

CLASSROOM PRACTICES AND STUDENT PROGRESS: RELATIONS BETWEEN
RATINGS OF CLASSROOM PRACTICES AND INDICATORS OF STUDENT
LEARNING IN READING

by

ERIN M. FUKUDA

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2016

DISSERTATION APPROVAL PAGE

Student: Erin M. Fukuda

Title: Classroom Practices and Student Progress: Relations Between Ratings of Classroom Practices and Student Learning in Reading

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy, and Leadership by:

Gina Biancarosa	Chairperson
Gina Biancarosa	Advisor
Keith Hollenbeck	Core Member
Joe Stevens	Core Member
Juliet Baxter	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School
----------------	-----------------------------

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2016

© 2016 Erin M. Fukuda
This work is licensed under a Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License.



DISSERTATION ABSTRACT

Erin M. Fukuda

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

June 2016

Title: Classroom Practices and Student Progress: Relations Between Ratings of Classroom Practices and Indicators of Student Learning in Reading

The state of Oregon, like many states, requires its districts' teacher evaluation systems to include measures of student learning as well as a measure of teachers' professional practice. State guidelines require use of state test data in assessed grades as one of the measures, but allow districts flexibility in which additional assessments to use and which source of information to prioritize when evaluating teachers. This study used existing data from one school district to compare students' performance on a state reading and literature assessment to their performance on reading curriculum-based measures, and the degree to which measures of teaching practices relates to both types of student outcomes. Results are interpreted with consideration of how the district implements their measure of teaching practice. Results from this study may help inform decisions the district will face as they continue to refine their teacher evaluation system in accordance with state guidelines, while elucidating challenges that such systems pose.

CURRICULUM VITAE

NAME OF AUTHOR: Erin M. Fukuda

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene

DEGREES AWARDED:

Doctor of Philosophy, Educational Leadership, 2016, University of Oregon

Master of Education, Education Studies, 2005, University of Oregon

Bachelor of Science, Education Studies, 2004, University of Oregon

AREAS OF SPECIAL INTEREST:

Teaching and learning

Metacognition

Development of expertise

Classroom dispersion models

PROFESSIONAL EXPERIENCE:

Research Associate, Educational Policy Improvement Current, 2013-Current

Research Assistant (GTF), IES Grant CFDA 84.305A. Dr. Keith Zvoch and Dr.

Joseph Stevens, Co-Principal Investigators, University of Oregon, 2010-2013

Academic Liaison, Springfield Public Schools, 2009-2010

Teacher, Springfield Public Schools, 2006-2009

GRANTS, AWARDS, AND HONORS:

Gates Millennium Scholar, 2000-2015

Jean DuRette Professional Development Scholarship, 2012

PUBLICATIONS:

Lench, S., Fukuda, E., & Anderson, R. (2015). *Essential skills and dispositions: Developmental frameworks for collaboration, communication, creativity, and self-direction*. Lexington, KY: Center for Innovation in Education at the University of Kentucky.

Conley, D., McGaughy, C., Davis-Molin, W., Farkas, R., & Fukuda, E. (2014). *International Baccalaureate Diploma Programme: Examining college readiness*. Bethesda, MD, USA. International Baccalaureate Organization.

Beghetto, R., Barbee, B., Brooks, S., Franklin-Phipps A., Fukuda, E., Hood, D., Raza, N., Uusitalo, N., & White Eyes, C. (2013). Light bulbs, cat hair, and Bill Evans: Representations of creativity in popular media. *Psychology of Popular Media Culture*, 2(4), 188-206.

ACKNOWLEDGMENTS

I would like to thank the members of my committee as well as Dr. Kathleen Scalise for their patience and guidance that helped shape a vague idea into a fully executed study. In addition, special thanks to Dr. David Conley, who informed my view of policy to include the implicit effects on behavior and thought behind every explicit rule or regulation. To the Bill and Melinda Gates Foundation and a certain high school teacher, without your support, I would not have had the opportunity to traverse and understand the multiple facets of education that influence the work that teachers in every classroom and I engage in. Last, yet of the greatest importance, I wish to express sincere gratitude to my family, and friends who have become family, for their ongoing patience and unwavering support throughout a long and challenging journey.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION	1
II. LITERATURE REVIEW.....	2
National Landscape of Teacher Evaluation	2
Contributions of Teachers as Inputs	5
Input-output Contributions.....	6
Input-output Limitations	7
Contribution of Teachers Through Processes	11
Process-outcome Contributions	11
Predictive Validity Across Settings	13
Observer Reliability	14
Process-outcome Limitations.....	15
The Role of Context.....	16
The Current Study.....	20
Research Questions.....	22
II. METHOD.....	24
District Population	24
Participants.....	25
Measures	27
District's Measure of Teaching Practices.....	27
Oregon Assessment of Knowledge and Skills (OAKS): Reading/Literature ..	29
easyCBM™	31

Passage Reading Fluency (PRF).....	31
Multiple-choice Reading Comprehension (MCRC).....	33
Data Screening.....	34
Missingness.....	34
Testing Assumptions.....	36
Analyses.....	36
Research Questions 1 Through 3	37
Research Questions 4 and 5	40
III. RESULTS	42
Analytic Sample.....	42
Data Screening.....	45
Missingness.....	47
Testing Assumptions.....	51
Model Specification.....	52
Research Question One.....	53
Research Question Two	55
Research Question Three	56
Research Question Four.....	60
Research Question Five	61
IV. DISCUSSION.....	65
Limitations	65
Uniqueness and Similarity Among Multiple Reading Measures.....	67
Contribution of Classroom Practices to Student Outcomes and Growth.....	69

Dispersion Patterns and the Matthew Effect.....	74
Future Studies	75
V. CONCLUSION.....	77
APPENDICES	79
A. DISTRICT'S STANDARDS FOR TEACHING PRACTICES.....	79
B. DISTRICT'S SUMMATIVE TEACHER EVALUATIONS FORM.....	81
C. DESCRIPTION OF CBM VOCABULARY ASSESSMENT AND MSSINGNESS.....	83
D. HISTOGRAMS OF CLASSROOM PRACTICE DOMAIN AVERAGES.....	85
E. HISTOGRAMS, Q-Q PLOTS, AND BOXPLOTS OF OAKS GAINS AND MCRC AND PRF SCORES BY SEASON.....	86
F. SCATTER PLOTS OF CLASSROOM READING MEANS AGAINST DOMAIN SCORES ACROSS SEASONS.....	91
G. BIVARIATE SCATTER PLOTS: PREDICTED READING SCORES AGAINST RESIDUALS	94
H. SCATTERPLOTS OF MAHALANOBIS DISTANCES AGAINST EXPECTED CHI-SQUARE DISTRIBUTION FOR OAKS GAINS, MCRC, AND PRF	95
REFERENCES CITED.....	96

LIST OF FIGURES

Figure	Page
1. Scatter plots of average classroom PRF means by domain standards, for fall, winter, and spring.....	91
2. Scatter plots of average classroom MCRC means by domain standards, for fall, winter, and spring.....	92
3. Scatter plots of average classroom OAKS gains by domain standards, for fall, winter, and spring.....	93

LIST OF TABLES

Table	Page
1. District demographics: Number and percentage by year	24
2. Start-of-year, fourth and fifth grade student enrollment in district school sites	25
3. Combined number and percent of fourth and fifth grade students designated as LEP or took the OAKS extended assessment, by year	27
4. OAKS Reading/Literature achievement standards	30
5. Proportions of 4th and 5th grade students representing select demographic categories in district population and respective samples	43
6. Summary of classroom composition: Class size and proportion of students by demographic indicator, $N = 945$	44
7. Range of classroom means, standard deviations, and missingness for reading scores/gains	46
8. Descriptive statistics of reading outcome data by assessment occasion.....	47
9. Proportion of participant students missing assessment data, $N = 945$	47
10. Differences in MCRC performance between students with and without data for other MCRC test occasions	48
11. Relationship between school and missing data, by assessment time point.....	50
12. Variance in OAKS attributed to differences between classrooms.....	54
13. Variance in PRF and MCRC attributed to differences between classrooms	54
14. Correlations between estimates of average classroom OAKS gains, CBM spring scores, and CBM growth ($N = 35$).....	55
15. Proportion of variance explained after the addition of classroom practices.....	56
16. OAKS gains and the contribution of classroom practices	57
17. MCRC outcomes, growth, and the contribution of classroom practices	58
18. PRF outcomes, growth, and the contribution of classroom practices.....	59

19. Variance in dispersion patterns of reading scores over time, between classrooms	60
20. Proportion of dispersion pattern variance explained after the addition of classroom practice scores.....	61
21. PRF fall dispersion, dispersion patterns over the year, and the contribution of classroom practices, ($N = 35$)	62
22. MCRC fall dispersion, dispersion patterns over the year, and the contribution of classroom practices, ($N = 35$)	63
23. Proportion of students missing vocabulary data, by data patter, $N = 945$	82
24. Association between school and missing vocabulary data, by assessment occasion.....	83

CHAPTER 1

INTRODUCTION

Teachers influence student learning. Inherent within this fundamental assumption is the idea that differences between teachers lead to differences in student learning. This foundational belief about the teaching profession motivated the development of teacher preparation courses (Labree, 2008), establishment of teacher licensure (Labree, 2008), and enactment of policies such as the 'Highly Qualified' hiring requirements within the No Child Left Behind Act (2001). For almost half a century, researchers have strived to understand the relationship between teachers and student learning through two primary perspectives: 1) teachers and their characteristics as what goes *into* a classroom, or 2) teacher practices that inform the *process* of learning.

