Indiana's Teacher Evaluation System: A Four-Year Analysis

Indiana Teacher Appraisal and Support System (INTASS)

Submitted to the Indiana State Board of Education

Hardy Murphy, Ph.D.

Sandi Cole, Ed.D.

February 1, 2017

INDIANA INSTITUTE ON DISABILITY AND COMMUNITY
CENTER ON EDUCATION AND LIFELONG LEARNING

Executive Summary

The research findings presented in this report are part of an ongoing study of the implementation and impact of Educator Evaluation reform in the state of Indiana beginning with the passage of Senate Bill 1 in 2012. This legislation required research based rubrics for rating the effectiveness of teachers and the use of student learning as one aspect of the evaluation process. This report is one of a series of reports of research supported by grant funding from the Indiana State Board of Education and the Joyce Foundation.

Research questions concerning implementation practices, plan quality, student, educator, classroom, school, and districts demographics, ratings of instructional effectiveness and annual state assessment outcomes served as the framework for the analyses conducted. A data share agreement with the Indiana Department of Education and Indiana University that included educator ratings (summative ratings determined from evaluator observations of teaching and individual growth measure (IGM) ratings determined from student state assessment outcomes), teacher evaluation plan characteristics and an array of student, educator, and district demographics provided the data used in the analyses. This data was used to run multiple statistical analyses to determine possible relationships associated with the data obtained through this agreement.

Key findings:

1) There is a discrepancy between Individual Growth Model ratings of teacher effectiveness and the summative ratings given teachers by their evaluators.

2) There is not an identified difference in teacher ratings and student outcomes associated with the growth weight used in evaluation plans.

3) Student poverty level as designated by Free and Reduced Lunch status is the single most powerful predictor of teacher evaluation ratings and student learning outcomes.

4) There is some evidence of a relationship with the development and implementation of high quality plans with teacher effectiveness ratings and student outcomes.

5) There appears to be a relationship between teacher mobility, teacher experience, district percentage of students on free and reduced lunch, and teacher evaluation ratings.

6) There are evident distinctions in student learning outcomes and teacher ratings based upon classroom characteristics.

7) There is a relationship between prior year student assessment outcomes and teacher evaluation ratings.

8) There are relationships between teacher demographics and evaluation ratings and principal ratings and teacher evaluation ratings.

9) Changes in the state accountability system either inadvertently or by design impacted the consistency and quality of educator plan development and implementation.

10) The current teacher evaluation model does not effectively account for student demographics.

Recommendations from these research findings address 1) the need for continued implementation with clear guidelines and requirements for the implementation of educator evaluator training; 2) the development of a classroom based teacher evaluation growth model, or the use of statistical controls, to account for student demographics with more validity, and 3) further research into the relationship between evaluator ratings and student learning outcomes. The implications of these findings for the discussion regarding current and future legislation involving teacher evaluation cycles and other requirements, a review of the state A to F accountability system and its relationship with teacher evaluation, and making adjustments to the process for awarding state performance grants to teachers are also identified. Areas for future research to further investigate questions critical to the development of teachers for improved instructional quality and improvement of student learning outcomes are also

identified in the alignment of the findings with current topics in the discussion of school improvement in Indiana.

Introduction

In July of 2015, the Indiana State Board of Education (SBOE) and the Joyce Foundation engaged the Indiana Teacher Appraisal System of Supports (INTASS) in a multi-faceted research analyses of the changes to educator evaluation required by the passage of Senate Bill 1 in 2012. This research included 1) a review of plan quality and compliance with the law's requirements and 2) a comprehensive research of factors related to educator evaluator ratings and student outcomes.

The purpose of this report is to share the results to date of this ongoing and comprehensive research of factors related to educator evaluator ratings and student outcomes in Indiana. The report is organized into seven sections: Introduction, Research Questions, Indiana History of Accountability, Methods, Results, Conclusions and Recommendations.

Changes to teacher evaluation resulting in a process formally structured around best practices in goal setting, instructional observation and feedback and the incorporation of student learning outcomes emerged in the school reform literature in the last decade. Research documentation of school and student characteristics and their relationship to student outcomes generated a data based framework for the development of federal and state accountability systems.

With the advent of Race to the Top and its requirements and allowances for state waivers in qualifying for competitive grant funds came the use of educator/teacher evaluation and student outcomes in the accountability framework. The response to changes in the process of educator evaluation incorporating student learning outcomes led to significant challenges in an effort to become a viable component of the school improvement process. How to make the process equitable across subject areas and grade levels stressed the concept of fairness and

differential treatment for those in subject areas and grade levels that were part of the state accountability assessment system and those who were not.

The reliability and validity of student outcome measures and the metrics used for calculating growth and proficiency, evaluator training and interrater reliability, and the appropriateness of a multitude of evaluation rubrics threatened the credibility of ratings that were given through these new processes. Linking the ratings to compensation heightened adversarial rhetoric in a contentious debate and precipitated an ethics driven commentary that questioned both the motive and impact of these new evaluation processes upon the teaching and learning experiences across the K-12 spectrum.

Educator apprehension and debate in the policy and research community about the impact of student income on the perception and effectiveness of teachers, independent of student outcomes further complicates the understanding of these new evaluation approaches and their possible benefit. Ongoing dialogue among those involved in the actual process of evaluation- superintendents, principals and teachers- that indicates similarly skeptical views on some of the same critical issues begs the question of whether the intent of legislation and policy will be realized in a process of implementation with fidelity unless major concerns are addressed. Similarly, a redefinition of the federal role in accountability represented by every Student Succeeds Act, including its requirements and expectations for teacher evaluation, only underscores the impact of changing policy and guidance at the state level upon interpretation and implementation at the local level.

Finally, this research is being conducted over a time of considerable turbulence in the Indiana accountability system with changes in standards, assessments, and leadership happening in an often erratic fashion. In spite of these issues, Indiana's students have performed notably better on the bi-annually administered National Assessments of Educational Progress (NAEP). In comparison with performance of other states, this improvement since the passage of the teacher evaluation legislation in 2012 has earned recognition in the state and national press in each of the NAEP administrations over the past four years. The results of this research are intended to shed further insight about the important relationships of educator and student

demographics, district and school characteristics, and teacher evaluation practices with educator ratings and student learning. Finally, this report offers a set of recommendations for moving forward.

Indiana History of Evaluations, State Assessment and Accountability and impact Upon Research

The effect of a constantly shifting accountability environment and recurring problems in the state's management of its accountability system presented an ongoing challenge in reaching comparability among data files across multiple years consisting of information from educators, students, districts, schools and classrooms in the state. Data characteristics and teacher evaluation plan environment differed in each research year.

There were test changes in the Indiana State Testing for Educational Progress (ISTEP) three of the four years accompanied by recurring problems in ISTEP scoring. There were pervasive issues with timeliness of results and providing test results to schools for use in instructional planning. Further, at the local level, implementation of teacher evaluation plans was inconsistent across the four years of the research. This inconsistency in plan development and implementation is due in part to conflicting interpretation of policy and guidance is documented in previous reports and policy briefs (INTASS Plan Review report to SBOE, 2016).

Additionally, the hold harmless provision in accountability changed the climate of evaluation expectations and consequences of student performance. The introduction of new methodologies, e.g., technology and open-ended response questions further complicated the administration of the state's accountability test.

Changes in state standards shifted the instructional emphasis in classrooms across the state, and changes in incentives for educator performance, e.g., competitive grants based upon district school improvement in student test scores, contributed to a general sense of unfairness in the evaluation plan experience. These changes in educator evaluation, and state assessment for accountability over the years of the research is illustrated in Appendix A, Table 1.

Additional problems were presented as a result of ambiguity in legislative language and policy practices. This ambiguity resulted in erratic submission of data from schools and districts across the state and "looseness" in the guidance provided districts in their management and implementation of teacher evaluation created additional challenges in the questions that could be researched in methodologically sound ways and the identification of analyses that met acceptable research standards. Changing standards, different tests for accountability, less than stellar vendor performance management of state assessments, differences in interpretation of which students should be included in accountability and teacher evaluation, and how students were assigned to teachers for accountability presented comparability problems in samples that confounded many analyses.

Similarly, inconsistencies in data entry and coding across years created pervasive complications for file comparability at the state level. The practice of only requiring summative ratings to be reported from local districts meant that student learning outcomes could only be analyzed for those teachers receiving IGM ratings and not those for whom student learning was determined through other methods. Further, the relative differences in weights for different student learning measures in the evaluation plan design was beyond the methodology of this study, because those learning outcomes are not included in the data provided from districts. Similarly, the fact that some districts did not implement the learning outcomes requirement and others did not do so with individual teachers, but rather interpreted the requirement as being met by using A to F grades at the school level for all teachers meant that not all districts could be treated in the same way in the analyses.

The aforementioned presented a challenge in the development of a research methodology that might yield results with valuable insights. Although it is important to mention this important limitation at the outset, the findings of this report do provide insight into the impact of teacher evaluation in Indiana that enable useful recommendations for improving the teacher evaluation process in Indiana.

