the way that growth is defined by a student growth percentile. In what follows we illustrate representative responses from principal rationales for their choices of most and least successful teachers, in that order. (For a larger sample of principal responses, selected in proportion to the frequencies found overall, see Appendix B.)

Table 10. Summary of Principal Responses to Question “Why did you identify this teacher as least/most successful in increasing student academic achievement?”

Principal's Rationale for Selecting Teacher	Most Successful	Least Successful
Describes Teaching Practices (Input)	71%	83%
References Student Test Outcomes (Output)	46%	32%
Mentions Student Growth within Year	18%	10%
Mentions SGP as basis for rating	4%	3%
Response References Both Inputs & Outputs	18%	16%
Response Only References Inputs	53%	67%
Response Only References Outputs	28%	17%
Number of Principal Responses	262	258

Rationales for “Most Successful” Teacher Choices

When principals did mention test outcomes as a basis for a least/most successful rating, it was typically mentioned with respect to “pass rates” on CRCTs or EOCTs.

> “All of [name removed]’s students passed the math, ELA, reading, social studies CRCT with a 90 to 100 percent met or exceeded rate.”

> “She had a high percentage of her students pass the EOCT.”

> “Low level of student achievement each year as measured on standardized test.”

Among those principals that referenced student test outcomes as a rationale, many appeared to interpret “least/most successful in increasing student achievement” to mean teachers for whom there was an upward or downward trend in the percent of students “passing” their CRCT or EOCT over time.

> “[name removed] is able to adjust instruction to meet the needs of diverse learners. This year the students' mean percentage of meeting and exceeding on the CRCT increased by 21 percent.”

For principals that did invoke “student growth” in their responses (18% of the time for most successful rationales, 10% for least successful), it was often not clear what they meant. For example “her students always make growth” and “Her students’ scores show significant increase each year” were representative responses. A small proportion (about 4%) did explicitly or implicitly reference results from the state growth model, either mentioning an SGP by name or referring to it in terms of the bubble charts provided by the GADOE.

> “Her test scores are always high and on the statewide longitudinal data system her bubbles are in the green and upper right quadrant.”

In all, it was relatively infrequent for a principal to provide a rationale for selecting a teacher as most successful that focused exclusively on outputs in the form of student test score outcomes without any mention of input characteristics. This happened just 28% of the time in rationales provided for choosing a teacher.

It was more common for principals to provide rationales that emphasized input characteristics of the teachers. When principals provided rationales for teachers they selected as most successful, 71% of the time they pointed to characteristics of what they considered good pedagogy, and 53% of the time this was all that they included in their rationale for selecting the teacher:

> “The teacher builds a positive rapport with the students. She studies and finds various research based instructional strategies to help the students succeed in Mathematics. The teacher models, ask questions, and allows the students to work on their own or in a group setting.”

> “She teaches to provide students with an understanding of the standard, how it relates to the real world and uses higher order thinking skills. She refers back to the standard during the lesson and provides exemplars for the students to self assess. She uses formative assessments regularly during the lesson and adjust her

> instruction based on the feedback she receives from the students. She uses small group instruction where she differentiates the instruction to the individual student."

> "Highly reflective; changes teaching approaches if something's not working or could work better; great mentor teacher."

> "[name removed] knows her content. She studies and is very prepared for each lesson she teaches. She introduces the lessons to students in a variety of ways. Her lessons are engaging and fun. She follows up the lessons with engaging activities and she conferences with her students on a regular basis as well as with their parents. She has a close relationship with her student's parents and with her class."

> "This teacher annually demands the best from her students and herself. She refuses to accept excuses or to allow students to fail. She builds relationships with parents and students, going beyond what is expected during the school day to make home visits and tutor students. She is respected by parents and students for her firm but consistent demeanor."

Some recurring themes were the ability of the most successful teachers to establish rapport with their students, have high expectations for students, differentiate instruction, make data driven decisions using formative assessment, and in general, a willingness to go "above and beyond."

Rationales for Least Successful Teacher Choices

Principals were even more likely to point to input characteristics when explaining why they selected a teacher as least successful. They did this 83% of the time overall, and in 67% of responses this was all that they referenced.

> "She does not use data to drive her instruction. She insists upon using the Review for the Test method. While the students do pass the test, I believe it is not due to

> her efforts completely. Most of the students she taught had CRCT scores that indicated that they would be successful. However, there were [also] CRCT scores for students that indicated that they may struggle. There was no RTI or differentiation. And, those students who were targets for possible failure--- were indeed not successful. The level of growth was not sufficient. Also the rigor in the classes was not present in both reading and writing."

> "I believe that [name removed] is not as successful increasing student academic achievement due to her not being a very good classroom manager and due to her not investing much time in preparing for her students."

> "Student academic achievement data demonstrated that the teacher's influence on their learning was very low. His classroom management was deplorable and not conducive to student learning, and he was reluctant to participate in professional development activities that generally helped teachers improve their job performance in ways that positively affected student academic achievement."

> "She refuses to change her instructional practices. She feels that her experience affords her the right not to do what it takes in order to meet the demands of educating students in the 21st century."

> "She is a bit lazy and scattered. She lacks focus and passion. She always has an excuse."