This study examined one school district's teacher evaluation system by utilizing an input-output lens to estimate learning within classrooms as represented by different reading measures and a process-outcome lens to examine relationships between student learning and teacher evaluation scores, used as proxy indicators of classroom practice. Additionally, findings are interpreted with consideration of how the district implements its teacher evaluation system.

CHAPTER II

LITERATURE REVIEW

Applying a common analogy (Teddlie & Reynolds, 2000), consider a classroom as a box that produces learning as its *output* or *outcome*. *Input-output* research examines student learning in relation to what we're *putting into* the box (e.g., who the teacher is, his or her years of experience, his or her degree(s), textbooks). *Process-outcome* research examines student learning in relation to the *processes occurring within* the box (e.g., what the teacher does, teacher-student interactions, instructional practices, how textbooks are used). Each approach to examining how teachers influence student learning has independently contributed pieces to a critical understanding that (a) teachers do contribute to student outcomes in varying degrees, and (b) differences in some teaching practices can contribute to this variation. Though an increasing number of states' teacher evaluation *policies* reflect this understanding, integrating both pieces into a cohesive system that measures and improves (or sustains) teaching quality is more complex in *practice*.

National Landscape of Teacher Evaluation

Measures reflecting the input-output and process-outcome perspectives on the contribution of teachers have gained increased national attention within the last several years. Changes in federal education policy allowed approved states flexibility in meeting certain requirements of the No Child Left Behind Act (2001), if in exchange the states and their local education agencies agreed to address four principles regarding instructional quality and student academic achievement (U.S. Department of Education, 2012). One principle, *Supporting Effective Instruction and Leadership*, requires

development of a teacher evaluation and support system that includes use of “multiple valid measures in determining performance levels, including as a significant factor data on student growth for all students (including English Learners and students with disabilities), and other measures of professional practice (which may be gathered through multiple formats and sources, such as observations based on rigorous teacher performance standards, teacher portfolios, and student and parent surveys)” (U.S. Department of Education, 2012, p.3). Such a system must also, “provide clear, timely, and useful feedback, including feedback that identifies needs and guides professional development” (U.S. Department of Education, 2012, p.3). To meet these requirements, local education agencies need to identify both the degree to which teachers contribute to their students’ learning as well as descriptions of their practices to guide professional development: in other words, both inputs and processes. Currently, 43 states, the District of Columbia, and Puerto Rico have been approved for NCLB flexibility (U.S. Department of Education, n.d.).

Changes in state policies reflect pressure by both the federal government and policy and research reports (e.g., Darling-Hammond, 2012; Bill & Melinda Gates Foundation, 2013; National Council on Teacher Quality, 2009) to incorporate both measures of the effects of teachers as inputs, *and* their classroom processes. Though teacher observation tools have commonly been used for teacher evaluation (Goe, Bell, & Little, 2008), 44 states and the District of Columbia now require use of observation measures for teacher evaluation purposes, compared to 30 states in 2009 (Doherty & Jacobs, 2013; National Council on Teacher Quality, 2009). Over the same period, the number of states requiring teacher evaluations to include measures of student

achievement has climbed from 15 states to 40 states plus the District of Columbia, as of 2013, (Doherty & Jacobs, 2013). Though the federal government endorses the use of multiple measures, limited guidance has been provided about how to incorporate both into a cohesive and valid teacher evaluation system. Twenty-three states and the District of Columbia attempted to address this issue by specifying the weight which student achievement or growth is given within teacher evaluation systems (Doherty & Jacobs, 2013). Required weights range from 20% to 50% or more of a teacher's evaluation and policies vary by the types of student assessments utilized, the grade levels identified, the number of years of achievement data analyzed, and how student outcome models are specified (Doherty & Jacobs, 2013). Little research beyond the work of Mihaly, McCaffrey, Staiger, and Lockwood (2013) has been done to examine the effect of different weighting schemes.

While specifying weights might bring transparency and clarity to how evaluation decisions will be made, explicit connections between districts' measures of teaching practices and the student academic outcomes selected for teacher evaluation often remain unknown. The two types of measures might provide redundant conclusions on the contribution of teachers, opposing conclusions, or somewhat complementary conclusions while still providing unique information, with the latter as the obvious policy intention. For example, a teacher evaluator might utilize student achievement or growth data to inform his or her perceptions of a teacher's success, and turn to evidence of observed instructional practices to explain such success and/or provide direction for improvement. Using the information in this way assumes the teaching practices described by classroom observation protocols contribute to the specific student outcomes included within the

teacher evaluation system. Though a sizable amount of attention has been paid to the types of measures to use for teacher evaluation, far fewer studies have examined the relations between measures of teaching practices and student outcomes on the specific academic assessments included within a district's teacher evaluation system.

Setting the stage for the current study, the question of relations between input-output and process-outcome measures are first considered conceptually by describing defining properties, contributions, and limitations of input-output research and of process-outcome research. Then, consideration is given to contextual elements that might foster or inhibit the relationship between input-output and process-outcome measures. The study is then described as a means of investigating this relationship empirically, by examining relations between student learning over time and the quality of classroom practices as determined by a school district's teacher evaluation system.

Contribution of Teachers as Inputs

Early efforts to examine the outcomes of education as a result of what we put *into* schools have been described as input-output studies (Teddlie & Reynolds, 2000; Hanushek, 2010). Such work attempts to model direct relationships between classroom inputs and education 'outputs' such as student learning (Hanushek, 2010; Monk, 1992). Inputs are the people and things *put into* classrooms (e.g., 30 students, teacher A, curriculum) and their related characteristics (e.g., students from low income families, a teacher with a Master's degree, scripted curriculum). Beginning with the Coleman report (Coleman et al., 1966), such studies have made a sizable contribution to the field of education by empirically examining the fundamental belief that teachers influence student learning and the extent to which that influence varies (Hanushek, 1986).

Input-output contributions. Many studies have indicated that teachers do influence student achievement, as well as change in achievement, and that a noticeable amount of variation in the degree of teacher influence exists (e.g., Hanushek, 1992; Huang & Moon, 2009; Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Sanders & Horn, 1998). Nye and colleagues (2004) synthesized 17 comparison analyses of teacher effect magnitude and reported that 7% to 21% of variance in student achievement gains was related to teachers. The authors' own analysis, utilizing more complex multilevel modeling and a pre-existing dataset of students randomly assigned to classrooms, resulted in teacher contributions to students math achievement gains near the 11% median of the other studies. Although 11% may not seem sizable, the authors noted that accounting for 10% of variance in student achievement translated into an increase of approximately one third of a standard deviation in student achievement gains for every one standard deviation increase in teacher effectiveness. Teacher contributions were smaller for reading achievement gains, ranging between 6% to 8%, yet for both subjects the variance due to differences in teachers was noticeably larger than the variance between schools.

Beyond providing empirical evidence of teachers' contributions to student learning, input-output research introduced consideration of student outcomes as a formal aspect of teacher evaluation. For example, the Tennessee Value-Added Assessment System utilizes statistical modeling of student test scores over time and was originally developed to determine school system, school, and teacher effectiveness (Sanders & Horn, 1998). This system and other value-added models view each teacher as a classroom input and the teacher's average classroom achievement, gain, or growth as the

output, with the added caveat that an individual teacher's quality can be determined by comparing his or her average “effect” to that of other teachers. Though some states specifically use value-added models (e.g., Tennessee, Florida), 80% of states require some indication of student academic learning as part of teacher evaluations. (Doherty & Jacobs, 2013). Public education began as a provision initiative (Fowler, 2009) and the input-output perspective necessitated examinations of the effects on students in relation to what is provided. Input-output research provided evidence of teachers’ contributions to student learning at-large and encouraged consideration of the extent to which each teacher contributes to their students’ learning. However, quantifying teacher contributions in a manner that allows for valid interpretations of teacher effectiveness and explaining variation in effectiveness continues to pose significant challenges.

Input-output limitations. Research examining value-added models has illuminated some factors that could affect the validity and reliability of statistically equating teacher quality with their students’ outcomes. Though initial teacher value-added models utilizing vertically scaled assessments reported relatively stable results over time (Sanders & Horn, 1998), controversy remains regarding model reliability and validity. Accounting for student background characteristics showed little impact on Tennessee Value Added Assessment System estimated teacher effects (Ballou, Sanders, & Wright, 2004). However, Newton, Darling-Hammond, Haertel, and Ewart (2010) found significant negative correlations between teachers’ rankings and the proportion of their students who were English language learners, received free lunch, or were Hispanic even after accounting for student demographics within ordinary least squares regression and multilevel models. Hill, Kapitula, & Umland (2011) also reported correlations

between teachers' value-added scores and the population of students they teach.

Moreover, there is some evidence that individual teachers' value-added scores may vary as a result of ceiling effects or other test properties (Kodel & Betts 2009), the timing and environmental conditions of test administration (Corcoran, Jennings, & Beveridge, 2011; Papay, 2011), and measurement error of the student achievement assessments utilized (Kodel & Betts, 2009; Papay, 2011). Some of these factors were raised as possible explanations for differences in teacher value-added estimates using either criterion- or norm-referenced test data (Stuit, Berends, Austin, & Gerdeman, 2014). Lockwood and McCaffrey (2007) provided further evidence in the role assessment selection could play in teacher value-added estimates, reporting variation in teachers' value-added effects for different math achievement measures to be larger than the variation of estimates across teachers. Additionally, the variation in value-added effects as a result of different math achievement measures was greater than the variation due to model specification choices. These findings suggest that different tests may illustrate different degrees of student learning independent of how the results are statistically or descriptively analyzed. Though the input-output studies described here focused on the validity and reliability of value-added models, they describe factors such as classroom demographics and properties of student outcome measures that may have ramifications for any teacher evaluation system incorporating measures of student learning.