Research Questions

To address many of the significant questions concerning teacher evaluation and to some extent, school accountability, a set of research questions was developed that included plan quality and characteristics, educator and student demographics, and district, school, and classroom profiles. The following research questions guided this study:

1. Is there a relationship between evaluation plan quality and teacher evaluation ratings-summative and/or IGM?
2. Is there a relationship between evaluation plan quality and student learning/assessment outcomes?
3. Is there a relationship between teacher evaluation ratings and student outcomes?
4. Is there a relationship between teacher IGM ratings and Summative ratings?
5. Is there a relationship between teacher/evaluator/ student demographics and teacher evaluation ratings?
6. Is there a relationship between teacher/evaluator/student demographics and student outcomes?
7. Is there a relationship between classroom, school and district demographics and teacher ratings?
8. Is there a relationship between classroom, school or district demographics and student outcomes?

The results of the analyses conducted for this report attempt to answer these questions for teacher evaluation in the state of Indiana and are intended to help in the repackaging of educator evaluation as a constructive tool for successful teaching and learning. In order for this to happen, critical decisions regarding legislation, policy and guidance will be required. It is hoped that the findings in this report will be helpful.

Method

As mentioned in the introduction to this section, changes in the accountability system that impacted the comparability of student test results and teacher evaluation outcomes across years made the results of longitudinal analyses across years questionable. A within years method was opted for to research questions concerning relationships among the multitude of variables included in the research questions. This allowed for the variable relationships to be analyzed for each year of the data set and then compared with every other year. Although not a longitudinal analysis per se, comparing the results of each year allowed a determination of whether or not the results of any particular year were consistent across years. Further, looking at similarities and differences between those districts and schools that began implementing plans in response to the legislation from 2012 allowed some inquiry of the effect of time upon variable relationships across the years of the data sets.

Additionally, decisions had to made regarding data to be included or excluded in order to account for variations in data quality. In addition to changes in the testing and accountability environment and a lack of replicability across years that created incomparability across the files in the data sets, variations in interpretations of guidance and policy required adjustments in which data was to be used in an analysis or sets of analyses.

Adjustments were also made in analysis models to align student test results and the accountability system with teacher evaluation requirements, and to address demographic differences across districts schools and classrooms. For instance, the accountability system requires an attendance of at least 162 days in order for a student test result to be included in the accountability results for districts and schools. How to account for anomalies due to fluctuations at the extremes of the student performance distribution so that analyses would actually reflect the true relationship of instruction and student performance was factored in some analyses. In the end, a variety of demographic variables, plan attributes, student performance results, and district characteristics were analyzed in an array of variable and

model configurations to ensure that all systematic influences upon evaluation ratings and student outcomes were accounted for.

The size of the samples and the large number of analyses required the use of levels of significance to account for the possibility that significance could actually occur because of random and chance rather than any valid relationship between the variables being researched. To account for this and reduce the chance of attributing significance where it does not exist, very conservative values of .01 and .001 were used to attribute significance. It is important to establish that the research presented in this study is descriptive and establishes associations between the variables in question and not causality between and among the questioned relationships.

Note: In this report Free and Reduced lunch was used to determine Social Economic Status (SES).

Data Collection

The general data set included files obtained from the Indiana Department of Education through a Data Share Agreement covering the school-years 2012-2016. The data in the files consisted of student, teacher, and principal demographic information, results of ISTEP assessments, and educator evaluation ratings. For teachers in the accountability grades and subject areas, both IGM and summative ratings were available. For other teachers, only summative ratings were in the files.

No information in the files allowed for the identification of any individual whose data attributes were included in the analyses conducted to answer research questions. Other relevant data included district, school, and classroom characteristics, and local district teacher evaluation plans and plan components or attributes. The plan attributes were researched based upon a rating of plan quality by trained raters using a scale derived from the INTASS plan development rubric (INTASS Plan Review, 2016). Plans were ranked with other plans in the state based upon their overall score from the plan attribute ratings. Points associated with these rankings were used to conduct analyses of plan quality and teacher and student outcomes.

Data Composition

For each year, data from the IDOE on students, teachers, and districts were combined into a master file. This file includes data from the ENROLL dataset, which includes student demographics; the AD dataset, which includes information on student homelessness and high ability; the ES dataset, which includes information on student expulsions and suspensions; the ATTENDANCE dataset, which includes information on student attendance; the ISTEP+ dataset, which includes information on the students ISTEP performance; the CP dataset, which includes information on certified teachers positions; the ER dataset, which includes summative scores for teachers; the TEACHERS dataset, which includes IGM ratings for teachers; the GRADUATION RATES dataset, which includes graduation rates at the school level; the EXPENDITURES dataset, which includes yearly expenditures at the school level; and the GROWTH WEIGHTS dataset, which includes corporation-level growth weights applied to IGM ratings for the summative score for districts who were using growth weights in that year. The school and district-level datasets were linked to students and teachers by the IDOE CORPORATION ID, and the student and teacher datasets were linked using the EE dataset, which includes all linkages between students and teachers for IGM rating purposes, along with information about whether the teacher is accountable for the student's ISTEP+ ELA performance, Math performance, or both.

Data Cleaning

Before analyzing the data, the growth weights applied by districts to IGM ratings were mean-centered by year. The adjusted median growth percentiles used to determine IGM ratings were also calculated by (1) finding the median growth percentile of students assigned to each teacher, (2) finding the standard error of the growth percentiles of students assigned to each teacher, and (3) adding the median found in (1) to the 1.25 times the standard error found in (2). This is in alignment with the method of determining adjusted growth percentiles outlined by the IDOE for determining teacher IGM ratings. This number essentially represents the teacher's "best case scenario" for their student's growth, and is sensitive to both the growth of the students as well as the variation in growth in the classroom.

For data to be considered valid for analysis, the datasets could have no conflicting reports on the school and district IDs associated with a student and/or teacher. For analyses with ISTEP scale scores as the outcome variable, datasets were required to have no conflicting reports on district, school, or grade level.

Teachers who had job/role changes or had conflicting records between Period 1 and 2 of a given year were also eliminated from the analyses to eliminate uncertainty about the validity of our data. A teacher was considered to have changed jobs if the certified employee record indicated a change in their district or number of days contracted. A teacher was considered to have conflicting records if the prior employment code, degree code, first-year teacher indicator, salary, or percent of salary paid by Title I changed. While it is possible that a teacher could have earned a raise, completed their first year of teaching, or earned a degree during the year, these teachers were eliminated to remove uncertainty from the estimation of the effects of these variables on student and teacher outcomes.

Models

Analyses modeled the effects of a number of different combinations of variables on a number of different outcomes. Specifically, analyses investigated the roll of student demographics, teacher demographics, and district demographics on two measures of ISTEP performance (the student's scale score and the student's growth percentile) and on three measures of teacher performance (the IGM rating, the summative score, and the adjusted median growth percentiles used to determine the IGM rating). (See Appendix A, Table 2.)

Student demographics

Student demographic predictor sets include combinations of socio-economic status, special education status, ethnicity, English Language Learner status, attendance, grade (Middle vs. Elementary) and content area, where appropriate. Additional variables used in select models include an expulsion/suspension indicator, and/or the ISTEP growth percentile from the prior year and the ISTEP scores from the prior two years. Results using homelessness status and high ability status are also available but unreported.

Most analyses using student demographics as predictors also include Primary Exceptionality as a fixed effect, except when its inclusion causes the models to fail for mathematical reasons beyond the scope of this report. Primary Exceptionality is a categorical variable with over twenty levels, which adds excessive complexity to some of the logistic and multinomial models. Analyses using student demographics to predict ISTEP Scores also include student grade level as a fixed effect. However, because grade level was not found to be a significant factor in ISTEP growth, it is not included as a fixed effect for models using ISTEP Growth or Teacher Ratings as outcomes.

Educator demographics

Teacher demographic predictor sets include an indicator for whether they are a first-year teacher, experience (0-5 yrs., 6-10 yrs., 11-20 yrs., and 21+ yrs.), salary, percent of salary paid by Title I funds, teacher ethnicity, teacher gender, the number of days the teacher is contracted to work, the average grade served by the teacher, and whether the teacher is a middle school (7/8 grade) or elementary school (4/5 grade) teacher. These variables are included in analyses of teacher demographics relationships with student outcomes as well as their relationships with IGM ratings.

When analyzing teacher demographics and Summative Ratings, the teachers are further split into an analysis of just those with or those without an IGM rating, to investigate differences in how teachers summative ratings differ for teachers that are versus are not subject to IGM rating. These groups are also combined for a view of the overall relationship of teacher demographics with Summative Ratings.

These analyses are also restricted to teachers who are not listed as a principal or superintendent in the Certified Position file. An additional analysis was done using only principals, to investigate relationships between principal demographics and the summative ratings received by teachers at their schools (whom they presumably evaluate as part of the summative rating process). Principal demographics include principal ethnicity, principal gender, principal experience (0-5 yrs., 6-10 yrs., 11-20 yrs., and 21+ yrs.)-and the grade for which the

teacher being evaluated is responsible- i.e., whether the teacher being evaluated is a middle school (7/8 grade) vs. elementary school (4/5 grade) teacher. Lastly, we include the principal's summative rating as highly effective, effective, improvement necessary, ineffective, or not applicable (not evaluated).