In many of these responses we see the inverse of the rationales provided for the most successful teachers. The least successful teachers are those that do not have a rapport with their students, do not have high expectations, who are unable or unwilling to differentiate instruction, who do not use assessment for formative purposes, and who generally only do the minimum of what is required of them. In addition, a new factor that emerged rather clearly in these responses was the importance principals placed on classroom management skills, something that was almost never mentioned as a notable feature among teachers selected as most successful. The implication seems to be that classroom management is regarded as necessary but not sufficient for a teacher to be effective. Lastly, in explaining why teachers were selected as least successful, principals were much more likely to bring up perceived character flaws (e.g., lazy, stubborn) or to note mediating factors (e.g., poor health).

In summary, we conclude that very few principals were selecting teachers as least or most successful at increasing student achievement solely on the basis of information at their disposal about student growth percentiles. This is not to suggest that principals were typically picking teachers without an awareness of student test performance. To the contrary, it seems likely that even when they did not include it in their written rationales, many principals are aware of the student test score trends of the teachers they had selected. And in any given year, teachers with students who perform well above or below average on achievement tests are also likely to have student growth percentiles that are above or below average. This probably explains some portion of the agreement between principal ratings and MeanGPs. On the other hand, our analysis of their written responses also provides support for the assumption that most Georgia principals were making holistic judgments on the basis of both inputs and outputs when asked to rate their teachers, and often it appears that the inputs (observations of teacher practices) were the driving force behind a principal's rating.

Implications

The results from this study show that both aggregated student growth percentiles and total TAPS scores are associated with principal judgments about teacher efficacy. Specifically, when restricted to distinctions principals make between teachers who require minimal vs. maximal PD support, or teachers who are the least or most successful at increasing student achievement, we find practically and statistically significant differences in the mean MeanGPs of teachers. The differences in mean TAPS scores are smaller, but also statistically significant. This can be taken as evidence in support of the validity of student growth percentiles and TAPS scores as a basis for inferences about teacher efficacy. However, it is important to appreciate the limitations of this finding relative to the context of this study. When restricted to teachers who appear to be at the tails of a hypothetical efficacy distribution, principals, student growth percentiles and TAPS scores are likely to converge. The results from this study cannot be used to argue that MeanGPs or TAPS scores are well-suited for making distinctions between teachers near the middle of this hypothetical distribution.

One actionable result from this study is the finding that the TAPS thresholds used to designate teachers as "Needing Development," "Proficient," or "Exemplary" do not appear to be well-aligned to the principal judgments elicited by our survey. Specifically, the range of TAPS scores that define the Proficient category essentially encompasses the full population of teachers that principals rated as either least or most successful at increasing student achievement.

A limitation of this study is that we did not ask principals to provide detailed ratings on multiple aspects that could be used to characterize teacher competencies. For example, in their survey Jacob & Lefgren (2007) asked principals to rate teachers with respect to eight characteristics that were frequently mentioned by our sample of principals in their open-ended responses: dedication and work ethic, organization, classroom management, role model for students, student satisfaction with teacher, parent satisfaction with teacher, positive relationship with colleagues, positive relationship with administrators. In addition, in that survey principals distinguished between teachers thought to be effective in math and teachers thought to be effective in reading. Although we very consciously decided not to ask these more detailed questions in order to maximize principal participation, having these additional responses would have allowed for more fine-grained analyses than we were able to conduct here.

On the other hand, one strength of our survey was that it gave us the opportunity to let principals express their perspectives on extremes of teacher efficacy in their own words. In doing so, it becomes evident that just as there is variability in teacher efficacy, there may be just as much variability in principal efficacy. Principal rationales for the teachers they selected as most or least effective varied both in terms of the quantity and quality of evidence that was referenced. A majority of principals focused primarily on input characteristics of teachers. When principals did focus on output characteristics, there was little sign that they were attuned to within grade evidence of student growth. This may indicate the need for concerted professional development to ensure that principals know how to interpret student growth percentiles and the aggregate statistics derived from them, and how they differ from other indicators of student progress.

References

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–65.

Betebenner, D. (2009). Norm- and Criterion-Referenced Student Growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