Another primary limitation of input-output research has been a lack of consensus in determining the characteristics of teachers that are associated with student achievement or achievement gains. Researchers have reported conflicting results regarding the significance of teacher certification type. For example, when Croninger, King Rice,

Rathbun, & Nishio (2007) compared types of certification grouped as “regular or alternative” vs. “none, temporary, provisional, emergency or probational” they found certification status was not a significant predictor of either reading or math achievement (p. 316). However, when Palardy and Rumberger (2008) examined student achievement gains, full certification was found to be a significant, positive predictor of student learning in both reading and math. Such mixed findings are typical of these efforts.

Similarly, conflicting results have been reported on the influence of teacher education when defined as possession of an advanced degree. Though an earlier meta-analysis of such studies (Greenwald, Hedges, & Laine, 1996) found possession of a Masters degree to be a positive factor related to student outcomes, several, more recent studies have reported advanced degrees to be a non-significant factor (Chingos & Peterson, 2010; Coninger et al., 2007; Nye et al., 2004; Rivkin et al, 2005).

Hypothetically, these conflicting findings could reflect an increase over time in the proportion of teachers with a Masters degree, to the point where the variable is no longer useful to differentiate teachers and their class outcomes. One study examined teacher education in a more nuanced way by exploring relationships between features of teacher preparation programs and student outcomes (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009). Among their findings, the authors report that teachers from programs that provide opportunities to engage in the practices of teaching and with more oversight during student teaching showed larger student gains during their first year of teaching. Collectively, these mixed results raise a need for more fine-grained examinations of differences in teacher training beyond what is reflected by broad classifications such as degrees obtained.

Teaching experience, defined as the number of years a teacher has taught, has also been a consistently inconsistent predictor of student outcomes. An early review of studies analyzing education processes identified 109 analyses of teaching experience. Only 40 studies reported statistically significant findings, and seven of those studies found negative effects (Hanusheck, 1986). Some of this inconsistency may be due to the more nuanced relationship of teaching experience and student outcomes. Recent studies have reported non-significant relationships between the aggregate number of years a teacher has taught and student achievement gains (Chingos & Peterson, 2010; Huang & Moon, 2009), but the number of years a teacher has taught at their current grade level has been found to be significant (Huang & Moon, 2009). Additionally, the relationship between student achievement gains and teacher experience may differ when experience is examined in terms of broader phases of a teacher's career. Chingos and colleagues (2010) reported increased gains in student achievement associated with increased years of teaching experience for early career teachers, while slight decreases in student achievement gains were reported for teachers near the end of their career.

Input-output studies including those using value-added models have contributed empirical examinations of the fundamental assumption that teachers influence student learning and whether the degree of influence differs. In doing so, such studies have encouraged consideration of the effects of teachers and emphasized the end goal of improving student outcomes. However, a reliable process for measuring the effect of individual teachers remains unclear and the input-output approach has produced inconsistent findings regarding what it is about some teachers that allow them to influence student learning more than others. This research approach does not address the

critical point of determining what each teacher does or does not do to influence student learning. Without understanding which teaching practices are related to differences in student outcomes, little feedback can be provided to sustain or improve teachers' practice and students' learning.

Contribution of Teachers Through Processes

Identifying teacher behaviors, including instructional practices, associated with student academic outcomes has been the defining feature of process-outcome (also known as process-product) research (Medley, 1977; Teddlie & Reynolds, 2000). As early as the 1950s, classroom observation was utilized as a means of identifying practices associated with teachers whose students exhibited academic success (Brophy & Good, 1986; Medley, 1977). The depth and breadth of this work is evidenced by multiple reviews of process-outcome research that summarize consistent findings (e.g., Medley, 1977; Rosenshine, 1971), unexpected findings (Brophy & Good, 1986), or overarching limitations and criticisms (e.g., Gage & Needels, 1989; Rosenshine, 1971). For example, Rosenshine (1971) described consistent positive relationships between student achievement gains and frequent use of probing follow-up questions, while also positing non-linear relationships with other teacher behaviors, and calling for more consideration of contextual elements that may inform interpretation of teacher behavior. The extensive body of early process-outcome research provided the field of education a path toward improving student outcomes by examining classroom practices.

Process-outcome contributions. One of the primary contributions of process-outcome research has been the development of tools that describe teaching practices that generalize across contexts and can be used to guide classroom observations. The initial

wave of process-outcome studies informed the early development of a multitude of classroom observation tools (Brophy & Good, 1986; Medley, 1977), as documented by the numerous volumes of Simon and Boyer's anthology series *Mirrors for Behavior* (1967 – 1974). The focus of observation tools evolved from affective systems that attended to emotional climate to more comprehensive measures of practices related to cognitive instructional objectives (Brophy & Good, 1986; Simon & Boyer, 1967), changes that are reflected in modern teaching standards and classroom observation frameworks used today.

For example, the Interstate Teacher Assessment and Support Consortium (InTASC) affiliated with the Council of Chief State School Officers developed Model Core Teaching Standards that are “based on our best understanding of current research on teaching practice” and “describe what effective teaching that leads to improved student achievement looks like” (Council of Chief State School Officers, 2013, p. 3-4). A review of articles addressing teacher knowledge, teacher disposition, or teacher performance, and published in peer-reviewed journals between 1990 and 2011 was conducted specifically to distinguish areas within the updated InTASC standards that were supported by empirical research from those supported by “normative agreements” (Youngs, 2011). Similarly, the Danielson Framework for Teaching and the Praxis III classroom observation assessment it was based on utilized approximately 150 educational theory and research documents including literature reviews on effective beginning teaching (Reynolds, 1992 is one example) during the development process (Danielson, 2007; Dwyer, 1994). Scales within another common measure of classroom processes, the Classroom Assessment and Scoring System (CLASS), were either adapted from existing

observation measures or based on multiple literature reviews as cited in La Paro, Pianta, & Stuhlman (2004). Though many other examples exist, the teaching framework and observation tools described here illustrate the breadth of process-outcome work and the extent of influence it has had on how good teaching practices are defined and measured today.

Conceptually, these observation tools are helpful in that they summarize good teaching practices in more general terms to facilitate implementation across a variety of settings. However, the strength of the research base used to develop observation tools is of little concern if the descriptions of good practice are (a) too general to capture differences between teachers that relate to student outcomes, (b) too specific to use across contexts, or (c) are not clear enough to be used reliably. What follows are examples of empirical predictive validity, generalizability, and reliability findings with respect to two commonly utilized classroom observation measures, the CLASS and the Danielson Framework for Teaching.

Predictive validity across settings. More recent process-outcome studies reported promising indications of predictive validity after examining relations between teachers' performance on classroom observation tools and student outcomes data. The Instructional Support component of the CLASS significantly predicted pre-kindergarten students' post-test (Mashburn, et al., 2008) and gain scores (Howes, et al., 2008) in receptive language, expressive language, rhyming, and letter naming, as well as post-test scores in math skills (Mashburn, et al., 2008). In pre-kindergarten classrooms where teachers displayed higher Emotional Support, their students exhibited decreased behavior problems and gains in social competence (Mashburn, et al., 2008). Similarly, Kane and

colleagues (2011) examined the relation between multiple cohorts of third to eighth grade students' achievement data and their teachers' Overall Classroom Practice score, as measured by a modified version of the Danielson Framework. For all grades, Overall Classroom Practices scores predicted students' end-of-year reading achievement after accounting for prior achievement. The same held true for math, excluding third grade for which student achievement data in math were not available. Such work begins to address the assumption that the teaching practices described by these tools relate to student outcomes and are generalizable.

Validity studies have also reported that relationships between *categories* of teaching practices and student outcomes may generalize across a variety of settings. Though the CLASS was originally designed around two constructs, Emotional Support and Instructional Support, further validation efforts established a three-factor model to better represent the variation in teacher-student interactions within preschool, elementary, and secondary U.S. classrooms (Allen, et al., n.d.; Hamre, Pianta, Mashburn & Downer, 2007), as well as kindergarten classrooms in Finland (Pakarinen, et al., 2010). Findings indicated that the three constructs of Emotional Support, Classroom Organization, and Instructional Support were relevant across different grades and not necessarily limited to cultural contexts within the United States.

Observer Reliability. Underlying the ability to establish relationships between teaching practices and student outcomes is the degree to which observers can reliably score what they see. For example, if two observers have different interpretations of what exemplifies “use of questioning and discussion techniques” (Sartain, Stoelinga, & Brown, 2011, p.6), they might indicate different levels of teacher performance even if they were

to observe the same lesson. Inconsistency between raters could alter the extent to which teaching practices relate to student outcomes. A high degree of reliability between observers provides us with more confidence in the validity findings previously described.

In their study of the CLASS, utilizing data from four large-scale national and regional projects focused on classroom-observation, Hamre and colleagues (2007) reported that certification of participating observers required at least 80% agreement with master coders on videotaped classroom footage. Though Kane et al. (2011) described a similar training process with respect to the Danielson Framework utilized in the Cincinnati Public Schools Teacher Evaluation System, reliability criteria were not provided. Efforts to examine the implementation of the Danielson Framework as part of the Chicago Public Schools Excellence in Teaching Pilot included analyses of rater reliability between school administrators and expert observers conducted after training and evaluations occurred (Sartain et al., 2011). In general, school administrators and expert observers provided similar proportions of Unsatisfactory and Basic ratings, but the proportion of Distinguished ratings was higher for administrators than expert observers. When a high level of practice was observed, administrators were more likely to identify it as Distinguished, while expert observers were more likely to call it Proficient. Reliability findings from these studies suggest a generally shared understanding of the teaching practices within the CLASS and the Danielson Framework and what different degrees of implementation of those practices look like. Modest to strong reliability results also help support the credibility of previously described relationships established between student outcomes and the CLASS and Danielson descriptions of good practices.