District demographics

The most frequently included variable used to assess the effect of district demographics is the mean-centered growth weight applied by the district in determining summative ratings. District weight on growth is included in all analyses of student demographics on student and teacher outcomes, teacher demographics on teacher outcomes, principal demographics on summative ratings, and district demographics on student and teacher outcomes. District weight on growth is also included in the special analyses of fourth and fifth grade intact classrooms. Other district demographics included only when specifically focusing on the effect of district demographics on student and teacher outcomes are: district average ISTEP score, district average ISTEP growth percent, district graduation rate, district expenditures on students, district percentage of students on free and reduced-price lunch (FRL), and whether the student or teacher is in elementary (4/5 grade) or middle (7/8 grade) school.

Model specifications (See Appendix A, Table 2)

A variety of models were employed in the analyses. Most models use one observation per student (for student demographic predictors) or teacher (for teacher demographic predictors). But, when analyzing the effect of student demographics on teacher outcomes, a second dataset was constructed that aggregates the student data at the teacher level. For example, one set of specifications includes the individual demographics of each student in predicting teacher ratings (while accounting for teacher effects nested within school effects), while another looks at the percentages of students assigned to a teacher that are on FRL, are in special education, are English language learners, are minority ethnicities, etc., in predicting teacher ratings. We call these models using percentages rather than individual observations "aggregate" models.

For models using student demographics as predictors, there are eight specifications employed.

- Full model includes all valid data, and is a generalized linear mixed model that includes school effects nested within district effects. Furthermore, students must have attended school a minimum of 162 days, except in the aggregate models.

- Elem vs. Middle filters the full model to include only students in 4th, 5th, 7th, and 8th grades. Since 6th grade is considered elementary in some districts and middle school in others, 6th graders are not included in this specification.

- Middle 68% of State filters the full model to include only students whose ISTEP scores (for ISTEP scores as outcome) or ISTEP growth percent (for all other outcomes) were in the middle 68% of the state ISTEP scores or growth percents, respectively.

- Middle 68% of State + Elem vs. Middle includes filters on the full model described in both of the aforementioned specifications.

- Middle 68% of District filters the full model to include only students whose ISTEP scores (for ISTEP scores as outcome) or ISTEP growth percent (for all other outcomes) were in the middle 68% of their district's ISTEP scores or growth percents, respectively. This means that students in particularly weak or strong districts that were not included in the middle 2/3 of the state but who represent the "average" student within their district are included in this model, resulting in a wider variety of ISTEP performance and a looser definition of the "average student" in Indiana.

- Middle 68% of District + Elem vs. Middle includes filters on the full model described in both of the aforementioned specifications.

- ELA Accountable filters the aggregate dataset previously described to include only the effects of student demographics on IGM ratings for teachers who are accountable for the ELA performance of those students, and include the mean ISTEP growth and scores of their students.

- MATH Accountable filters the aggregate dataset previously described to include only the effects of student demographics on IGM ratings for teachers who are accountable for the math performance of those students, and include the mean ISTEP growth and scores of their students.

For models using teacher demographics as predictors, there were ten specifications employed.

For teacher demographic effect on student outcomes:

- ELA model looks at the effect of teacher demographics only on student ELA performance.

- ELA + Elem vs. Middle filters the ELA model to include only students in 4th, 5th, 7th, and 8th grades.

- ELA + Prior Year(s) Performance adds variables to the ELA model for prior ISTEP performance in ELA. For the ISTEP score outcome, this includes the prior year as well as two years prior. For the ISTEP growth outcome, this includes the prior ISTEP growth in ELA.

- MATH model looks at the effect of teacher demographics only on student ELA performance.

- MATH + Elem vs. Middle filters the ELA model to include only students in 4th, 5th, 7th, and 8th grades.

- MATH + Prior Year(s) Performance adds variables to the ELA model for prior ISTEP performance in ELA. For the ISTEP score outcome, this includes the prior year as well as the two years. For the ISTEP growth outcome, this includes the prior ISTEP growth in ELA.

For teacher demographic effect on teacher ratings:

- Full model is the same as previously described.

- Elem vs. Middle is the same as previously described.
- ELA Accountable is the same as previously described.
- MATH Accountable is the same as previously described.

For principal demographic effect on teacher summative ratings:

- Full model is the same as previously described.
- Elem vs. Middle is the same as previously described.

For models using district demographics as predictors, there were four specifications employed.

- Full model is the same as previously described, except it includes teacher effects nested within school effects. School data may be correlated due to membership in a district; teachers data may be correlated due to membership in a school; and students data might be correlated due to teacher assignment.
- Elem vs. Middle is the same as previously described, except it includes teacher effects nested within school effects.
- Middle 68% of State is the same as previously described, except it includes teacher effects nested within school effects.
- Middle 68% of District is the same as previously described, except it includes teacher effects nested within school effects.

For models focused only on fourth/fifth grade intact classrooms, there were three specifications employed.

- 4^{th}/5th Grade Only + No Special Ed Teachers + Math and ELA Accountable includes only teachers and students in fourth grade classrooms where the teacher is accountable for both the student math and ELA

- The dataset is further filtered to include only students who are linked to only one teacher. Furthermore, the teacher must be linked to either 15-27 students (one set of results) or 20-32 students (alternate set of results). These teachers must not be identified as special education teachers in the CP file. Restrictions also require the students to have attended at least 162 days, and teacher effects are nested within school effects with no effect for district.

- + ELA Growth filters the previous model to include only student data attached to an ELA score. In other words, a teacher can be listed in the EE file as accountable for a student's ELA and MATH performance, but the student has only one or the other performance recorded in the ISTEP+ file. For the aggregate datasets, prior year(s) ELA ISTEP scores and growth are also included.

- + Math Growth filters the previous model to include only student data attached to a MATH score. For the aggregate datasets, prior year(s) Math ISTEP scores and growth are also included.

For models investigating demographics and ISTEP scores, the scores were considered to be continuous, and a linear model was assumed. For models investigating demographics and ISTEP growth, the scores were also considered to be continuous, and a linear model was assumed. A model with outcomes bounded between 0 and 100 was also considered, but the unbounded model resulted in a much lower Akaike Information Criterion (AIC), which suggests that lower values demonstrate better, more parsimonious models. For models investigating the effect of demographics on the Adjusted Growth Percents used to determine IGM ratings, the adjusted growth percents were considered to be continuous, and a linear model was assumed. For models investigating demographics and IGM ratings, a multinomial model was assumed, with an ordinal relationship between ratings of 1, 2, 3, and 4. Teachers with a rating of 0, indicating the IGM rating was not applicable, were not included in the analyses. For models investigating demographics and summative ratings, a logistics model was assumed, grouping the teachers rated 1-3 together and comparing them to teachers rated a 4. This was done because of the very low percentage of teachers rated a 1 (less than 1%) or a 2 (less than 3%).

Plan Quality/Attribute Investigation

Plan quality, defined as how many of the plan attributes specified in the INTASS plan evaluation rubric were present in a particular plan, and plan attributes, grouping of major rating scale components by sub-category, were analyzed as an additional predictor of student and teacher outcomes. The dataset includes indicators for the presence of 37 plan attributes in 223 district plans for how teachers will be evaluated.

For models using the plan attributes as predictors, four model specifications were employed.

- Full model is the same as originally described.
- Elem vs. Middle is the same as originally described.
- Middle 68% of State is the same as originally described.
- Middle 68% of District is the same as originally described.

The 37 attributes were analyzed at three levels.

- Totals use the total score out of 37 attributes as the only predictor.
- High Level uses the total score for each category 1.0 through 12.0 as predictors.
- Mid-Level uses the total score for each subcategory (e.g., 4.1, 4.2, 4.3, etc.) as predictors.

Each analysis was performed on three sets of districts.

- All Districts include all 223 districts whose plans were investigated
- Early Adopters includes only the districts who put plans in effect in 2012
- Later Adopters include only the districts who put plans in effect after 2012

For all levels, an additional variable was included in the analysis of the plan attributes effects on ISTEP Growth that indicated whether or not the district had been identified as one of the six "highly effective" districts. This variable was added to research in some degree the impact of implementation with fidelity as these districts distinguished themselves in the IDOE monitoring process and were recommended by IDOE staff for their implementation of high quality plans with fidelity and were recognized by the SBOE for their efforts.

Statistical Significance

For a single model, an alpha of 1% ($p < 0.01$) was used to identify findings that were only marginally significant, versus an alpha of 0.01% ($p < 0.001$) for those that were highly significant. These should be contrasted with conventional measures of 0.1 for marginally significant results, and 0.05 for highly significant. The threshold was lowered to account for the large sample size within the data set, which has a tendency to deflate the p-values that determine statistical significance. Moreover, multiple models were run for each outcome measured. Further definitions were adopted to define strong, medium, and weak support for the statistical findings that were based upon the consistency of statistical significance across all models.

Statistically Meaningful

The multiple models and analyses often provided a wide array of findings that differed in significance and in relationships to the variables being analyzed.

The following scale was used to label variable significance across the analyses conducted:

None: Significant in none of the specifications*years

Weak: Significant in 25% or less of the specifications*years

Medium: Significant in 75% or less of the specification*years

Strong: Significant in over 75% of the specifications*years.

Results

Relationship between student demographics, student outcomes and teacher evaluation ratings.