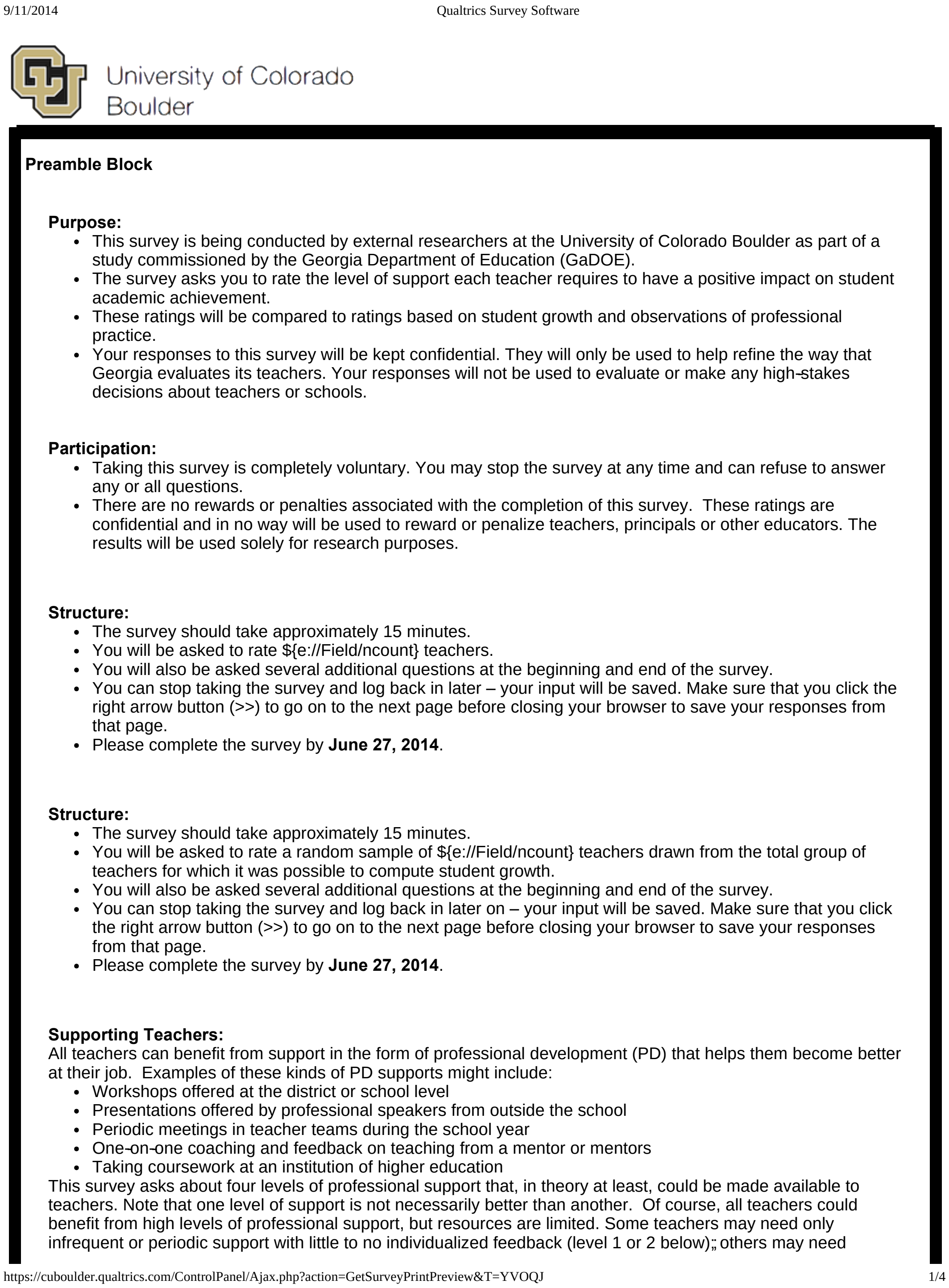
Black, P., & Wiliam, D. (1998). Inside the Black Box : Raising Standards Through Classroom Assessment By Paul Black and Dylan Wiliam. *Phi Delta Kappan*, *80*, 139–144, 146–148.

Briggs, D. C., & Weeks, J. P. (2011). The Persistence of School-Level Value-Added. *Journal of Educational and Behavioral Statistics*.

Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, *26*(1), 101–136.

Shang, Y., Betebenner, D. & Van Iwaarden, A. (in press). Covariate Measurement Error Correction for Student Growth Percentiles Using the SIMEX Method. *Educational Measurement: Issues & Practice*.

APPENDIX A: Principal Survey



9/11/2014 Qualtrics Survey Software

University of Colorado Boulder

Preamble Block

Purpose:

- This survey is being conducted by external researchers at the University of Colorado Boulder as part of a study commissioned by the Georgia Department of Education (GaDOE).
- The survey asks you to rate the level of support each teacher requires to have a positive impact on student academic achievement.
- These ratings will be compared to ratings based on student growth and observations of professional practice.
- Your responses to this survey will be kept confidential. They will only be used to help refine the way that Georgia evaluates its teachers. Your responses will not be used to evaluate or make any high-stakes decisions about teachers or schools.

Participation:

- Taking this survey is completely voluntary. You may stop the survey at any time and can refuse to answer any or all questions.
- There are no rewards or penalties associated with the completion of this survey. These ratings are confidential and in no way will be used to reward or penalize teachers, principals or other educators. The results will be used solely for research purposes.

Structure:

- The survey should take approximately 15 minutes.
- You will be asked to rate ${e://Field/ncount} teachers.
- You will also be asked several additional questions at the beginning and end of the survey.
- You can stop taking the survey and log back in later – your input will be saved. Make sure that you click the right arrow button (>>) to go on to the next page before closing your browser to save your responses from that page.
- Please complete the survey by **June 27, 2014**.

Structure:

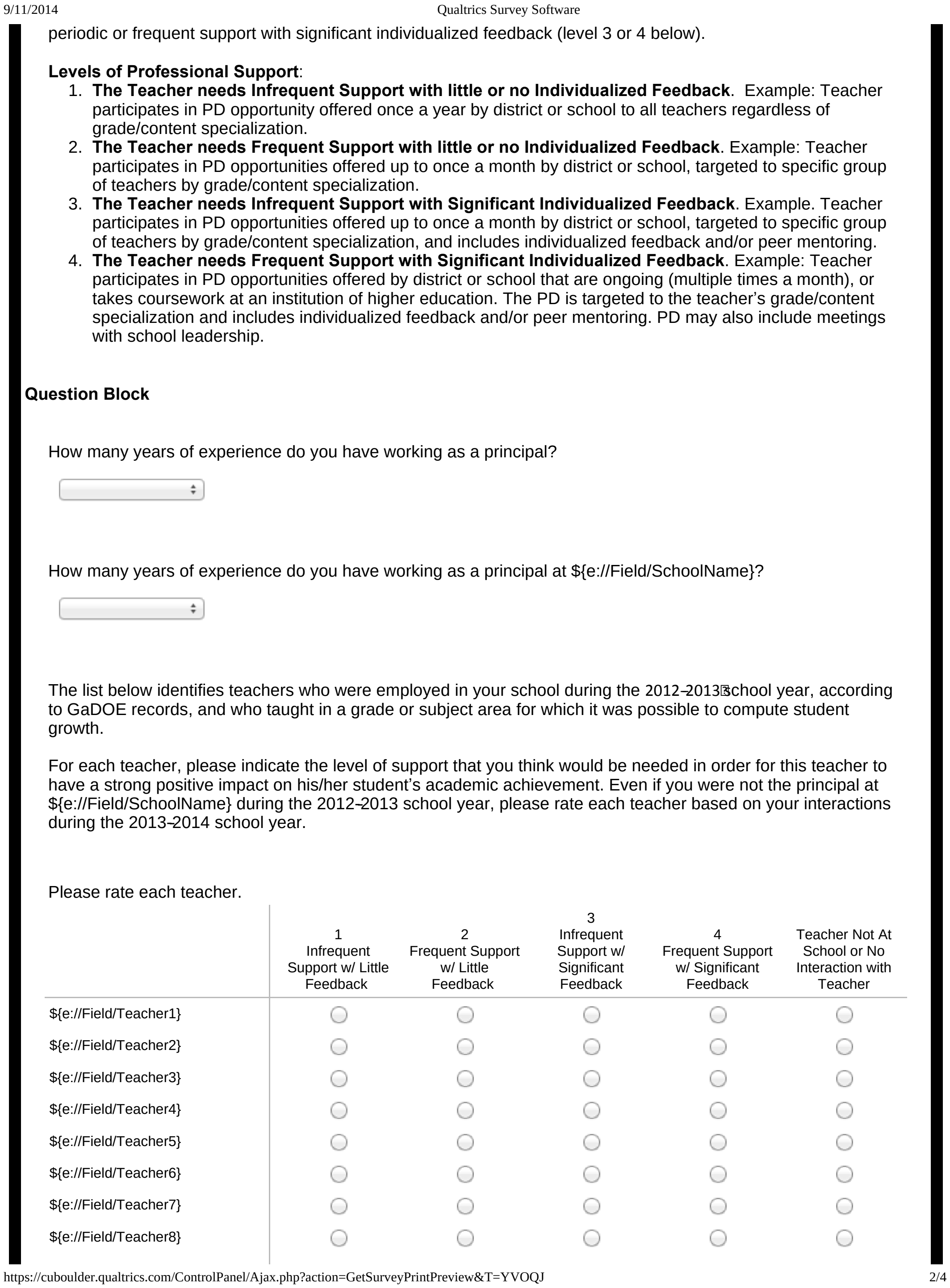
- The survey should take approximately 15 minutes.
- You will be asked to rate a random sample of ${e://Field/ncount} teachers drawn from the total group of teachers for which it was possible to compute student growth.
- You will also be asked several additional questions at the beginning and end of the survey.
- You can stop taking the survey and log back in later on – your input will be saved. Make sure that you click the right arrow button (>>) to go on to the next page before closing your browser to save your responses from that page.
- Please complete the survey by **June 27, 2014**.

Supporting Teachers:
All teachers can benefit from support in the form of professional development (PD) that helps them become better at their job. Examples of these kinds of PD supports might include:

- Workshops offered at the district or school level
- Presentations offered by professional speakers from outside the school
- Periodic meetings in teacher teams during the school year
- One-on-one coaching and feedback on teaching from a mentor or mentors
- Taking coursework at an institution of higher education

This survey asks about four levels of professional support that, in theory at least, could be made available to teachers. Note that one level of support is not necessarily better than another. Of course, all teachers could benefit from high levels of professional support, but resources are limited. Some teachers may need only infrequent or periodic support with little to no individualized feedback (level 1 or 2 below);; others may need

https://cuboulder.qualtrics.com/ControlPanel/Ajax.php?action=GetSurveyPrintPreview&T=YVOQJ 1/4



periodic or frequent support with significant individualized feedback (level 3 or 4 below).

Levels of Professional Support:

1. **The Teacher needs Infrequent Support with little or no Individualized Feedback**. Example: Teacher participates in PD opportunity offered once a year by district or school to all teachers regardless of grade/content specialization.
2. **The Teacher needs Frequent Support with little or no Individualized Feedback**. Example: Teacher participates in PD opportunities offered up to once a month by district or school, targeted to specific group of teachers by grade/content specialization.
3. **The Teacher needs Infrequent Support with Significant Individualized Feedback**. Example. Teacher participates in PD opportunities offered up to once a month by district or school, targeted to specific group of teachers by grade/content specialization, and includes individualized feedback and/or peer mentoring.
4. **The Teacher needs Frequent Support with Significant Individualized Feedback**. Example: Teacher participates in PD opportunities offered by district or school that are ongoing (multiple times a month), or takes coursework at an institution of higher education. The PD is targeted to the teacher's grade/content specialization and includes individualized feedback and/or peer mentoring. PD may also include meetings with school leadership.

Question Block

How many years of experience do you have working as a principal?