Process-outcome limitations. Despite increasing efforts to validate tools like the

CLASS and the Danielson Framework for Teaching, some findings indicate limitations in the degree to which a single framework or measure of teaching practices is generalizable across various contexts. For example, examinations of the CLASS tool for secondary schools (CLASS-S) found an additional factor, Regard for Adolescent Perspectives, related to end-of year test scores, after accounting for class size, student demographics and prior achievement (Allen et al., n.d.). The Regard for Adolescent Perspectives scale was a secondary-specific form of the Regard for Students Perspective Scale, which took into consideration “adolescents’ needs for a sense of autonomy and control, for an active role in their learning, and opportunities for peer interaction” (Allen, et al., n.d., p.2). Such a modification of an observation instrument suggests a need for context-relevant flexibility in how some teaching practices are operationally defined.

Potential limits to the generalizability of scales across subject areas have also been reported. When comparing teachers with similar Overall Classroom Practices scores on a Danielson-based observation tool, different types of teaching practices related to student achievement in different subject areas (Kane et al., 2011). Classroom management was relevant with respect to students’ math achievement but not reading achievement, while the ability to engage students in discussion and questioning was related to student achievement in reading but not math (Kane et al., 2011). Despite evidence relating a teaching practice to student achievement in one subject, assumptions that the relationship is generalizable to other subjects may not be tenable.

The Role of Context

The mixed predictive validity findings with respect to teacher observation tools underscore the context-dependent nature of validity arguments (Kane, 1992; Messick,

1989). Applying Messick's (1989) validity frame, the findings from each study do not definitively determine the validity and reliability of the observation tools themselves. Instead the validity and reliability findings are in reference to scores on those observation tools for specific participants within each particular setting. The need to modify operational definitions of teaching practices and the inconsistent predictive validity findings across subject areas may reflect "lack [of] validity for *uniform* use across the varied circumstances but not necessarily lack [of] validity for differential use or even validity of meaning in specific contexts" (p.15, emphasis in original). Like any assessment score, a classroom observation rating is "a function not only of the items, tasks, or stimulus conditions, but of the *persons* responding and the *context* of measurement" (Messick, 1989, p.14, emphasis in original). Contextualized decisions including how to operationally define teaching practices for different student populations, grade levels, and subject areas have ramifications for the validity of teacher observation scores within the unique contexts of individual schools and districts. Despite promising validity evidence of classroom observation scores, one of the most challenging limitations of process-product research and the observation tools might be dependence on the implementation decisions made in context.

As Sartain and colleagues (2011) noted, "Policymakers face a range of decisions that include, but go beyond, which observation tool to select. Many of these decisions have the potential to contribute to, or impede, successful implementation at scale" (p.29). As part of their report summarizing findings from teacher evaluation system pilot, five key considerations when designing a teacher evaluation system were discussed: how classroom observations inform final evaluation decisions, observation logistics, training,

principal engagement, and accountability and feedback for evaluators. What follows is an example of the impact that different design decisions, whether purposive or not, can have on the validity and reliability of classroom observation and teacher evaluation results.

The extent to which training and ongoing support is provided to observers and evaluators is one of the contextual factors raised by several researchers (e.g., Kane et al., 2011; Kimball, White, Milanowski, & Borman, 2004; Sartain et al, 2011). The modest to strong agreement between observers, reported by several studies (see Hamre et al., 2007; Kane Taylor, Tyler, Wooten, 2011; Sartain et al., 2011), utilized data from teacher evaluation systems similar to what Kane et al., (2011) described as “high-quality”, where substantial training was provided to observers, and teachers may have been observed multiple times over the course of the school year. For Kane et al. (2011) and Hamre et al. (2007), all observers were required to participate in extensive training and then demonstrate their ability to achieve the same rating outcomes as experts before they were allowed to observe teachers independently. School administrators participating in the Chicago Excellence in Teaching Pilot (Sartain et al., 2011) received three days of summer training, four half-day professional development sessions throughout the school year, and regular opportunities to meet with other principals to discuss implementation issues. The three external observers for the Excellence in Teaching Pilot received even more extensive training including the initial three-day training, follow-up support for specific components of the Framework, and practice observation opportunities within classrooms. Without these high-quality training efforts to obtain consistency between observers, the classroom observation ratings might have been less reliable.

In contrast, lack of an emphasis on training was identified among potential reasons for the inconsistency between teachers' observation ratings and student achievement outcomes in Washoe County (Kimball et al., 2004). The teacher observation data utilized represented seven evaluation components from Danielson's Planning and Preparation, and Instruction domains. Out of the nine analytic combinations of teacher evaluation scores predicting either third, fourth, or fifth grade student performance in reading, math, and reading/math combination, only four were statistically significant. In addition to research design limitations utilizing only seven of the 23 teacher evaluation components and issues with student achievement data, the researchers described aspects of the evaluation system adoption that might have impacted its validity. Based on prior research with the district, the researchers noted that evaluator emphasis on improving staff morale through praise and growth might have lowered the reliability of ratings, due to the low stakes application of the teacher evaluation system. Other contextual factors were the inconsistency of evidence gathered by evaluators, lack of emphasis on evaluator training support to ensure consistency and accuracy, and the lack of differentiation in standards for content-specific pedagogy (Kimball et al., 2004).

The considerations required to create a valid teacher evaluation system are numerous. Through decades of process-outcome research, classroom observation tools continue to be developed, tested, and refined to help schools and districts identify the presence of teaching practices that relate to student outcomes. Yet, adopting a research-based and tested *tool* is not enough. Districts have a multitude of implementation and integration decisions to make, each of which can affect the validity and reliability of their *evaluation system*. The degree and consistency of training and support throughout the

adoption and use of teacher observation tools can play a key role in how meaningful results of those tools are to teachers and administrators. Decisions around how to model student learning have the potential to influence results. Additionally, both input-output and process-outcome research suggest that the choice of student outcome measures matter with respect to both the amount of student learning attributed to teachers and the strength of relationship between observed instructional practices and those student outcomes.

The current study. These are among the considerations required of school districts across the country, including those within the state of Oregon. During the 2012-2013 and 2013-14 school years, 14 Oregon school districts piloted one of two methods for combining student learning and growth in summative evaluations of teachers: either a weighted percentage, or a matrix determining a “summative rating for professional practice and professional responsibilities correlated with a score based on progress toward student learning and growth goals and aligned with a professional growth plan” (Oregon Department of Education, 2013b, p.1). The results from the pilots were used to refine the State’s guidelines for teacher evaluations, and were included in an amendment to meet Flexibility waiver requirements. All school districts were required to develop or modify their teacher evaluation systems to meet State evaluation requirements and begin implementation during the 2013-2014 school year. Each district needed to include evidence of both teachers’ practices and their students’ learning on multiple measures. Decisions on what to include require knowing whether academic assessments indicate the same degree of student success, and whether measures of teaching practices relate to those student outcomes.

This study sought to address these questions with respect to one school districts' system, by comparing students' performance on a criterion-based state test to their performance on currently used curriculum-based measures. The degree to which their established measure of teacher practices relates to both types of student outcomes will also be examined. Inherent within an evaluation system that examines professional practices (within or beyond the classroom) in relation to student learning, is an assumption that measured teacher practices affect classroom practice, which in turn affects student learning. Though this study used teacher evaluation scores, the term *classroom practice(s)* is used synonymously with *teacher practices* to reflect this accountability inference. Specifically, this study examined the extent to which classroom practices are predictive of average differences between classrooms in average student performance, and of change in performance over time. This study also examined differences between classrooms with respect to the *Matthew Effect*.

The term "the Matthew Effect", coined in 1968 by sociologist Robert Merton, references a gospel passage and is often used to describe systematic effects reflecting the analogy of "the rich get richer and the poor get poorer". Merton introduced the term in reference to "the accruing of greater increments of recognition for particular scientific contributions to scientists of considerable repute and the withholding of such recognition from scientists who have not yet made their mark", (p. 3). In the field of education, the Mathew Effect is often reflected in achievement gaps or differential average growth trajectories between groups typically viewed as advantaged (e.g., white, male, and or middle class) and those typically viewed as less advantaged (e.g., racial or ethnic minorities, females, or students from low income households). This study takes a

different approach by explicitly modeling classroom variance as the dependent variable, and changes in classroom variance over the school year. Such an approach provides a high level view of the degree to which entire classrooms of students are becoming more similar or dissimilar in passage reading fluency and multiple choice reading comprehension over time. It allows for tests of the extent to which classrooms differ in spread or *dispersion* of student performance at the start of school, and differences in dispersion changes over the school year. Dispersion changes could take on a variety of different patterns. For instance increases in dispersion over time (a fan pattern) indicate a growing gap in student performance. Conversely decreases in dispersion over time (a narrowing pattern) indicate the gap in student performance is shrinking. In other classrooms, the amount of variance might not change significantly over time (a consistent pattern), indicating the extent of differences among students in a class remains relatively unchanged. The following research questions were addressed:

Research Questions

1. How much variance in reading, as measured by the Oregon Assessment of Knowledge and Skills (OAKS) and curriculum-based measures (CBMs), is attributable to differences between classrooms?
2. To what extent is average class performance in reading similar when measured by OAKS and reading CBMs?

3. To what extent does variance between classroom practice scores on four standards domains predict classroom differences in students' reading performance on OAKS and CBMs¹?
4. To what extent do classrooms differ in changes in within-classroom variance in CBM scores over time?
5. If classrooms differ in the change of within-classroom variance over time, to what extent do classroom practice scores on four standards domains predict changes in within-classroom variability on reading fluency and reading comprehension CBMs?