1. Free and Reduced lunch has a strong negative relationship with:
    a. Student ISTEP growth scores
    b. Growth percentile used for IGM
    c. IGM ratings for teachers
    d. Summative ratings for all teachers
2. Ethnicity (Black, Hispanic, Multiracial) has a strong negative relationship with:
    a. ISTEP scores
3. Ethnicity (Black and Hispanic) has a strong negative relationship with
    a. Growth percentile used for IGM
4. Ethnicity (Black) has a medium negative relationship with:
    a. ISTEP growth
    b. IGM ratings
5. Ethnicity (Hispanic) has a medium negative relationship with:
    a. IGM ratings
    b. Summative ratings
6. Ethnicity (Hispanic) has a weak negative relationship with:

a. Student ISTEP growth scores

7. Special Education has a medium negative relationship with:

    a. Summative ratings

8. Special Education has a weak negative relationship with:

    a. Student ISTEP growth scores

9. Special Education has a strong negative relationship to:

    a. ISTEP scores

10. Limited English proficient has a strong negative relationship with:

    a. ISTEP scores

11. Limited English proficient has a medium positive relationship with:

    a. ISTEP growth

    b. Growth percentiles used for IGM ratings

12. Fluent English proficient has a strong positive relationship with:

    a. ISTEP scores

    b. ISTEP growth

13. Fluent English proficient has a medium positive relationship with:

    a. IGM teacher ratings

    b. Growth percentile used for IGM ratings

14. Student ethnicity is correlated with proficiency but the results weaken when we look at growth. Computing growth isn't as sensitive to student ethnicity.

15. Student attendance has a medium positive relationship with IGM ratings.

16. District weight on growth has a medium positive relationship with summative ratings but a weak negative relationship with IGM and growth percentiles.

Table 3: Analysis of Relationships of Student Demographics with Student and Teacher Outcomes

	Student Outcomes		Teacher Outcomes									
	ISTEP Score [A]	ISTEP Growth [B]	GPs used for IGM Ratings* [B]	[†C]	1-4 IGM Ratings [B]	[†C]	1-3 vs. 4 Summative Ratings: All Teachers [B]	[†C]	Teachers w/IGM [B]	[†C]	Teachers w/o IGM [B]	[†C]
District Weight on Growth	Medium –	None	Medium –	None	Medium –	Weak –	Medium +	Mixed	Medium +	Mixed	Medium +	None
Student Demographics												
Free/Reduced SES (Full price baseline)	Strong –	Strong –	Strong –	Strong –	Strong –	Strong –	Strong –	Strong –	Strong –	Strong –	Strong –	Medium –
Ethnicity (White baseline)												
American Indian	Strong –	None	Medium –		None		Weak +		None		Medium +	
Asian or Pacific Islander	Strong +	Strong +	Strong +		Strong +		Medium +		Medium +		Medium +	
Black	Strong –	Medium –	Strong –		Medium –		Mixed		Mixed		Strong –	
Hispanic	Strong –	Weak –	Strong –		Medium –		Medium –		Medium –		Medium –	
Multiracial	Strong –	None	Mixed		Mixed		Mixed		Mixed		Medium –	
Native Hawaiian or other PI	Mixed	None	Weak +		Weak +		None		None		Mixed	
English Language Learner (Not ELL baseline)				Medium +		Medium +		Weak +		Weak +		None
Fluent English proficient	Strong +	Strong +	Medium +		Medium +		Mixed		Mixed		Mixed	
Limited English proficient	Strong –	Medium +	Medium +		Mixed		Mixed		Mixed		Mixed	
Native English speaking immigrant	Medium +	Medium +	Medium +		Medium +		Strong +		Strong +		Medium +	
Special Ed (General Ed baseline)	Strong –	Weak –	None	Strong –	None	Strong –	Medium –	Weak +	Medium –	Medium +	Medium –	Medium –
ELA content area (Math baseline)	Mixed	Weak –	Mixed		Mixed		Mixed		Mixed		Weak –	
Middle school student (Elementary baseline)	Strong +	None	Strong –	Strong –	Weak –	Medium –	Medium –	None	Medium –	None	None	None
Average days attended				Medium +		Medium +		None		None		None
Student's ISTEP Score 1 yr prior				Weak +		Strong +		Strong +		Strong +		

* Calculated as $m_{GP} + 1.25 * se$ where m_{GP} = median growth percent for the teacher and se = the standard error of the growth percents for all n students assigned to the teacher, as outlined in the IGM Calculation Template.xlsx located at http://www.doe.in.gov/evaluations/student-growth

† Student data aggregated at the teacher level

A Models include Full model; Elem vs. Middle School; Middle 68% of State; Middle 68% of State * Elem vs. Middle School; Middle 68% of District; Middle 68% of District * Elem vs. Middle School

B Models include Full model; Elem vs. Middle School; Middle 68% of State; Middle 68% of District

C Models include Full model; Elem vs. Middle School; ELA Accountable; Math Accountable

Relationship between teacher demographics, student outcomes and teacher evaluation ratings

1. Weak positive relationship of district weights and summative ratings and weak negative on IGM ratings.

2. Strong positive relationship between student prior ISTEP scores:

    a. Summative ratings

    b. IGM ratings.

3. Student attendance has a strong positive relationship with:

    a. Student ISTEP growth scores

4. Student attendance has a medium positive relationship with:

    a. IGM ratings

    b. Growth percentiles used for IGM ratings

5. Student attendance has a weak positive relationship with:

    a. Summative ratings

6. Salary has a medium positive relationship with:

    a. Student ISTEP growth scores

7. Salary has a strong positive relationship with:

    a. IGM ratings

8. Salary has a weak positive relationship with:

    a. Summative ratings

9. First year teaching has a strong positive relationship with:

    a. Summative ratings

10. First year teaching has medium positive relationship with:

    a. IGM ratings

11. Teachers 6-10 years and 11-20 years have a strong positive relationship with:

    a. Summative ratings

12. Teachers 21+ years have a medium negative relationship with:

a. Student ISTEP growth scores

b. Growth percentiles used for IGM ratings

c. IGM ratings

13. Teachers 21+ years have a medium positive relationship with:

a. Summative ratings

14. Male teachers have weak negative relationship with:

a. ISTEP growth

15. Male teachers have medium negative relationship with:

a. Growth percentiles with IGM

b. IGM ratings

16. Male teachers have a strong negative relationship with:

a. Summative ratings

17. Middle school teachers have a strong negative relationship with:

a. ISTEP scores

18. Middle school teachers have a medium negative relationship with:

a. Growth percentiles for IGM

19. Middle school teachers have a weak negative relationship with:

a. IGM ratings

Table 4: Analysis of Relationship of Teacher Demographics with Student and Teacher Outcomes

	Student Outcomes				Teacher Outcomes						
	ISTEP Score		ISTEP Growth		GPs used for IGM Ratings*		1-4 IGM Ratings		1-3 vs. 4 Summative Ratings		
	With Salary[D]	Without Salary[D]	With Salary[D]	Without Salary[D]	With Salary[C]	Without Salary[C]	With Salary[C]	Without Salary[C]	All Teachers[C]	Teachers w/IGM[C]	Teachers w/o IGM[E]
District Weight on Growth					None		Weak –		Weak +	Weak +	None
Student's ISTEP Score 1 yr prior	Strong +				Medium +		Strong +		Strong +	Strong +	
Student's ISTEP Growth 1 yr prior			Strong +								
Teacher Demographics											
Average days their students attended	Strong +		Strong +		Medium +		Medium +		Weak +	Medium +	None
Days contracted	Medium –		Medium –		None		Weak –		None	Weak +	None
% salary paid by Title I	Medium –		None		Weak –		Weak –		Weak +	Weak +	None
Salary Amt	Strong +		Medium +		Medium +		Strong +		Weak +	Weak +	None
First year teaching	Weak +	Weak +	Medium +	Medium +	Medium +	Medium +	Medium +	Medium +	Strong +	Strong +	Medium +
Experience (0-5 yrs baseline)											
6-10 years	Weak +	Medium +	Weak +	Medium +	Weak +	Medium +	Weak +	Medium +	Strong +	Strong +	None
11-20 years	Weak –	Medium +	Medium –	Weak –	Weak –	Weak +	Weak –	None	Strong +	Strong +	None
21+ years	Weak –	Medium +	Medium –	Mixed	Medium –	None	Medium –	None	Medium +	Medium +	None
Ethnicity (White baseline)											
American Indian	Mixed		Weak +		Weak +		Medium +				
Asian or Pacific Islander	Weak –		Weak +		Weak –		Weak +				
Black	Medium –		Weak –		Weak –		Weak –				
Hispanic	Medium –		Weak –		None		None				
Multiracial	Weak –		None		None		None				
Native Hawaiian or other PI	Mixed		Mixed		Medium –		Mixed				
Teacher gender (Female baseline)	Mixed		Weak –		Medium –		Medium –		Strong –	Strong –	None
Average grade served	Mixed								Medium –	Medium –	Weak –
Middle school teacher (Elementary baseline)	Strong –		None		Medium –		Weak –		None	None	None

* Calculated as $m_{GP} + 1.25 * se$ where m_{GP} = median growth percent for the teacher and se = the standard error of the growth percents for all n students assigned to the teacher, as outlined in the IGM Calculation Template.xlsx located at http://www.doe.in.gov/evaluations/student-growth

C Models include Full model; Elem vs. Middle School; ELA Accountable; Math Accountable

D Models include ELA model; ELA + Elem vs. Middle School; ELA + Prior Year Performance; MATH model; MATH + Elem vs. Middle School; MATH + Prior Year Performance

E Models include Full model; Elem vs. Middle School

Relationships between District demographics, student outcomes and teacher evaluation

1. District weight has no discernable relationship to any variable included in analysis.
2. There is a strong relationship to District average ISTEP scores and student ISTEP scores.
3. There is a strong relationship to District ISTEP scores and IGM ratings.
4. There is a strong relationship between District ISTEP growth and student ISTEP growth.
5. There is a strong relationship between District ISTEP growth and growth percentiles used for IGM.
6. There is a strong relationship between District ISTEP growth and IGM ratings, summative ratings for IGM teachers.
7. There is a medium relationship between District average ISTEP scores and ISTEP growth/growth percentiles used for IGM.