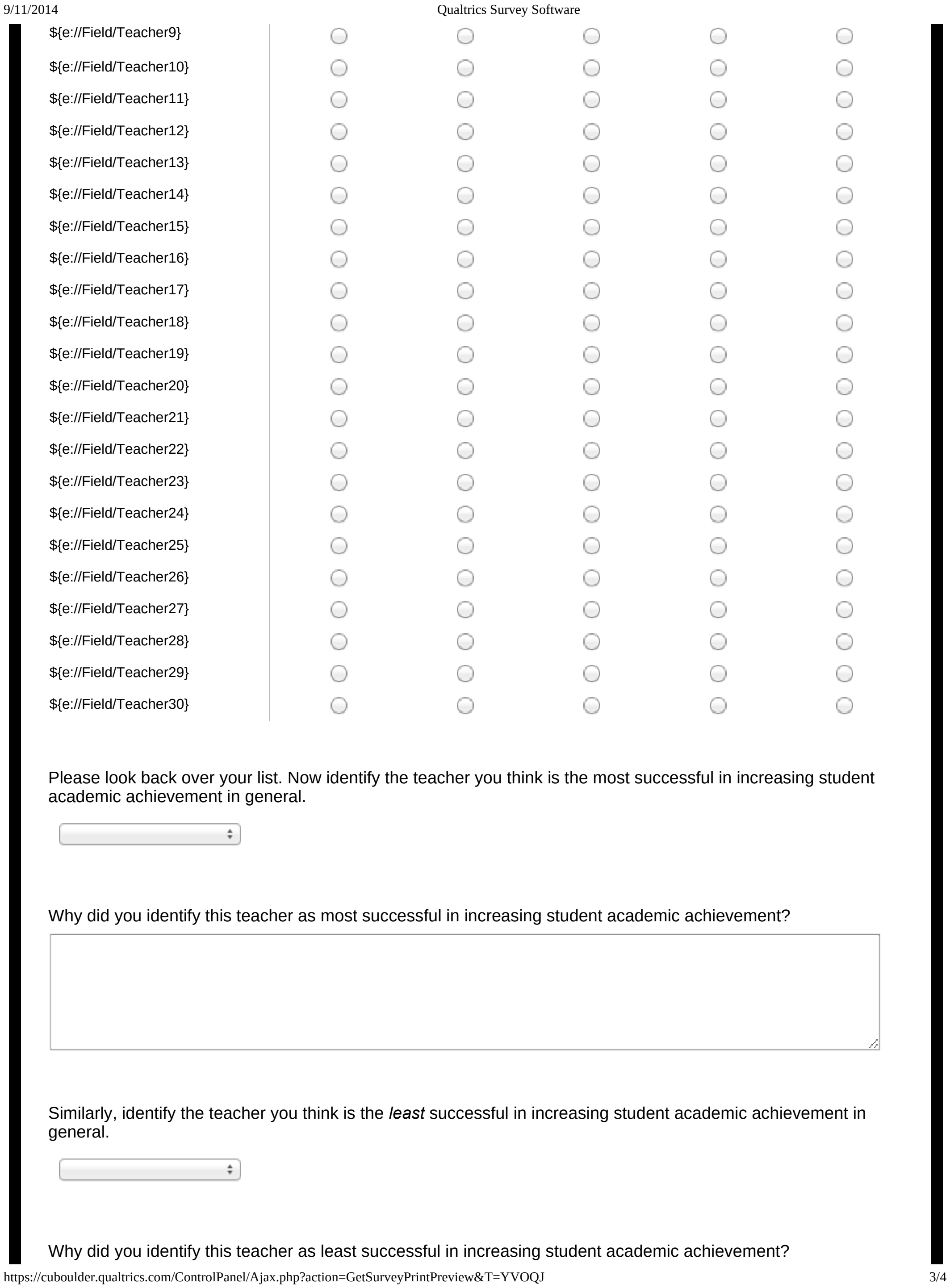
How many years of experience do you have working as a principal at ${e://Field/SchoolName}?

The list below identifies teachers who were employed in your school during the 2012‐2013school year, according to GaDOE records, and who taught in a grade or subject area for which it was possible to compute student growth.

For each teacher, please indicate the level of support that you think would be needed in order for this teacher to have a strong positive impact on his/her student's academic achievement. Even if you were not the principal at ${e://Field/SchoolName} during the 2012‐2013 school year, please rate each teacher based on your interactions during the 2013‐2014 school year.

Please rate each teacher.

	1 Infrequent Support w/ Little Feedback	2 Frequent Support w/ Little Feedback	3 Infrequent Support w/ Significant Feedback	4 Frequent Support w/ Significant Feedback	Teacher Not At School or No Interaction with Teacher
${e://Field/Teacher1}	○	○	○	○	○
${e://Field/Teacher2}	○	○	○	○	○
${e://Field/Teacher3}	○	○	○	○	○
${e://Field/Teacher4}	○	○	○	○	○
${e://Field/Teacher5}	○	○	○	○	○
${e://Field/Teacher6}	○	○	○	○	○
${e://Field/Teacher7}	○	○	○	○	○
${e://Field/Teacher8}	○	○	○	○	○