The research questions were addressed within the context of one school district and its process for assessing classroom practice and student learning. Evaluating the practices of general education teachers has served as the primary component of the district's evaluation system. In line with Oregon's teacher evaluation regulations, the partner school district also incorporated student scores on a criterion-referenced state assessment and curriculum-based measures in grades where the state assessment is administered. Administration of both types of assessments has been a regular part of operations within the school district, but not as a component of their teacher evaluation system, until recently. Analyses of student academic data and indicators of classroom practices from prior years could help inform decisions the district continues to face as they implement and refine their new system.

¹ Analyses of overall ratings "Satisfactory" or "Not Satisfactory" were curtailed due to lack of variability.

CHAPTER II

METHOD

District Population

The sample for this study was drawn from the district population of teachers and students between the academic years of 2010-2011 through 2012-2013. As seen in Table 1, approximately 60% of students within the district were eligible for free or reduced price lunch, with a slight increase in 2012. The student population was primarily White, with the percent of minority students remaining between 31% and 33%.

Table 1

District Demographics: Number and Percentage by Year

Demographic Category	2010		2011		2012	
	n	%	n	%	n	%
Free or Reduced Lunch						
Free Lunch Eligible	5,851	54.0	5,838	54.1	6,101	55.4
Reduced Lunch Eligible	799	7.4	633	5.9	862	7.8
Total Eligible	6,650	61.4	6,471	59.9	6,963	63.2
Race/Ethnicity						
White	7,306	67.4	7,236	67.0	7,514	68.2
Black	119	1.1	122	1.1	137	1.2
Hispanic	1,981	18.3	2,080	19.3	2,108	19.1
Asian/Pacific Islander	179	1.7	172	1.6	170	1.5
American Indian/ Alaskan Native	259	2.4	234	2.2	202	1.8
Multi-Ethnic	993	9.2	952	8.8	887	8.1
Unknown	0	0.0	0	0.0	0	0.0
Total Minority	3,531	32.6	3,560	33.0	3,504	31.8
Total	10,837	100.0	10,796	100.0	11,018	100.0

A total of 20 elementary schools and non-profit centers served district students between 2010 and 2013, though six of these school sites were not open for all three academic years. Combined enrollment of fourth and fifth students at the start of school ranged from seven students to 194 students across years and school sites. The number of students enrolled at each site and the total number of student across sites is presented in Table 2, by academic year.

Table 2

Start-of-Year, Fourth and Fifth Grade Student Enrollment in District School Sites

School	Year			School	Year		
	2010	2011	2012		2010	2011	2012
School 1	64	55	--	School 11	14	--	--
School 2	19	--	--	School 12	163	174	178
School 3	128	140	121	School 13	150	151	167
School 4	133	112	98	School 14	170	191	166
School 5	117	113	135	School 15	10	14	12
School 6	24	--	--	School 16	13	16	12
School 7	106	107	108	School 17	181	194	179
School 8	12	10	7	School 18	54	72	71
School 9	110	109	108	School 19	120	138	144
School 10	94	83	--	School 20	--	--	114
Total	1682	1680	1620	School Range	10-181	14-194	7-179

Note. Data obtained from the Oregon Department of Education. Missing enrollment numbers reflect schools sites not in operation.

Participants

The sample included fourth and fifth grade teachers who were evaluated between 2010-2013 and their students from the evaluation year. Evaluations for tenured teachers

were required once every three years, while annual evaluations of new teachers were required during their first three years. A new teacher with multiple years of evaluation data was treated as a unique case for each year to reflect differentiated expectations within the district evaluation tool. Contracted teachers with multiple years of data were treated similarly since the study includes data from a three year span and established teachers are typically not observed more than once every three years unless concern warrants additional observations. Such circumstances would suggest provision of additional support corresponding to the extra evaluation with the intent of changing the teacher's practices.

Students of these teachers were included if they had a fourth or fifth grade state test score, a prior year state test score, and/or at least one fourth or fifth grade CBM score. Students designated Limited English Proficient (LEP) were included based on the same data criteria previously described. A student identified as Limited English Proficient comes from an environment “where a language other than English has had a significant impact on the individual’s level of English language proficiency” and has difficulty understanding or communicating in English to the point of inhibiting proficiency on State assessments, success in classrooms where English is the language of instruction, or the ability to participate fully in society (ODE, 2009a, p.10). For these reasons, some students designated as LEP might not have taken the OAKS or certain CBM assessments at certain time points, and thus would not have that data to include in this study. OAKS data provided by the district did not include scores for students who took the Oregon’s Extended Assessment. This assessment is based on alternate achievement standards and produce test results that are not comparable to results from the

general OAKS assessments (ODE, 2011a). Table 3 describes the district population of students identified as having taken the OAKS Extended Assessment or designated LEP. However, as previously noted, all students from a classroom with evaluation data and with test data for one or more analyses (945 students) were included in this study.

Table 3

Combined Number and Percent of Fourth and Fifth Grade Students Designated as LEP or Took the OAKS Extended Assessment, by Year

OAKS Designation	2010-11		2011-12		2012-13	
	n	%	n	%	n	%
Extended Assessment	35	2.1	35	2.1	32	2.0
Limited English Proficient	144	8.6	141	8.4	106	6.5

Note. n reflects the combined number of 4th and 5th grade students for each category. Percentages are based on combined start-of-school year enrollment for Grades 4 and 5, during 2010, 2011, and 2012, which was 1682, 1680, 1620 students respectively.

Measures

Classroom practice and student outcome data were analyzed *only for the year in which a teacher's practices were documented*, which for most teachers was once during the three-year period. In rare instances when teachers were evaluated more than once (i.e., a new teacher for whom annual evaluations were required for the first three years), the teacher and associated student data were treated as a unique case for each year.

District's measure of teaching practices. The partner school district's evaluation system was based on modified version² of the Danielson Framework for

² Modifications were specific to the partner school district; a common practice according to Milanowski (2011) and encouraged by Danielson (2007).

Teaching³. All teachers were evaluated on 15 standards, though their contract status (first-year, second-year, or third-year/contracted) determined the number of performance targets assessed to operationalize each standard (See Appendix A). For example, to meet Standard 1: Knowledge of Content, a third-year or contracted teacher would need to demonstrate (a) “effective command of the subject to guide student learning”, (b) effective use of “instructional resources, including technology, to communicate content knowledge”, and (c) taking “an active role in adapting new content standards and frameworks to their teaching”. A second-year teacher would be accountable for the first two performance targets, while a first-year teacher would be accountable for only the first performance target. Evidence of a teacher’s practices was obtained through classroom observations, discussion with the teacher, and analysis of artifacts such as lesson plans. First- and second-year teachers were observed at least twice a year and evaluated annually. Contracted teachers were observed and evaluated a minimum of once every three years, with additional observations typically indicative of a cause for concern.

The Summative Evaluation Form (see Appendix B) includes designations of *Exemplary*, *Proficient*, *Basic*, and *Unsatisfactory* for each standard organized within four domains: *Planning and Preparation*, *Classroom Environment*, *Instruction*, and *Professional Responsibilities*. Performance on each standard was coded where *Exemplary* = 3, *Proficient* = 2, *Basic* = 1, and *Unsatisfactory* = 0. This study used domain averages of teachers’ summative evaluation ratings for each of the 15 standards to keep domain scores on the same scale since domains have differing numbers of items. In other words,

³ Beyond Sartain, Stoelinga, and Brown (2009), there is limited, if any, research on the construct, content, or criterion-validity of the Framework. Instead most research focused on relationships between teachers' scores on the framework and student test scores.

scores on standards within each of the four domains were averaged so as not to overly weight domains that included more standards. As a result, the teaching practice scores could range from 0 to 3 for all four domains. Each domain average served as a proxy indicator of classroom practice.

Oregon Assessment of Knowledge and Skills (OAKS): Reading/Literature.

The OAKS Online Reading/Literature assessment was a criterion-referenced and computer adaptive state test. The Reading/Literature assessment was administered in Grades 3-8 and in Grade 10. Fourth grade students were evaluated across five categories: Vocabulary, Read to Perform a Task, Demonstrate General Understanding, Develop an Interpretation, and Examine Content and Structure: Informative Text; all categories but the last were required in third grade (ODE Office of Assessment and Information Services 2011a, 2011b, 2012a, 2012b). Each category represented between 12% and 28% of the 45 multiple-choice questions required of fourth grade students, and the 40 questions provided in third grade. Each item was a question or statement that required completion and students selected from four answer choices. The first item is of average difficulty for the specified grade. The assessment algorithm then selects subsequent items based on the number of items already presented for each content strand, the degree the item maximizes precision in identifying the student's proficiency, and the student's ability based on performance on earlier items (ODE, 2009b). Raw scores were analyzed using the one-parameter logistic item response theory or Rasch model and then converted into a scale or Rasch unit (RIT) score. Students had up to three testing opportunities

between October and May to take the OAKS test⁴. The Rasch scale scores were interpreted using state proficiency cut-points that translated scale scores into four proficiency categories. The two highest proficiency categories for the Reading/Literature achievement standards and the approximate range of scores for Grades 3, 4, and 5 are provided in Table 4 (ODE, 2009c, 2010, 2011b, 2011c, 2011d, 2012a, 2012b, 2013a).

Table 4

OAKS Reading/Literature Achievement Standards

Year	Grade 3			Grade 4			Grade 5		
	Meet	Exceed	Range ^a	Meet	Exceed	Range ^a	Meet	Exceed	Range ^a
2012-13	211	224	192-245	216	226	197-250	221	230	201-246
2011-12	211	224	189-244	216	226	196-246	221	230	201-247
2010-11	204	218	188-240	211	223	196-246	218	230	201-246
2009-10	204	218	188-234	211	223	195-243	218	230	203-248

Note. Scores are on a Rasch Unit scale ($M = 200$, $SD = 10$) and are comparable within the same content area and grade (ODE, 2009a).