8. There is a medium relationship between District percent FRL and IGM ratings/summative ratings.

9. There is a relationship between student prior year ISTEP scores and growth percentiles used for IGM, IGM ratings, and summative ratings.

Table 5: Analysis of Relationships of District Demographics with Student and Teacher Outcomes

	Student Outcomes		Teacher Outcomes				
	ISTEP Score [B]	ISTEP Growth [B]	GPs used for IGM Ratings* [B-alt]	1-4 IGM Ratings [B-alt]	1-3 vs. 4 Summative Ratings		
					All Teachers [B-alt]	Teachers w/IGM [B-alt]	Teachers w/o IGM [B-alt]
District Demographics							
District weight on growth (2013-2016)	None	None	None	None	Mixed	Mixed	None
District average ISTEP Score	Strong +	Medium –	Medium –	Strong –	Mixed	Mixed	Weak +
District average ISTEP Growth	Weak –	Strong +	Strong +	Strong +	Strong +	Strong +	None
District graduation rate (2012-2015)	Weak –	None	None	None	Weak +	None	None
District expenditures per students (2012-2015)	None	None	Weak +	None	None	None	None
District %age of students on FRL	Weak +	Weak +	Weak –	Medium –	Medium –	Medium –	Medium –
Total enrollment			Medium +	Weak +	Medium –	Medium –	Weak –
Prior year ISTEP scores			Strong +	Strong +	Strong +	Strong +	
Middle school student/teacher (Elementary baseline)	None	None	Weak –	None	Medium –	Strong –	Medium +

* Calculated as $m_{GP} + 1.25 * se$ where m_{GP} = median growth percent for the teacher and se = the standard error of the growth percents for all n students assigned to the teacher, as outlined in the IGM Calculation Template.xlsx located at http://www.doe.in.gov/evaluations/student-growth

B Models include Full model; Elem vs. Middle School; Middle 68% of State; Middle 68% of District

B-alt Models include Full model; Elem vs. Middle School; Middle 68% of State; Middle 68% of District with alternate controls

Relationships between Principal characteristics, student outcomes and teacher evaluation ratings

1. There is a weak positive relationship with district weight and growth and teacher summative ratings.

2. There is a strong negative relationship between principals who are rated effective and teachers summative ratings (Principals rated effective rate their teachers lower than principals rated HE).

3. There is a medium negative relationship between principals rated improvement necessary and teachers summative ratings. (Principals rated Needs Improvement rate their teachers lower than principals rated HE)

4. There is a medium negative relationship between principals rated ineffective and teachers summative ratings. (Principals rated Ineffective rate their teachers lower than principals rated HE)

5. Principal gender, ethnicity or experience have no significant relationship with teacher ratings.

6. There is a weak relationship between direct expenditures per student and growth percentiles used for IGM.

Table 6: Analysis of Relationship of Principal Demographics with Teacher Evaluation Ratings

	Teacher Outcomes 1-3 vs. 4 Summative Ratings All Teachers[E]	
District Weight on Growth	**Weak**	**+**
Student's ISTEP score 1 yr prior	**Strong**	**+**
Principal Demographics		
Average days their students attended	**None**	
Principal Summative Rating (Highly Effective baseline)		
Effective	**Strong**	**–**
Improvement necessary	**Medium**	**–**
Ineffective	**Medium**	**–**
Not applicable (not evaluated)	**Medium**	**–**
Principal Experience (0-5 yrs baseline)		
6-10 years	**None**	
11-20 years	**None**	
21+ years	**None**	
Principal Ethnicity (White baseline)		
American Indian	**Strong**	**–**
Asian or Pacific Islander	**Medium**	**+**
Black	**Weak**	**–**
Hispanic	**None**	
Multiracial	**Weak**	**–**
Native Hawaiian or other PI	**None Sampled**	
Principal gender (Female baseline)	**None**	
Average grade served	**Medium**	**–**
Middle school teacher (Elementary baseline)	**None**	

E Models include Full model; Elem vs. Middle School

Relationships between classroom composition and student and teacher outcomes

Fourth and fifth grade classrooms with high free and reduced lunch versus low free and reduced lunch:

1. There is a strong relationship with the proportion of students on free and reduced lunch in a classroom and student and teacher outcomes.

2. Teachers in classrooms with higher percentages of student on free and reduced lunch, teachers are rated less well on IGM and summative ratings.

3. Teacher IGM and summative ratings are lower in classrooms with higher percentages of students on free and reduced lunch.

Fourth and fifth grade classrooms free and reduced lunch distribution decile comparisons

1. There are "break" points in classroom composition where identifiable differences in student performance and teacher ratings are associated with percentages of students on free and reduced lunch.

2. Using decile distinctions of classrooms by % of free and reduced lunch:

    a. Decile comparisons show distinct differences between classrooms based on percent of students on free and reduced lunch. In most models, there is a point increasing FRL no longer mattered. Typically, around the 7^{th} decile or 65%.

3. In the Figure A below you can see a decline in average IGM rating as the proportion of students on free and reduced lunch increases in classrooms. Statistical comparisons detect significant impacts immediately upon changing deciles in several models. This is not as pronounced in all models. Sometimes the second and third deciles are about the same as the first. (In this analyses of "like" classrooms, the narrower and conservative split of deciles was chosen in an effort towards caution.) That is to say that classrooms in the second decile (18% - 28.5% FRL) have statistically lower IGM scores than classrooms in the first decile (0% - 18%). Similarly, we find that classrooms with 65% or more of students on free and reduced lunch show no statistically significant differences between one another. This suggests there may be immediate detectable effects on IGM scores arising from the presence of low SES students, in a classroom, but the effect of additional low SES students tapers off as the proportion grows. (See Appendix A for

Figure B: 4th/5th classroom mean growth percentile by FRL decile and C: mean scale score by FRL decile.)

Figure A: 4th/5th Classroom FRL decile and IGM Mean

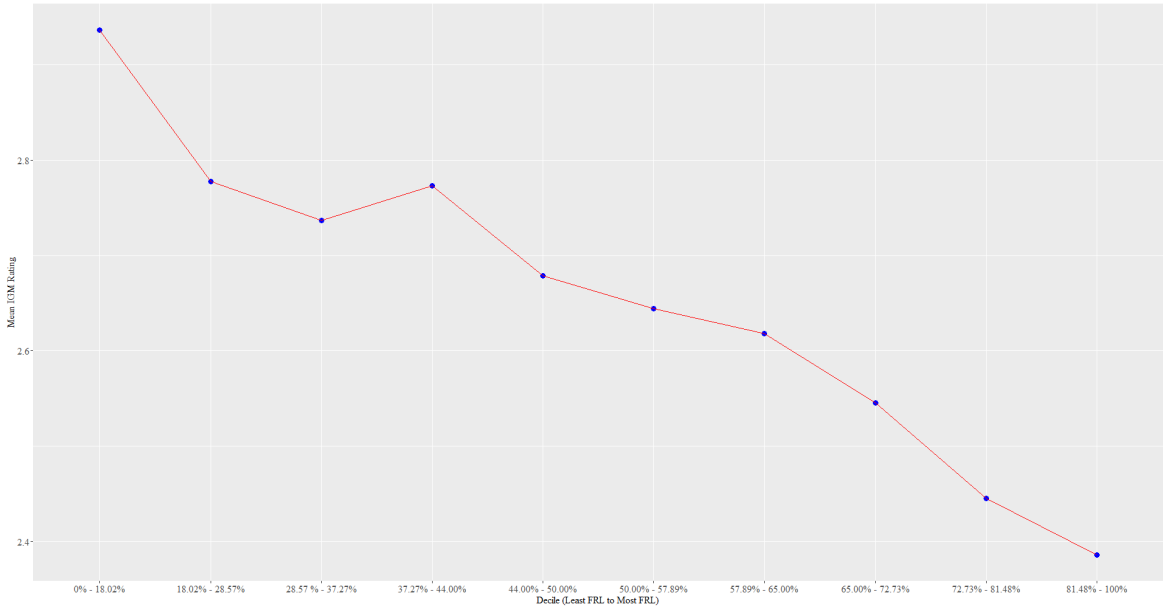

Table 7: Analysis of Relationship of 4th and 5th Grade Intact Classrooms with Student and Teacher Outcomes