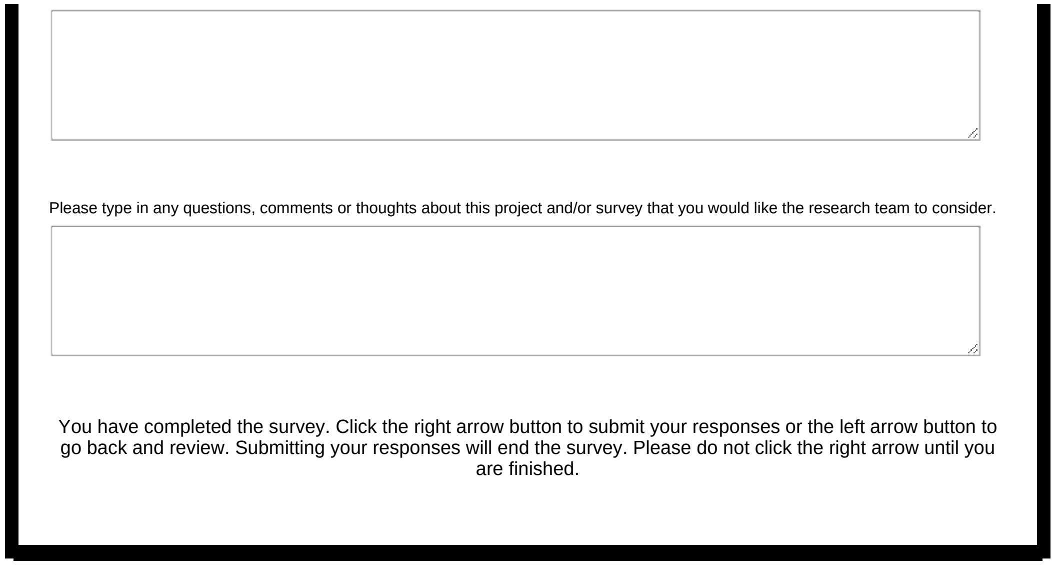
${e://Field/Teacher9}	○	○	○	○	○
${e://Field/Teacher10}	○	○	○	○	○
${e://Field/Teacher11}	○	○	○	○	○
${e://Field/Teacher12}	○	○	○	○	○
${e://Field/Teacher13}	○	○	○	○	○
${e://Field/Teacher14}	○	○	○	○	○
${e://Field/Teacher15}	○	○	○	○	○
${e://Field/Teacher16}	○	○	○	○	○
${e://Field/Teacher17}	○	○	○	○	○
${e://Field/Teacher18}	○	○	○	○	○
${e://Field/Teacher19}	○	○	○	○	○
${e://Field/Teacher20}	○	○	○	○	○
${e://Field/Teacher21}	○	○	○	○	○
${e://Field/Teacher22}	○	○	○	○	○
${e://Field/Teacher23}	○	○	○	○	○
${e://Field/Teacher24}	○	○	○	○	○
${e://Field/Teacher25}	○	○	○	○	○
${e://Field/Teacher26}	○	○	○	○	○
${e://Field/Teacher27}	○	○	○	○	○
${e://Field/Teacher28}	○	○	○	○	○
${e://Field/Teacher29}	○	○	○	○	○
${e://Field/Teacher30}	○	○	○	○	○

Please look back over your list. Now identify the teacher you think is the most successful in increasing student academic achievement in general.

Why did you identify this teacher as most successful in increasing student academic achievement?

Similarly, identify the teacher you think is the *least* successful in increasing student academic achievement in general.

Why did you identify this teacher as least successful in increasing student academic achievement?



Please type in any questions, comments or thoughts about this project and/or survey that you would like the research team to consider.

You have completed the survey. Click the right arrow button to submit your responses or the left arrow button to go back and review. Submitting your responses will end the survey. Please do not click the right arrow until you are finished.

APPENDIX B: Examples of Principal Open-Ended Survey Responses

Note: Any references to specific names of teachers have been removed and some responses have been shortened to keep the identity of principal and teacher anonymous.

Illustrative Rationales Provided By Principals for Teachers Rated Most Successful

Input Characteristics Only [54% of responses]

1. The teacher was able to identify areas of weaknesses and develop plans/activities which addressed those weak areas. She is very creative, data driven and always thinking of ways to make the connection.

2. --- uses a variety of strategies to enhance instruction such as technology, self directed learning and small group activities.

3. ---- is very reflective. She takes every opportunity to learn something new or refine tried and true practices. [She and] I have worked on her implementation of small flexible groups during her reading instruction. While this idea is certainly not new, she has sought to improve her use of flexible groups in an increasingly rigorous manner. She he has taken constructive feedback through her TKES evaluations and other informal observations seriously as a result, her instruction has improved. She takes every bit of feedback to heart, reflects on it, analyzes it, and acts upon it (666)

4. The teacher builds a positive rapport with the students. She studies and finds various research based instructional strategies to help the students succeed in Mathematics. The teacher models, ask questions, and allows the students to work on their own or in a group setting. (667)

5. She premeditates what she teaches to provide students with an understanding of the standard, how it relates to the real world and uses higher order thinking skills. She refers back to the standard during the lesson and provides exemplars for the students to self assess. She uses formative assessments regularly during the lesson and adjust her instruction based on the feedback she receives from the students. She uses small group instruction where she differentiates the instruction to the individual student. (679)

6. Challenges the students with a rigorous learning environment. She uses student assessment data to drive instruction. She has incorporated cross-curricular lessons within her teaching strategies. (682)

7. ---- is an EIP teacher who teaches every child at their skill level. He finds multiple ways to get the information across to the students and stops at nothing to help them. (689)

8. She teaches, plans and searches for extra resources to make lessons full of information and interesting. She interjects technology, labs, guest speakers, nonfiction books and hands on activities to make learning real. (695)

9. ----- maintains relationships with his students and is an inclusion teacher. His work and planning with the special needs teacher gives him strategies to help all students in the classroom. He is organized, orderly and respected by the students. (696)

10. The student becomes very involved in his/her own learning an takes ownership. This teacher has a gift of making this subject become relevant and students become excited about learning while being challenged at the highest level. (723)