^aRange includes highest score in the 1st percentile through the lowest score in the 99th percentile.

Reliability of the OAKS Online Reading/Literature assessment was examined through several different approaches (ODE, 2007). Standard error curves indicated reliable scores across the ability range except in the ends of the distribution, and consistent amounts of error regardless of ethnicity, LEP, or Special Education designations. High achievement classification reliability was reported, ranging from 84 to 99 percent.

⁴ There were only two opportunities for third and fourth grade students in 2012-13 as the state transitioned to the new once a year common assessments.

Oregon Department of Education (2007) reported high correlations ranging from .71 to .82 between students' reading performance on the Oregon state test and their performance on the California Achievement Test, Iowa Test of Basic Skills, the Northwest Evaluation Association Subject Test, and the Lexile Scale. Correlations specific to fourth grade students' State reading scores were .77 with the California Achievement Test and .78 for the NWEA Subject Test. Gain scores defined as the difference between the student's current grade OAKS score and their prior grade OAKS score were calculated and used in all OAKS analyses.

easyCBM™. Data from two easyCBM™ measures were used in this study: Passage Reading Fluency and Multiple Choice Reading Comprehension. Findings from a confirmatory factor analysis, that also included data from a Vocabulary measure not used in this study⁵, suggested that the measures are related, but also contribute individually to reading proficiency (Sáez et al., 2010). Fit indices from the three-factor model were adequate for benchmark data from fall (CFI = .973, TLI = .985, RMSEA = .02) and spring (CFI = .972, TLI = .985, RMSEA = .025). Researchers reported moderately high correlations between the three measures for fall ($r = .71 - .76$) and spring ($r = .71 - .76$) administrations. Correlations across benchmark time points were high for Passage Reading Fluency ($r = .88-.90$) and moderate for Multiple Choice Reading Comprehension ($r = .61-.64$).

Passage Reading Fluency (PRF). The Passage Reading Fluency measures assess students' ability to accurately read a narrative text of approximately 250 words long (Alonzo & Tindal, 2007). Multiple forms are available for progress monitoring and

⁵ Analysis of Vocabulary data was not possible due to substantial missingness (approximately 45% of spring scores) associated with school. See Appendix C.

seasonal benchmarking purposes. Trained assessors administer the measure to students individually. The assessor reads the directions on his or her copy to the student and reviews all proper nouns in the passage (Alonzo & Tindal, 2012). Students are asked to read as much as they can within 60 seconds and timing begins when the student reads the first word. The assessor indicates words read incorrectly on his/her own test protocol. If students hesitate for more than three seconds, the assessor provides the word and counts it as an error. Omitted words are also marked as errors. Self-corrections within three seconds are noted but not counted as errors and inserted words are ignored. The assessor marks the last word read at the end of 60 seconds, then notes the total words read, number of errors, and total correct words (Alonzo & Tindal, 2012).

Analyses of fourth grade Passage Reading Fluency forms showed moderately strong test-retest reliability when administered one week apart. Correlations ranged from .86 to .96 across nine forms (Alonzo, Anderson, Park, & Tindal, 2012). Bivariate correlations between alternate versions of the nine forms revealed moderate to strong relationships from .83 to .98. Predictive and concurrent validity analyses have found fall, winter, and spring PRF to account for 45%, 41%, and 43% of the variance in the Oregon Assessment of Knowledge and Skills (OAKS), respectively (Sáez et al., 2010). PRF scores for students eligible for special education services explained slightly more variance in OAKS scores, 46%, 47%, and 47% respectively, while fall, winter, and spring PRF scores for English language learners accounted for slightly less variance in OAKS – 36%, 40%, and 37%. Sáez and colleagues (2010) also examined the predictive validity of students' PRF growth in relation to their OAKS scores. Students were classified into quartiles based on their fall PRF score. For students in the first quartile, PRF growth was

moderately correlated with OAKS scores. Correlations were weaker for the second- ($r = .28$), third- ($r = .27$), and fourth ($r = .16$) quartiles. Results from a ROC analysis were used to establish cut-points of 96-, 111-, and 126 words read correctly for fall, winter, and spring benchmarks that would identify students at-risk of not meeting OAKS cut-scores (Park, Anderson et al., 2011). The cut-points were developed to maximize sensitivity while keeping specificity above .7 and were found to be relatively stable across two randomly selected groups of fourth grade students (Park, Anderson et al., 2011; Park, Irvin et al., 2011).

Multiple-choice Reading Comprehension (MCRC). The easyCBMTM reading comprehension measure was designed for group administration (Park et al., 2012). Students first read a narrative fiction story of approximately 1500 words then answer 20 multiple-choice questions based on the story (Alonzo, Liu & Tindal, 2007). The multiple-choice questions were scaled using Rasch modeling and assess literal, inferential, and evaluative comprehension (Anderson et al., 2014). For each question a question stem is provided along with three answer choices. Each question is worth one point for a total of 20 possible points. Fall, winter, and spring cut-points for identifying students at risk were set at 10, 13, and 13 points respectively based on results from Receiver Operating Characteristics (ROC) curve analyses (Park, Anderson et al., 2011). Determinations were based on maximizing sensitivity, while keeping specificity above .7. Further analyses using two randomly selected groups from the original analysis suggested cut-points were relatively stable (Park, Anderson et al., 2011).

Park and colleagues (2012) examined the reliability of the multiple-choice reading comprehension measures using a one-way repeated measures ANOVA and found mean

performance across the sample of forms was not significant, indicating reliable degrees of difficulty. Split-half coefficients ranged from .38 to .67 across the six Grade 4 forms. Examinations of the differential functioning of items for students at or below the 23rd percentile to students at or above the 78th percentile resulted in identification of only two items across all forms that exhibited higher proportions of correct responses for low-performing students, though the difference between groups performance was .03 or less. Collectively, these results indicated that difficult items function appropriately. Tindal, Nese, & Alonzo (2009) reported moderate correlations of .66, .61 and .65, and .61 and .59 for fall, winter and spring scores and the OAKS across two districts. Fall, winter, and spring MCRC scores accounted for approximately 60%, 53%, and 60% of variance in OAKS respectively. Sáes and colleagues (2010) reported correlations from .48 to .53 between students fall to winter slopes and performance on OAKS across quartiles of students classified by their fall MCRC score.

Data Screening

Prior to analysis, data were examined for the extent of missingness, the degree to which missingness was random, and the extent to which the analytic sample differed from the population of Grade 4 and Grade 5 students. Data were also examined for univariate and multivariate normality. The data screening process is described below. Findings from data screening procedures and analyses are reported in the results section, and validity ramifications due to non-normal data, missingness, and how missingness was addressed are described in as limitations within the disussion section.

Missingness. Teachers with no evaluation data and their associated students were excluded from the sample. Students with only one OAKS score were excluded from

OAKS gain score analyses ($n = 66$), reducing that analytic sample from 945 students to 879 and representing a loss of approximately seven percent of the potential sample. Students missing all scores for PRF ($n = 26$) or MCRC ($n = 37$) CBMs were excluded from respective analyses. However, all students with at least one data point were included in easyCBM analyses because of the strength of HLM growth models for missing repeated measures data when using full maximum likelihood (Raudenbush & Bryk, 2002), as was done in the current study. The final analytic samples consisted of 919 students for PRF and 908 students for MCRC analyses.

Omitted students were compared to included students on all dependent variables to determine whether they were significantly different or if they were missing at random. Patterns of missingness by time point were examined. In instances where more than 2% of students represented a specific pattern (e.g., missing all but fall scores), those students were compared to peers without missing data on other variables to determine the extent to which they differed significantly based on extant data.

Frequency counts and percentages were requested for all other variables. For variables with missing data greater than 2%, t -tests were conducted to examine differences in each outcome variable based on missingness. Results from all analyses were used to determine whether students were missing completely at random, missing at random, or not missing at random. These analyses are reported in the Results section. However, substantial missing data, how those data are treated especially the use of listwise deletion for OAKS data, potential issues with internal validity of results, analytic power, and inferences based on results, as well as limits to generalizability are noted in the Discussion section.

Testing assumptions. OAKS gain scores and fall, winter, and spring CBM scores for Passage Reading Fluency and Multiple Choice Reading Comprehension were examined for univariate normality, skew, and outliers using descriptive statistics, box-plots, and histograms. Distributions of average classroom domain scores were examined. Scatterplots of average domain and student scores were examined for multivariate linearity and residuals were examined for homoscedasticity. After specifying growth models in HLM, tests of homogeneity of Level 1 variance were conducted (Raudenbush & Bryk, 2002). Residuals for all levels of the HLM models were exported and examined in SPSS for consistency with assumptions underlying multilevel models⁶. Additionally, a scatterplot of the Mahalanobis distances against the expected chi-square distribution values from the higher level residual files was requested along with a fit line to ensure consistency with assumptions underlying multilevel models.

Analyses

Descriptive statistics including means and standard deviations for each CBM measure in fall, winter, and spring were calculated for the full sample and for each classroom to describe classroom level changes over time. Gain scores were computed for the OAKS and means and standard deviations were examined overall and by class. Means and standard deviations were calculated for indicators of overall teaching designations and for teaching practice scores for each of the four standards domains. Classroom practice scores for each domain were grand mean centered across classrooms to facilitate meaningful interpretation of the intercept.

⁶ Findings with regard to examining assumptions are described in the Results section.