	STUDENT OUTCOMES		TEACHER OUTCOMES					
	ISTEP Score F	ISTEP Growth F	GPs used for IGM Ratings* F	GPs used for IGM Ratings* †F	1-4 IGM Ratings F	1-4 IGM Ratings †F	1-3 vs. 4 Summative Ratings F	1-3 vs. 4 Summative Ratings †F
District Weight on Growth	Medium –	None	None	None	None	None	Medium –	Medium –
Classroom size	Medium +	None	Weak –	Weak –	Weak +	None	Medium +	Medium +
Student Demographics								
Free/Reduced SES (Full price baseline)	Strong –	Strong –	Medium –	Medium –	Medium –	Weak –	None	Medium –
Ethnicity (White baseline)				None		None		None
American Indian	Medium –	Weak –	None		None		None	
Asian or Pacific Islander	Strong +	Strong +	Medium +		Medium +		Medium +	
Black	Strong –	Medium +	Weak –		Weak –		Medium –	
Hispanic	Strong –	None	None		None		None	
Multiracial	Medium –	Mixed	Weak +		None		None	
Native Hawaiian or other PI	None	None	None		None		None	
English Language Learner (Not ELL baseline)				None		None		None
Fluent English proficient	Strong +	Medium +	None		None		None	
Limited English proficient	Strong –	Mixed	None		None		None	
Native English speaking immigrant	None	Weak +	None		None		Medium +	
Special Ed (General Ed baseline)	Strong –	Weak –	Weak –	None –	Medium –	None	None	Medium +
Expelled or Suspended (baseline=No)	Strong –	Strong –	Medium –	Medium –	Medium –	Medium –	Medium –	None
ELA content area (Math baseline)	Mixed	Strong –	None		None		None	
Average days attended				Weak +		Weak +		None
Student's ISTEP Score 1 yr prior				Medium +		Medium +		Strong +
% FRL in the Classroom, by Deciles								
1st decile (low % on FRL - Wealthy Classroom)	Baseline	Baseline	Baseline					
2nd decile (low % on FRL - Wealthy Classroom)	Strong –	Medium –	Medium –		Medium –		None	
3rd decile (low % on FRL - Wealthy Classroom)	Strong –	Medium –	Medium –		Medium –		None	
4th decile (average % on FRL)	Strong –	Medium –	Strong –		Strong –		Medium –	
5th decile (average % on FRL)	Strong –	Strong –	Strong –		Strong –		Medium –	
6th decile (average % on FRL)	Strong –	Strong –	Strong –		Strong –		Strong –	
7th decile (average % on FRL)	Strong –	Strong –	Strong –		Strong –		Medium –	
8th decile (high % on FRL - Poor Classroom)	Strong –	Strong –	Strong –		Strong –		Strong –	
9th decile (high % on FRL - Poor Classroom)	Strong –	Medium –	Strong –		Strong –		Strong –	
10th decile (high % on FRL - Poor Classroom)	Strong –	Strong –	Strong –		Strong –		Strong –	

* Calculated as $m_{GP} + 1.25 * se$ where m_{GP} = median growth percent for the teacher and se = the standard error of the growth percents for all n students assigned to the teacher, as outlined in the IGM Calculation Template.xlsx located at http://www.doe.in.gov/evaluations/student-growth

† Student data aggregated at the teacher level

F Models include 4th Grade Only with No Special Ed Teachers who are both Math and ELA Accountable; + ELA Growth; + Math Growth

Variance between IGM ratings and summative ratings:

Analyses were conducted to determine the amount of inconsistency between Summative and IGM ratings. The results of these analyses show that 1) there are significant discrepancies between IGM and summative ratings and 2) the most significant discrepancies are between IGM's of Needs to Improve and Summative ratings of Highly Effective and Effective.

Table 8: Variance Between IGM Ratings and Summative Ratings

	2013 Percent of Teachers	2014 Percent of Teachers	2015*Percent of teachers
Summative rating of effective with an IGM ratings of Ineffective	13%	11%	12%
Summative rating of HE with an IGM ratings of Ineffective	5%	4%	6%
Summative rating of effective with an IGM rating of Needs Improvement	29%	33%	33%
Summative rating of HE with an IGM rating of Needs Improvement	19%	16%	23%

*Hold harmless year

Plan quality, Plan components and Attributes

A series of analyses was conducted to determine if there are relationships between plan quality and plan attributes with teacher ratings and student outcomes. These analyses were further refined by dividing the plans into 1) those receiving the highest plan total ratings in the 38 attributes of plan quality, 2) those six districts recognized for plan quality and implementation with fidelity, 3) those districts implementing the new teacher evaluation systems in 2012-13 school year designated as "early adopters," and those implementing as "late adopters" after 2012-13. The analyses conducted were:

- Summative ratings and their relationship with IGM ratings.
- Plan rubric attribute ratings and their relationship with educator evaluation ratings.
- Plan rubric attribute ratings and their relationship with educator IGM ratings.
- Relationships between students assigned and not assigned to educators and student demographics.
- Years of implementation and plan characteristic consistency.

In general, the only findings with significance occurring with any consistency happened with the analyses of plan quality and outcomes for the six recognized districts.

The findings for these analyses are:

1) The six districts recognized for their plan quality also had better student growth

2) There is a Medium to strong relationship between plan quality and teacher summative ratings for districts.

3) There is A strong relationship between prior year student ISTEP scores and teacher ratings for districts.

Table 9: Plan Quality and Teacher and Student Outcomes

	Student Outcomes				Teacher Outcomes	
	Including %FRL		Without %FRL		Including %FRL	Without %FRL
	ISTEP Scores	ISTEP Growth	ISTEP Scores	ISTEP Growth	Summative Ratings	Summative Ratings
District Demographics						
Plan Total	None	None	Mixed	Medium –	Strong +	Strong +
Highly Effective (Not Highly Effective Baseline)	Weak +	Weak +	Medium +	Medium +	None	None
Prior Year ISTEP Score (Mean)					Strong +	Strong +
Teacher Demographics						
Elementary School Teacher (Middle Baseline)	Strong –	None –	Strong –	None	None	None

Teacher mobility and evaluation ratings

In order to see if there were identifiable patterns in teacher mobility across and within districts, evaluation ratings and student outcomes a series of analyses were conducted that looked at changes in district and school assignments across the years of the study and the ratings and student outcomes associated with individual teachers. The results show that:

1. There is no relationship between teacher mobility and whether or not teachers are in an accountability grade or subject area.

2. Teachers tend to move from districts and with high percentages of students on free and reduced lunch to districts with lower percentages of students on free and reduced lunch.

3. There is not an association with teacher mobility from districts with higher percentages of students on free and reduced lunch to districts with lower percentages of student on free and reduced lunch and changes in IGM ratings.

4. There is a relationship between movement from districts with high percentages of students on free and reduced lunch to districts with lower percentages of students on free and reduced lunch and summative teacher evaluation ratings.

5. More experienced teachers are less likely to move districts.

Table 10: Teacher Mobility

TEACHER MOBILITY				
	Moved Districts	**Decrease in FRL**	**Change in Rating IGM**	**Summative**
Accountable Experience	**None** / **Strong –**			
Moved Districts		**Strong +**	**None**	**Strong +**

Conclusions

The research results offer insight into several critical questions concerning the implementation of teacher evaluation in Indiana since the passage of the teacher evaluation legislation in 2012. In addition to understanding the impact of teacher evaluation and outcomes associated with its implementation, data illustrative of the management of its rollout also provide a picture of how to more effectively address implementation on a going forward basis. Similarly, instability in the accountability environment and the impact upon compliance, confidence, and trust in systems designed to gage and improve the teaching and learning experiences in the state offer valuable information for use in planning the administration and management of these systems in the future.

The debate about accountability at the national level was also experienced in Indiana through the A to F grading system for schools. The apprehensive reaction to accountability became

further aggravated with a loss of credibility due to pervasive and ongoing problems with the administration and management of the test administration process for the state assessment, late reporting of test scores, errors in scoring, inaccurate budgeting of funds targeted for at risk students, and ambiguous language in legislation, guidance and policy. This led to inconsistent implementation of teacher evaluation from the moment that the 2012 legislation became law. The enduring instability of this environment suggests that we have yet to see what a consistently implemented, high quality teacher evaluation process will yield in the form of instructional effectiveness and student outcomes.

Researching teacher evaluation also surfaced issues concerning data integrity and file composition that made comparative analyses across years difficult even within the same school corporation. Further, considerable latitude in the interpretation of policy and guidance severely limited quality comparisons across school corporations. The process of assigning students to teachers is an example worth mentioning. Districts are given the responsibility to assign students to teachers for accountability and for teacher evaluation. However, the latitude that districts have in this process allows students to be assigned to every teacher in the school in an effort to create equality in the evaluation experience and universal responsibility. Likewise, some districts use A to F grades as their quantitative measure and assign this as the only student learning outcome for a teacher. Additionally, the data files showed that many students were not assigned to any teacher and their learning outcomes were not assigned to any teacher. There appears to be no readily available explanation to evaluate these practices.

Percent FRL (%FRL) and plan attributes (PA) are two very important variables central to the questions of the research. The strongest relationships appear to happen when FRL% is not accounted for. However, even when FRL% is included in the analyses, relationships appear across the six districts with high plan quality and fidelity of implementation. When the first model was run (the model without %FRL), plan attributes were a significant predictor of outcomes. After adding %FRL to the model the significance of PA is sometimes weakened or disappears. However, for the six districts rated most highly in plan development and implementation there is still a conditional relationship between plan attributes and outcomes

for students and teachers. Seeing a positive relationship between the six districts identified through the IDOE monitoring process for plan quality and later recognized through a process that included assessment of implementation and technical assistance is encouraging because the student and teacher outcomes validate the efforts for developing and implementing high quality teacher evaluation experiences in school districts. Additionally, although consistent significance for subcategories of plan attributes is not apparent in the current analyses it may be because the collective rather than the individual influence of the attributes is what we see as evident in the analyses and results of plan quality and outcomes.