11. At least 4 of the teachers listed above could have just as easily been identified as the teacher most successful in increasing student academic achievement because they are all apart of the same grade level and all provide the same level of support to students to include: consistent data monitoring, goal setting with students, driven desire for professional growth, consistent planning, creating informal and formal assessments on a weekly basis to drive instruction, communication with parents, building relationships with students, consistent team planning, and a driven desire to support students. (739)

12. Highly reflective; changes teaching approaches if something's not working or could work better; great mentor teacher (743)

13. This teacher successfully builds a great relationship with her students from the beginning of each year. She plans effective, differentiated lessons with careful thought put (757)

14. ---- knows her content. She studies and is very prepared for each lesson she teaches. She introduces the lessons to students in a variety of ways. Her lessons are engaging and fun. She follows up the lessons with engaging activities and she conferences with her students on a regular basis as well as with their parents. She has a close relationship with her student's parents and with her class. (758)

15. ----- encourages her students to think critically at all times. Students must justify, defend and explain process for answers. (759)

16. Commitment to learning more so that she could meet the needs of students. She has a desire to be better. (796)

17. She goes the extra miles for her students. She supports them with differentiation, small group instruction, technology, and always has a positive classroom environment. She is also a mentor for teachers that are struggling. (867)

18. ----- has a strong understanding of her content knowledge, but also she understands best practices and instructional practices. (882)

19. Approximately one third of this teacher's fifth grade class were special education students. She diligently made data driven decisions to differentiate her instruction. Every individual student was rigorously challenged to achieve at a high level. Individual goals were set. Strategies for helping students achieve those goals were often identified in consultation with the student. (891)

20. This teacher taught fourth grade and he used best practices and strategies that engaged the students. He was also very intentional about involving the students in their learning i.e. He established "I Can" learning goals and he met with his students to discuss the progress. (1032)

21. Plans instruction effectively for content mastery, pacing and transitions; consistently demonstrates accurate, deep and current knowledge of content; understands how to differentiate instruction to meet the needs of all students; consistently maximizes instructional time; and uses assessments effectively to inform instruction while also varying assessments to determine individual student needs and progress. (1037)

22. This teacher annually demands the best from her students and herself. She refuses to accept excuses or to allow students to fail. She builds relationships with parents and students, going beyond what is expected during the school day to make home visits and tutor students. She is respected by parents and students for her firm but consistent demeanor. (850)

Output Characteristics Only [28% of responses, references to within year student growth underlined]

1. Her test scores are always high and on the statewide longitudinal data system here bubbles are in the green and upper right quadrant. [715]

2. Her students always make growth. [771]

3. She had a high percentage of her students pass the EOCT. (794)

4. She has one of the highest student growth percentile measures. She is focused in her methods and committed to the success of her students. (795)

5. The students in her class consistently pass the CRCT and her SGP is above 70%. (809)

6. Her growth model data indicates that she is one of the teachers in my building who has contributed to the highest student growth, especially in mathematics. (812)

7. This is a high poverty school, with a high student rate of transition…math scores increased in 2012-2013 and science and social studies scores increased by double-digits. (718)

8. All of -----'s students passed the math, ELA, reading, social studies CRCT with a 90 to 100 percent met or exceeded rate. (845)

9. I personally think that the 8th grade math content is the hardest content area in middle school. -----, has made steady gains over the past 7 years with academic achievement for all of her students with an average of 93% meeting or exceeding the standard on the first CRCT test and 96% overall meeting after the retest. Her class that teaches for high school credit typically has a 100% passage rate on the end of course test. (853)

10. Her results on the CRCT (65% met or exceeded) versus the scores of the other teachers (51% met, 47% met). (877)

11. Because of the successes that the teacher has had with students of varied ability levels and identifications. (872)

12. I chose this teacher based on comparing student achievement results across a period of three years. (813)

Input and Output [18% of responses]

1. ---- accepted the responsibility of improving the math scores of our 5th graders 4 years ago. Each year she works diligently with professional development activities and speaking with other professionals. We had only a 75% success rate four years ago in fifth grade math as stated in the CRCT results. This year, 2014, she posted a 97% passage rate in math in fifth grade. She is the best teacher in the building

2. ---- was able to take ALL students and teach bell to bell. Whether I gave her a honors class or an inclusion class, she welcomed all students and took a personal interest in these students. She was very organized which allowed her to make the most of her instruction. Her test scores were more than passing. She often had high exceeding percentages

(honors or not) and her growth model was in the 60's if I remember correctly. She was a great model teacher and often willing to help in any way the school needed her. (710)

3. This teacher consistently taught CCGPS. She provided her students with rigorous and engaging activities, set high expectations for herself as well as her students; attended numerous PL opportunities; received her Gifted Endorsement. Additionally, 100% of her students to passed both the Reading and Math on the 2014 CRCT. (766)

4. Teacher/student rapport are vital to education. ---- believes in effective communication skills. She understands the 7th grade math curriculum and communicates high expectation for student learning. Critical thinking, collaborative groups, guided instruction, and research-based hands-on activities/strategies are utilized during classroom instruction. (827)

5. ---- had 89% growth for her students for the 2012-13 school year. She is extremely thorough and ensures that each student's needs are met on a daily basis. She uses data to inform her interventions and continually takes the students deeper into the content. (834)

6. Based on student test data, the teacher has been very effective in meeting the need of her students over the past three years. This teachers works extremely hard in keeping parents informed and involved about their children's progress during the school year. (839)