Research questions 1 through 3. Hierarchical linear modeling and correlation analyses were used to address the first three research questions. OAKS gains were examined using a two-level hierarchical linear model with student at Level 1 and their OAKS Reading/Literacy gain score as the outcome. The intraclass correlation coefficient (ICC) for this null model was calculated. The original plan included the addition of the overall practice designation at Level 2, after which the ICC would be calculated again, a pseudo- R^2 would be calculated, and classroom Empirical Bayes (EB) estimates would be exported for later correlational analyses. As discussed in the results section, the overall practice designation was not used due to limited variability. However the same procedures were used, adding classroom practice scores for each of the four standards domains at Level 2; see equations 1.1 and 1.2.

$$Y_{ij} = \beta_{0j} + r_{ij} \quad (1.1)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Tprac1})_j + \dots + \gamma_{04}(\text{Tprac4})_j + u_{0j} \quad (1.2)$$

where

Y_{ij} is the OAKS Reading/Literature gain score for student i in classroom j ;

r_{ij} is the residual for student i in classroom j ;

β_{0j} is the mean OAKS Reading/Literature gain score for classroom j ;

$(\text{Tprac1})_j$ through $(\text{Tprac4})_j$ in equation 1.2 represent teaching practice scores

for each of the four standards domains in classroom j ;

γ_{00} is the mean OAKS Reading/Literature gain score across classrooms;

γ_{01} through γ_{04} in equation 1.2 are the average relationship between teaching

practice scores on the four standards domains and OAKS gain scores

across classrooms;

u_{0j} is the residual for classroom j .

To examine CBM passage reading fluency and reading comprehension outcomes, a series of three-level models included the same student and teacher variables with time at Level 1, student at Level 2, and classroom at Level 3. In the final PRF model, the time variable at Level 1 (PRFseason) represented fall, winter, and spring administrations and was coded as -2.34, -1, and 0 respectively. The same classroom-level variables were entered at Level 3 in these models to yield the complete systems of equations 2.1 to 2.5. As with the OAKS models, the classroom EB residuals were exported for use in subsequent analyses. The intraclass correlation coefficient was calculated for the unconditional model and a Level 3 pseudo- R^2 was calculated after the addition of the teaching practice variables.

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}(\text{PRFseason})_{tij} + e_{tij} \quad (2.1)$$

$$\pi_{0ij} = \beta_{00j} + r_{0ij} \quad (2.2)$$

$$\pi_{1ij} = \beta_{10j} + r_{1ij} \quad (2.3)$$

$$\beta_{00j} = \gamma_{000} + \gamma_{001}(\text{Tprac1})_j + \dots + \gamma_{004}(\text{Tprac4})_j + u_{00j} \quad (2.4)$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101}(\text{Tprac1})_j + \dots + \gamma_{104}(\text{Tprac4})_j + u_{10j} \quad (2.5)$$

where

Y_{tij} is the PRF score at time t for student i in class j ;

$(\text{PRFseason})_{tij}$ is a time indicator that is -2 at fall, -1 at winter, and 0 at spring;

e_{tij} is the residual at time point t for student i in classroom j ;

π_{0ij} is the spring status for student i in classroom j , the expected spring PRF score;

π_{1ij} is the change in PRF performance for student i in classroom j across the three seasons;

r_{0ij} is the spring PRF residual for student i in classroom j ;
 r_{1ij} is the growth residual for student i in classroom j ;
 β_{00j} is the mean spring PRF score across students in classroom j ;
 β_{10j} is the mean PRF growth rate across students in classroom j ;
 $(Tprac1)_j$ through $(Tprac4)_j$ in equations 2.4 and 2.5 represent teaching practice scores for each of the four standards domains in classroom j ;
 γ_{000} is the grand mean PRF score for spring across all classrooms;
 γ_{001} through γ_{004} in equation 2.4 are the average relationships between teaching practice scores for each of the four standards domains and the grand mean spring PRF score across all classrooms;
 γ_{100} is the overall average PRF growth rate across all classrooms;
 γ_{101} through γ_{104} in equation 2.5 are the average relationships between teaching practice scores for each of the four standards domains and the overall average PRF growth across all classrooms;
 u_{00j} is the spring PRF residual for classroom j ;
 u_{10j} is the PRF growth residual for classroom j .

A similar system of questions was used for MCRC CBM scores, where time $(PRFseason)_{ij}$ was replaced with $(MCRCseason)_{ij}$ coded -2, -1, 0, and a fixed quadratic term $(MCRCquad)_{ij}$ was added, coded as 4, 1, 0 to more accurately represent MCRC growth.

Various components from the above models were used to address the first three research questions. Exported classroom-level residuals from the unconditional OAKs

and CBM models were used to address Research Question 1, regarding the degree to which OAKS and CBM scores provide similar indications of average class performance in reading. Specifically, the OAKS and CBM Level-2 intercept residuals were correlated with the passage reading fluency and reading comprehension CBM Level-3 intercept and slope residuals. Research Question 2 was addressed by comparing the ICCs calculated for the OAKS and CBM unconditional models, with a specific focus on the variance attributable to the classroom level. To address Research Question 3, pseudo- R^2 for OAKS and CBM were computed and compared for the models conditioned on teaching practice scores on each standards domain.

Research questions 4 and 5. To address the remaining research questions, classroom variance for CBM passage reading fluency and reading comprehension measures for each time point were used as the dependent variable in a new series of growth models. Dispersion of student scores within classrooms and the extent to which dispersion changed over time was examined using a two-level model where Level 1 was the variance in CBM scores at each time point and Level 2 was the classroom. Unlike previous models, CBM variance in fall served as the intercept since the research questions are in respect to slopes only (i.e., time will be coded 0, 1, 2). The ICC was calculated for the unconditional model to address Research Question 4 and determine the degree to which dispersion patterns varied between classrooms. Classroom practices were added at Level 2 as a predictor of the slope estimate and a pseudo- R^2 was calculated to answer Research Question 5. The final model is shown in the system of equation 3.1 through 3.3.

$$Y_{tj} = \beta_{0j} + \beta_{1j}(\text{CBMseason})_{tj} + r_{tj} \quad (3.1)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (3.2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{Tprac1})_j + \dots + \gamma_{14}(\text{Tprac4})_j + u_{1j} \quad (3.3)$$

where

Y_{tj} is the amount of dispersion in CBM scores at time t for

classroom j ;

$(\text{CBMseason})_{tj}$ is a time indicator that is 0 at fall, 1 at winter, and 2 at spring;

r_{tj} is the residual variance at time t in classroom j ;

β_{0j} is the CBM dispersion for classroom j in fall;

β_{1j} is the change in CBM dispersion for classroom j across time;

$(\text{Tprac1})_j$ through $(\text{Tprac4})_j$ in equation 3.3 represents classroom practice scores

for each of the four standards domains in classroom j ;

γ_{00} is the average CBM dispersion in fall across all classrooms;

γ_{10} is the average change in dispersion over time across all classrooms;

γ_{11} through γ_{14} in equation 3.3 are the average relationships between classroom

practice scores for the four standards domains and average change in

dispersion across classrooms;

u_{0j} is the fall dispersion residual for classroom j ;

u_{1j} is the change in dispersion residual for classroom j .

CHAPTER III

RESULTS

In this section, the final analytic samples are described along with data screening, patterns of missingness, and model specification decisions. Subsequently, results from the final models are described, organized by their respective research question.

Analytic Sample

Student reading data and indicators of teaching practice were obtained from a medium sized school district in the Northwest. The population of interest for this study included fourth and fifth grade teachers who were evaluated from 2010-2013 and their students from the evaluation year. Out of the 41 classrooms evaluated between 2010 and 2013, six⁷ were not linked to students and were excluded from the sample. A total of 945 students were included in this study. Student mobility, defined by the district as an address change but not necessarily a school or district change, was fairly high across the Grade 4 and Grade 5 students, with approximately 20% having six or more residence changes. Over 75% of students were eligible for free or reduced price lunch. Approximately 65% of students were White and over 20% were of Hispanic ethnicity. A summary of demographic characteristics of the 4th and 5th grade student population between 2010-2013, students assigned to teachers within the data file provided by the school district, the sample of participating students is provided in Table 5. Differences between the population and the analytic sample are also reported.

⁷ These cases might represent specialists who serve students from multiple classrooms (e.g., music or physical education teachers). Students' reading test scores might be associated with their general education teacher only.

Table 5

Proportions of 4th and 5th Grade Students Representing Select Demographic Categories in District Population and Respective Samples

Student Characteristics	Population (<i>N</i> = 5361)		Assigned to Teachers (<i>N</i> = 4646)		Assigned to Evaluated Teachers (<i>N</i> = 945)		χ^2
	Prop.	Miss.	Prop.	Miss.	Prop.	Miss.	
Female	.473	.003	.474	< .001	.490	.001	0.931
FRL Eligible	.775	.000	.783	.000	.863	.000	37.260***
SPED	.252	.000	.242	.000	.247	.000	0.107
LEP	.078	.003	.084	< .001	.107	.001	8.940**
Race/Ethnicity							
White	.658	.000	.655	.000	.631	.000	2.588
American Indian	.021	.000	.020	.000	.026	.000	0.944
Asian	.011	.000	.011	.000	.016	.000	1.729
Black	.011	.000	.010	.000	.008	.000	0.693
Hispanic Ethnicity	.209	.000	.216	.000	.241	.000	4.894*
Multiple Pacific Islander	.083	.000	.079	.000	.071	.000	1.550
# of Moves ^a							
0	.199	.000	.200	.000	.169	.000	4.614*
1	.171	.000	.172	.000	.151	.000	2.299
2	.131	.000	.133	.000	.133	.000	0.028
3	.105	.000	.108	.000	.114	.000	0.685
4	.088	.000	.088	.000	.097	.000	0.800
5	.074	.000	.072	.000	.080	.000	0.417
6+	.231	.000	.227	.000	.254	.000	0.124

Note. "Prop." is the proportion of students representing each characteristic. "Miss." is the proportion of students missing data for that demographic characteristic. For example, for gender and LEP status, .3%, less than .1%, and .1% of students were missing data in relation to the population of Grade 4 and Grade 5 students, the sample of students assigned to a teacher, and the analytic sample (students assigned to an evaluated teacher).