The research findings concerning a relationship between IGM and Summative ratings shows the impact of evaluator judgment upon the evaluation process and validates the tension between the rating of teachers receiving IGM ratings and those who do not. It also validates concerns about differences in the process between the two types of teachers. The fact that significantly more teachers are rated as ineffective or needs to improve with their IGM ratings than their summative ratings raises the question of the relationship between evaluator training, instructional effectiveness and the validity and reliability of evaluator ratings, student growth measures and the instructional rubrics being used.

The finding that the single most relevant predictor of teacher evaluation ratings and student outcomes is student free and reduced lunch status points toward the necessity for some accounting of this influence being included in the teacher evaluation rating process. Similarly, the fact that classroom makeup has a significant relationship with evaluation ratings is important and requires adjustment in the way that ratings are assigned. While quality of instruction should be the guiding factor in determining a teacher's evaluation, it would appear as though teacher evaluations are, at least in part, unfairly based upon the demographic composition of their classrooms.

While the impact of Student Economic Status (SES) on student outcomes may be difficult to eliminate, we can at least do a better job of accounting for it in evaluating the effectiveness of teachers. In this research undertaking a procedure for doing so was illustrated using the actual research data. A set of analyses was conducted by dividing all Indiana fourth and fifth grade

classrooms into deciles based upon their proportions of students on free and reduced lunch. Precise deciles vary from year to year and by the outcome in question, but the first decile consists of classrooms with roughly 0% to 17% of students on free and reduced lunch, and the last decile is roughly 83% to 100% of students on free and reduced lunch. Dividing classrooms in this fashion allows for an approximately equal number of classrooms within each decile. One such representation of these decile comparisons was shown in Figure A that illustrated the relationship between classroom FRL composition and IGM scores for 2016. This methodology approximates comparisons of like classrooms and an evaluation of teacher performance with their peers. Of course, this, like other methodologies involving static modeling, there are limitations that will have to be overcome.

However, we argue from a policy perspective that it is worth the time and effort to explore resolution of any limitations of method in order to better address the negative bias of student SES on teacher evaluation, in order to get an accurate read on instructional effectiveness. It can also help to address, in part, concerns of unfairness in the evaluation process because, in the current model, we are not comparing like classrooms in the rating of student learning outcomes.

The decile methodology described here suggests that one way to accomplish it would be to evaluate teachers in classrooms with similar proportions of low SES students. To this end, it does not matter whether the fifth and sixth deciles are statistically equivalent or different; it only matters that classrooms within the decile are similar with regard to their SES composition. Other breakpoints such as quintiles or quartiles could be used, but deciles do a satisfactory job of limiting the range of low SES proportions while maintaining reasonable numbers of classrooms for comparison. Likewise, the results analyzing student race and ethnicity, language status, and to a lesser extent disability status suggest similar within classroom interactions are happening that make the present growth model appear to be less valid with increasing diversity.

The finding that teacher ratings are also explained by prior student ISTEP performance is also worth commenting on. Teachers having students experiencing prior assessment success

realizing higher evaluations than teachers with classrooms having higher percentages of students on free and reduced lunch because these classrooms have fewer students experiencing prior success. This finding underscores the policy and philosophical dilemma of how to address the impact of tracking and clustering of students. From a local policy perspective in the teacher evaluation process, addressing tracking is worth deliberation in an effort to ensure a fair and valid rating of teacher effectiveness and reduce the impact of disadvantaging or advantaging some teachers over others in the evaluation process.

These findings suggest that something other than the current growth model based upon student to student cohorts is needed. As mentioned above, comparing like classrooms much in the way that student cohorts are compared in the current growth model is one option worth considering. Another option would be to use statistical controls to account for the effect of free and reduced lunch upon evaluation ratings. It is troubling to think that statistically accounting for student differences for some is akin to accepting less than equitable student outcomes based upon race and ethnicity. However, the fact that teachers leaving districts and schools with high percentages of students on free and reduced lunch receive higher summative and IGM ratings suggests that something other than teacher differences and instructional capability is at play.

It may be that teacher preparation programs have a part in the adequate preparation of teachers for classroom diversity. The fact that teachers with more education and qualifications seem to have better student outcomes, offers reason for supporting continued development of teachers in the recruitment and retention process. Policy and guidance at a different level of the teacher preparation process including post-secondary implications are important.

Evaluator and teacher characteristics significantly related to ratings and outcomes also shed light on the process and offer areas for additional inquiry. Determining why teachers with more than twenty years of experience appear to have negative student outcome indicators, even though teacher experience is significantly and positively related to evaluation ratings is important. Similarly, the interaction between evaluator and teacher experience and teacher ratings begs the question of rating validity. Why principals make up the difference in IGM

ratings with higher summative ratings is yet another critical question that underscores this point. Resolving these discrepancies will go a long-ways toward establishing credibility in the teacher evaluation process.

Similarly, although not a focus of this research, the fact that student characteristics explain more variance than either school or district differences has implications for the state accountability system and its letter grade format. The analyses would suggest that the A to F system is in reality grading/labeling students rather than the instructional processes in the respective schools and district corporations. The implications for school takeover and teacher incentive rewards are obvious.

Underscoring these issues are the findings regarding teacher mobility, experience, and ratings. The fact that inexperienced teachers seem to move to districts with lower percentages of students on free and reduced lunch is an important indicator of how, if not handled correctly, teacher evaluation can exacerbate the problem of teacher retention. The fact that teacher ratings improve without a corresponding increase in teacher IGM ratings suggests that there is a flaw in the system that needs to be addressed. However, the fact that more experienced teachers are rated more highly would appear to offer support for teacher mentoring with experience. Making these findings even more complicated are the results that show some relationship between principal and teacher evaluation ratings.

It appears that there is little evidence that plan weight is related to either teacher ratings or student outcomes. However, because only student outcome ratings for teachers receiving IGM ratings is available, the relationship of other student outcome measures in the evaluation process makes determination of learning weight an elusive undertaking. Compounding the problem is the lack of teacher awareness concerning plan components. Teacher awareness of the evaluation process is not something that is monitored when looking at district plans for compliance. Previous work in this area of Indiana's teacher evaluation system has shown that there is a significant lack of awareness on the part of teachers of evaluation plan processes, components, and criteria. Because of this, the impact of a particular weight is difficult to sift out from those factors having an influence upon individual motivation and collective effort. Yet, the

Indiana experience of continuous and significant improvement in NAEP results during the years since the implementation of evaluation reform in the state suggests that at the very least the presence of weight in evaluation and other reform components have not worked to the detriment of instructional improvement and student learning.

Notwithstanding the differences evident in implementation and practice, changes in educator evaluation and its role in accountability represent a theme that will continue to be a part of the school effectiveness discussion. In addition to resolving the questions concerning the validity and reliability of the ratings, the question of how to make the process supportive and helpful for educators seems to be a theme that resonates with all involved. Making this transition will require something more than a resolution of the metrics associated with the process and involve a repackaging of the message to one of support for teacher and student success. Additionally, the resolution of disparate political view-points in the educational reform environment will be necessary to change the tone of conflicting opinions regarding its purpose and the strategies offered as solutions.

Management competence will be a necessary component for the implementation of the large-scale aspects of accountability and teacher evaluation including holding the vendors responsible for test development and scoring accountability. In the end, all involved in this process will have to undertake the development and articulation of a new message of efficiency and support for educators backed up with policy and effective monitoring and assistance at the national, state and local levels in order to create consensus and support for this new paradigm in educator evaluation.

Recommendations

The following are recommendations based upon the findings from this report:

1. Ensure data integrity and file composition in order for research enabling informed decision making on policy and guidance to be valid and effective.

2. Ensure explicit guidance on assigning students to teachers for accountability and include this in the monitoring process.

3. Ensure vendor management and competence with impactful penalties enforced for lack of performance.

4. Review the current growth model to ensure that student characteristics are accounted for while maintaining high expectations for all students.

5. Consider exploring a teacher evaluation growth model that looks at like classroom comparisons as the basis for the student learning component.

6. Consider exploring a teacher evaluation classroom growth model that looks at growth based upon a classroom metric other than student growth percentiles.

7. Establish an effective monitoring system that includes an analysis of teacher evaluation ratings and student outcomes to inform state support for teacher development.

8. Provide ongoing support to ensure inter- rater consistency in the teacher evaluation process.

9. Elevate principal and superintendent evaluation quality assurance to the same level of importance and teacher evaluation.

10. Establish a set of criteria based upon research findings to evaluate plan development and plan implementation.

11. Establish incentives for districts to engage in a teacher evaluation process that focuses on teacher development through professional development.

12. Establish a process to ensure that teachers are involved in teacher evaluation as a collaborative process.

13. Consider establishing state provided incentives for teachers to accept teaching assignments in districts and classrooms with high percentages of students on free and reduced lunch.

14. Provide state support for professional development in culturally responsive instruction.

15. Provide funds for districts to engage in a state supported pilot project to revise and improve the teacher evaluation process.