7. She has taken ALL of the PL we have provided and improved her instruction over the last few years. She teaches students of great ability and SWD with the same skill and passion. Her scores demonstrate her success but it is obvious during the school year as well. (871)

8. ---- works diligently to take advantage of all opportunities afforded to him however in the classroom he is creative, supportive, and reflective. He is the type of teacher that takes an active role in students even beyond the classroom doors and his students know that he is 100% focused on their success. His scores 2013 growth model scores show that his students had high growth and high achievement. (879)

Illustrative Rationales Provided By Principals For Teachers Rated Least Successful

Input Characteristics Only [67% of responses]

1. Identifying misconceptions / Lack of reteaching and evaluating to address strengths and weaknesses / Lack of differentiated instruction / Identifying areas needed for professional development (610)

2. The teacher works with our SWD (Students with Disabilities) population. Although she is always willingly to try different strategies, I am not certain if she checks for frequent enough to make an impact on the students learning. (614)

3. Lack of use of data; instructional strategies consist of mostly lecture; limited opportunities for students to manipulate the content; not open to change in the best interest of the students (615)

4. Difficult to coach as a teacher. He was clearly hired before I arrived to coach men's soccer. unless we see significant improve with planning, instructional delivery and student achievement, he is aware that his coaching duties will expire after the 2014-2015 school year. He is on an informal professional development plan. (617)

5. ---- does not plan effective activities for students. She is very "text book" oriented which causes students to disengage from learning. (618)

6. She is unorganized and not timely with lesson planning, required reporting and promptness to work. (622)

7. The teacher is early in their career and still mastering the content and their delivery style (626)

8. She does not understand the concept of data analysis nor differentiating instruction. She has taught for 20+ years teaching content and is still teaching the content. If students do not get it, she moves onward. (633)

9. ---- is stubbornly old school. A noisy classroom makes her nervous. (635)

10. I believe that ---- is not as successful increasing student academic achievement due to her not being a very good classroom manager and due to her not investing much time in preparing for her students. (636)

11. Less than engaging personality / 2. "Old school" in delivery of instruction / 3. Energy level does not match middle schoolers needs / 4. Not as willing to try new, proven research-based instructional practices (645)

12. Teacher has great classroom management and the class looks as if the students are learning, but they are not performing on any assessment that is not created by the teacher (661)

13. SPED teacher and her students show the least amount of achievement. (665)

14. In addition to having a young child and no support system in the area where she lived, she encountered some medical issues which negatively impacted her effectiveness. (669)

15. The slightest classroom disturbance causes her to sit down and give up. (716)

16. She doesn't like students. (720)

17. She is a bit lazy and scattered. She lacks focus and passion. She always has an excuse. (682)

18. ---- has 20 years of experience, but does not behave like a confident veteran teacher. She needs constant support and does not appear to be able to consistently pull her weight. (690)

19. ---- is excellent at building relationships with his students. However, he could be much more effective at planning for instruction. His lack of adequate preparation prevents him from reaching his maximum potential. (800)

20. ---- had serious attendance issues. She was absent from work most of the school term. When she was present she was often unprepared and did not effective instruct her students. (806)

21. The teacher rarely participated in PD opportunities and did not respond to individualized feedback or modeling. (823)

22. She refuses to change her instructional practices. She feels that her experience affords her the right not to do what it takes in order to meet the demands of educating students in the 21st century. (764)

23. ---- was cited several times for not instructing students. He spent most of his time seated at his desk while students copied questions from the board. (994)

24. He doesn't tailor instruction to meet the individual needs of students. He just teaches the content and keeps going. (737)

Output Characteristics Only [17% of responses, references to within year student growth underlined]

1. Her <u>growth model data</u> indicated one of the least amount of contribution to student growth. (765)

2. Low level of student achievement each year as measured on standardized test. (791)

3. The students with disabilities did not perform at a growth rate commensurate with their peers. (748)

4. This teacher showed no progress in her CRCT results from last year to this year. (794)

5. This teacher had 27% of his students to pass the CRCT. (986)

6. <u>SGP measures</u> (830)

Both Input and Output [16% of responses]

1. She does not use DATA to drive her instruction. She insists upon using the Review for the Test method. While the students do pass the test, I believe it is not due to her efforts completely. Most of the students she taught had CRCT scores that indicated that they would be successful. However, there were 8th grade CRCT scores for students that indicated that they may struggle. There was no RTI or differentiation. And, those students who were targets for possible failure--- were indeed not successful. The level of growth was not sufficient. Also the rigor in the classes was not present in both reading and writing. (612)

2. ---- is the newest of teachers. He is hard working and loved by his students and co-workers. He has been extremely ill. (674)

3. Student academic achievement data demonstrated that the teacher's influence on their learning was very low. His classroom management was deplorable and not conducive to student learning, and he was reluctant to participate in professional development activities that generally helped teachers improve their job performance in ways that positively affected student academic achievement. (679)

4. ---- is learning how to meet the needs of SWD students in an inclusion setting. CRCT Scores increased from 0% meeting and/ or exceeding to approximately 50%. ---- received support and feedback from the principal, academic coach and SWD instructional specialist regarding co- teaching models and evidence- based learning strategies. (705)

5. This teacher's instructional style is very passive in nature. It lacks variety with respect to differentiation and students' aptitudes. As a result, many students who meet or exceed on high stakes assessment would show only "slow growth" relative to a database of similar students. (877)