Future Research

To date, the INTASS team has explored the relationship between district growth weight and student performance, district growth weight and IGM scores, discrepancies between IGM scores and summative ratings, the impact of student SES on student outcomes and IGM scores, comparisons between districts with high fidelity plan implementation to those without, and more. Extensive scope and depth are represented in the models constructed for this report yielding findings that can inform both research and policy on a going forward basis. The results to date are informative, but many questions remain including a reliable investigation into plan components and outcomes for students and teachers. Are some aspects of plan development and implementation being an important question that has implications for policy and practice A further exploration into the precise nature of the relationship between student FRL status and outcomes could also yield beneficial information as we move forward with educator evaluation. Such additional models might include exploring whether or not low SES status disproportionately affects certain types of teachers and students, or whether FRL exhibits patterns within schools receiving different levels funding.

Thus far, a great deal of emphasis has been put on determining which factors other than student SES uniquely explain variance in student outcomes, IGM scores, and other outcomes of interest. These relationships are often not as strong when we include FRL as an explanatory variable in our model. To date, how these variables are confounded with Free and Reduced Lunch is unclear. Perhaps, however, this approach isn't the best way to think of how to impact

student outcomes and RISE scores. Instead, perhaps we should look for variables that mitigate the effect of FRL, rather than uniquely explain variance. Such an approach would call for mediation models with random effects, and would allow us to answer questions centered on whether or not plan quality and fidelity of implementation actively reduce the negative effects of low SES on outcomes. It could also identify these variables as inputs to the teaching process that may have the desired impact of overcoming the negative impact of poverty upon learning outcomes. Research is still being done on the availability and applicability of such models to our data.

To accomplish the aforementioned, we suggest the creation of a research agenda that includes:

- Continued research of the relationship between high quality plan development and implementation and student outcomes
- A qualitative research design to incorporate educator experiences into the analysis of teacher evaluation processes and outcomes.
- Evaluation of the effect of technical assistance and support on the teacher evaluation process in Indiana.
- Additional research on the inter-relationships of classroom, school and district characteristics on teacher evaluation and student outcomes.
- Additional research on school and district resources and supports on teacher evaluation and student outcomes
- Research teacher evaluation impact on IGM vs. non-IGM teachers.

Research Team

- Hannah Bolte, Statistical Consultant/Lecturer, IU Department of Statistics
- Michael Frisby, Statistical Consultant, IU Department of Statistics

- Demetrees L. Hutchins, Management Analyst, IUPUI School of Education
- Gary R Pike, Professor, Higher Education & Student Affairs
- Sarah Pies, Research Associate, INTASS Project, Indiana University
- Hardy Murphy, Clinical Faculty IUPUI, Co-Director INTASS
- Sandi Cole, Director Center on Education and Lifelong Learning IUB, Co-Director INTASS

Special Acknowledgement: The research team acknowledges Hammad Rahman, Data Management Specialist, Indiana Department of Education, for the important contributions made to this research.

Appendix A:

Table 1: History of changes in accountability in Indiana

2011-12	2012-13	2013-14	2014-15	2015-16	2016-17
1. General Assembly enacts annual educator evaluations starting with the 2012-13 school year 2. Districts had one year to pick a model (unless extended contract) and to plan for full implementation. 223 districts had full implementation during the 2012-13 school year 3. First Year of Competitive Performance FY 12	1. First year of implementation of educator evaluations for 223 districts 2. New leadership takes office January 2013 3. Release of RISE 2.5 from IDOE due to ISTEP+ systemic glitches- summer 2013 4. Second year of Competitive Performance Grants FY 13	1. General Assembly eliminates Common Core and builds Indiana College- and Career-Ready Standards and Assessments 2. General Assembly changes the performance grant to $2M for Title I Focus and Priority Schools (not competitive) 3. December of 2013- first public release of final summative evaluations results to public	1. Late release of assessment results and triggers "hold harmless" on accountability and evaluations tied to ISTEP+ due to common drop in test scores 2. January 2015- second year of public release of final summative evaluations results to public 3. $2M for Title I Focus and Priority Schools (not competitive) continues 4. General Assembly adds $30M based on state assessment proficiency only 5. IDOE begins onsite monitoring of evaluation plan implementation due to ESEA Flexibility Waiver	1. First year of new Indiana College- and Career- Ready Standards and new vendor for assessments (Pearson) 2. General Assembly continues $30M based on state assessment proficiency only and continues $2M for Title I Focus and Priority Schools (not competitive) 3. Testing results come in fall to schools and public 4. IDOE released third year of final summative evaluation results to public in July 2016 5. Hold Harmless Provision taken for A-F letter grades and teacher evaluations due to new assessment/standards 6. First year of new A-F Accountability Model for schools released in the late fall of 2016 7. ESSA passed by Federal Government; evaluations no longer required through Federal Government 8. IDOE ends monitoring of district evaluation plan implementation after ESSA is passed	1. Third year of new Indiana College- and Career-Ready Standards and new vendor for assessments (Pearson) 2. General Assembly continues $40M based on state assessment proficiency only and continues $2M for Title I Focus and Priority Schools (not competitive) 3. New ISTEP+ Panel meeting in the fall of 2016 to provide suggestions for new ISTEP+

Table 2: Model descriptions

MODELS USING STUDENT AND DISTRICT DEMOGRAPHICS AS PREDICTORS	Full Model	Elem vs. Middle	Middle 68% of State	Middle 68% State + Elem vs. Middle	Middle 68% of District	Middle 68% District + Elem vs. Middle	ELA Accountable	MATH Accountable
FILTERS								
Student_attended_at_least_162_days	No for aggregates Yes for all others	No for aggregates Yes for all others	Yes	Yes	Yes	Yes	No	No
No_conflicting_records_on_district	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No_conflicting_records_on_school	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No_conflicting_records_on_grade	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Grades_4/5_and_7/8_only	No	Yes	No	Yes	No	Yes	No	No
Performance_in_Middle_68%	No	No	State	State	District	District	No	No
CONTROLS								
Primary_exceptionality_effects	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Grade_effects	Yes for ISTEP Score No for all others	Yes for ISTEP Score No for all others	Yes for ISTEP Score No for all others	Yes	Yes for ISTEP Score No for all others	Yes	No	No
District_effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nested_school_effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
ALTERNATE CONTROLS FOR DISTRICT EFFECTS ON TEACHER OUTCOMES								
Primary_exceptionality_effects	No	No	No		No			
Grade_effects	Yes for ISTEP Score No for all others	Yes for ISTEP Score No for all others	Yes for ISTEP Score No for all others		Yes for ISTEP Score No for all others			
District_effects	No	No	No		No			
School_effects	Yes	Yes	Yes		Yes			
Nested_teacher_effects	Yes	Yes	Yes		Yes			

Figure B: 4th/5th Classroom Mean Growth Percentile by FRL Decile

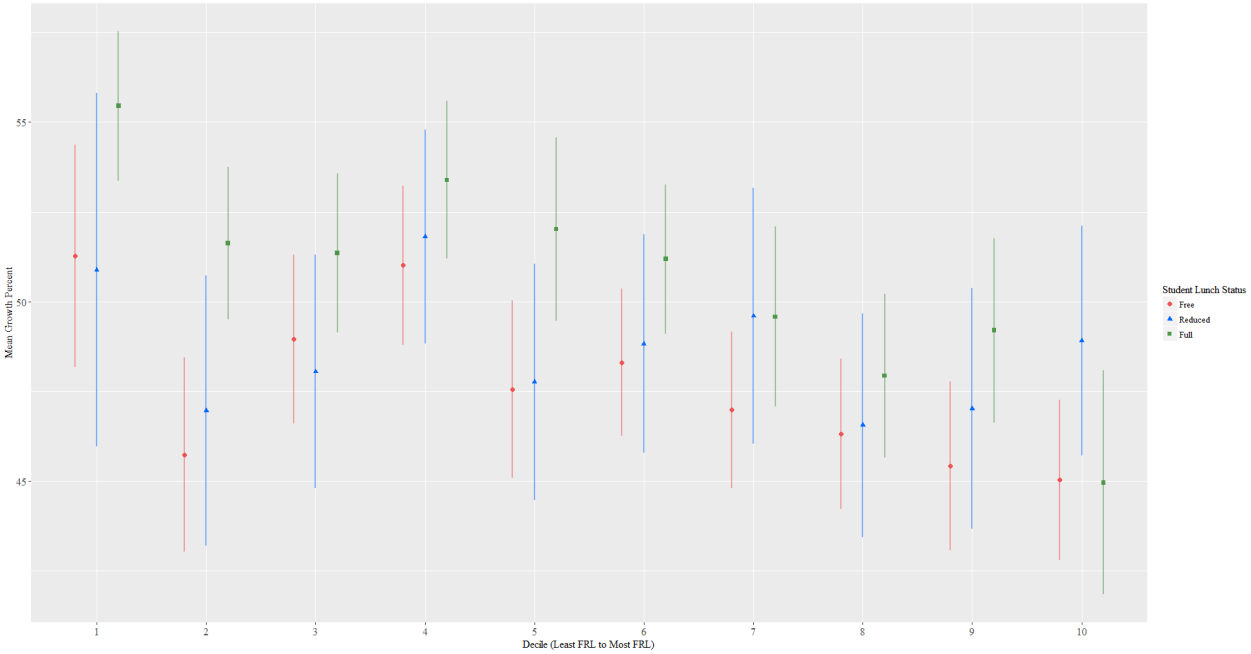

Figure C: 4th and 5th Classroom Mean Scale Score by Decile