^aMoves are residence changes, not necessarily a change in school.

p* < .05. *p* < .01. ****p* < .001

The analytic sample included significantly more students receiving free or reduced price meals, with limited English proficiency, and students of Hispanic ethnicity, as well as fewer students who had not changed residences. These differences have implications for the generalizability and validity of the study findings detailed in the Discussion section. As shown in Table 6, classroom composition varied widely across the 35 classrooms.

Table 6
Summary of Classroom Composition: Class Size and Proportion of Students by Demographic Indicator, N=945

Classroom Descriptive	Range	Minimum	Maximum	Mean
Class size	18	15	33	27
Student Demographics				
Female	0.29	0.38	0.68	0.49
FRL	0.58	0.42	1.00	0.87
SPED	0.43	0.05	0.48	0.25
LEP	0.26	0.00	0.26	0.11
No Residence Changes	0.33	0.00	0.33	0.17
6+ Residence Changes	0.53	0.09	0.62	0.26
White	0.47	0.41	0.88	0.63
Hispanic	0.43	0.00	0.43	0.24

Note. Student demographics presented as proportions can be translated into percentages. For example, a minimum proportion of 0.38 female students and a maximum proportion of 0.68 female students indicate that between 38 to 68 percent of students were female across all classrooms.

Data Screening

Out of the 35 evaluated teachers, all but one received an overall designation of satisfactory. For this reason, analyses of the overall classroom designations were dropped from this study. Data from the district's measure of classroom practice reflected patterns in teacher evaluation scores similar to those found in other studies that used a Danielson based classroom observation tool (e.g., Ho & Kane, 2013). For each of the 15 standards most teachers were rated as *Proficient*, a few as *Basic* or *Exemplary*, with only one occasion of *Unsatisfactory* performance for three standards. Domain averages were calculated for *Planning and Preparation* ($M = 2.09, SD = 0.36$), *Classroom Environment* ($M = 2.20, SD = .60$), *Instruction* ($M = 2.16, SD = .43$), and *Professional Responsibilities* ($M = 2.14, SD = .45$) to prevent overly weighting domains that included more standards. Similar to the distribution of domain standard scores, all domain averages were skewed (see Appendix D), where 88% or more of classroom averages were between a two (i.e., *Proficient*) and a three (i.e., *Exemplary*). Cronbach's alpha for the five *Planning and Preparation* items, three *Classroom Environment* items, three *Instruction* items, and four *Professional Practice* items were .85, .94, .76, and .84 respectively, indicating moderate to high internal consistency of domain items.

Data for OAKS, PRF, and MCRC were examined to understand the nature of the distribution of scores at each assessment occasion. Means, standard deviations, and the percent of students missing data for OAKS gains and each easyCBM measure in fall, winter, and spring were calculated for the full analytic sample to describe classroom level changes over time. To ensure the confidentiality of classrooms, statistics are summarized through minimums and maximums. As shown in Table 7, classrooms varied widely in

terms of average scores, the spread of scores (reflected by the standard deviation), and the percent of missing data.

Table 7.

Range of Classroom Means, Standard Deviations, and Missingness for Reading Scores/Gains

Assessment	Mean		SD		% Missing	
	Min	Max	Min	Max	Min	Max
OAKS gains	0.38	12.19	3.29	10.09	0	22.20
PRF						
Fall	83.83	152.54	26.44	55.68	0	14.00
Winter	100.41	172.36	26.72	58.38	0	14.00
Spring	104.96	176.53	29.24	63.10	0	100.00
MCRC						
Fall	9.82	15.20	2.03	5.15	0	19.20
Winter	11.08	17.70	2.04	4.71	0	19.00
Spring	9.50	16.93	1.97	5.25	0	100.00

Inspection of histograms and Q-Q plots of OAKS and PRF data suggested relatively normal distributions at each time point (See Appendix E). However, Shapiro-Wilk tests of normality were significant at each assessment occasion, except for PRF time points 2 and 3 (see Table 8), indicating those data were not normally distributed. For MCRC data, negative skew became more apparent as kurtosis increased over time, across

fall (kurtosis = -0.44, SE = 0.16), winter (kurtosis = 0.51, SE = 0.16), and spring (kurtosis = 0.73, SE = 0.18). These patterns might reflect some degree of ceiling effect due to the nature of assessment questions or differences in the sample of students with spring data, undermining assumptions that MCRC scores within the sample were normally distributed.

Table 8.

Descriptive statistics of reading outcome data by assessment occasion.

Assessment	N	Mean	Std. Deviation	Variance	Shapiro-Wilk
OAKS 1	881	217.20	11.08	122.84	0.991 ^{***}
OAKS 2	927	222.18	10.29	105.87	0.995 ^{**}
PRF fall	903	119.61	44.41	1972.33	0.995 ^{**}
PRF winter	916	136.57	43.01	1851.41	0.998
PRF spring	857	148.19	46.25	2138.64	0.999
MCRC fall	893	12.40	4.02	16.12	0.961 ^{***}
MCRC winter	906	14.33	3.80	14.48	0.929 ^{***}
MCRC spring	720	13.91	3.66	13.41	0.931 ^{***}

*** $p < 0.001$. ** $p < 0.01$.

Missingness. Proportions of missing data varied by reading measure and time point. Though fewer than seven out of every 100 students were missing data for most

assessment-by-time point combinations, much greater proportions of data were missing for the spring administration of the MCRC assessment (see Table 9).

Table 9

Proportion of Participant Students Missing Assessment Data, $N = 945$

Assessment	Time 1	Time 2	Gain Score/Time 3
OAKS	.068	.019	.070
PRF	.044	.031	.093
MCRC	.055	.041	.238

Note *Missing Gain Scores are applicable to OAKS, indicating the proportion of students who did not have scores at Time 1 and/or Time 2.

Patterns of missing data were examined using Little's MCAR test in SPSS version 22.0 (IBM Corporation, 2013) and were not significant across OAKS or PRF time points. However, MCRC scores across fall, winter, and spring were not missing completely at random, $\chi^2 = 41.43$ ($df = 9, p < .01$). Results from follow up t -tests indicated that students who were missing data at any one MCRC time point had significantly different scores than those who were not, across all time point combinations (see Table 10).

Table 10

Differences in MCRC Performance Between Students With and Without Data For Other MCRC Test Occasions

Occasion	With Score			Missing Score			<i>df</i>	<i>t</i>	^a <i>d</i>
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>			
MCRC1 miss									
MCRC 2	880	14.39	3.80	26	12.46	3.41	904	2.55*	0.53
MCRC 3	687	14.01	3.61	33	11.73	4.17	718	3.53***	0.58
MCRC2 miss									
MCRC 1	880	12.43	4.01	13	10.31	4.03	891	1.89	0.53
MCRC 3	696	13.96	3.65	24	12.46	3.75	718	1.98*	0.41
MCRC3 miss									
MCRC 1	687	12.13	4.09	206	13.29	3.63	374	3.91***	-0.30
MCRC 2	696	14.06	3.85	210	15.21	3.52	904	3.87***	-0.31

* $p < .05$, *** $p < .001$ ^a*d* = Cohen's *d*

To further investigate these significant patterns of missing data, chi-square tests were conducted to examine differences in missingness at each assessment time point by school and by demographic characteristics. Significant differences were found between the expected number of students missing data and the number observed, by school, for MCRC time point 2 and time point 3 (see Table 11). In both cases, a couple of schools were missing a large percentage of data. However, since the sampling strategy for this study was driven by the availability of classroom practice data, schools were represented by only one to six classrooms. For example, 100% of students at one school, represented by two classrooms, were missing MCRC time 3 data. Another school, represented by only one classroom, had 96.6% of students missing MCRC time 3 data. Collectively

these three classrooms accounted for over 40% of missing data at that time point. This pattern of a statistically significant amount of missing data by school also applied to OAKS time 1, and 2, as well as PRF time 3. With respect to PRF time 3, the 88 students missing data were spread across three schools, which collectively represented 11 classrooms. Despite the clear patterns of missingness by school, it is not clear whether that missingness represents a school-wide administrative policy or teacher decisions, given the limited number of classrooms representing the schools with the most missing data.

Table 11

Relationship Between School and Missing Data, by Assessment Time Point

Assessment	Not Missing		Missing		χ^2	<i>p</i>
	<i>n</i>	%	<i>n</i>	%		
OAKS 1	881	93	64	7	32.61	0.001
OAKS 2	927	98	18	2	24.52	0.017
PRF 1	903	96	42	4	16.66	0.163
PRF 2	916	97	29	3	24.08	0.020
PRF 3	857	91	88	9	371.71	< 0.001
MCRC 1	893	95	52	5	20.88	0.052
MCRC 2	906	96	39	4	28.68	0.004
MCRC 3	720	76	225	24	418.13	< 0.001

Significant differences in the proportion of missing data by assessment time point were also found in relation to certain demographic characteristics. Greater proportions of missing data for students eligible for free or reduced priced meals were found for PRF time point 3, and MCRC time points 2 and 3.

Testing assumptions. Data for MCRC and PRF were visually examined for linearity. Box plots and descriptive statistics including means and medians indicated that the average rate of growth for both PRF and MCRC was not consistent from fall to spring. Average MCRC scores for winter appeared slightly higher than spring scores. Average PRF gains from fall to winter were greater than gains from winter to spring. Shapiro-Wilk tests of normality were significant for OAKS gains ($SW = 0.988, df = 879, p < 0.001$); fall PRF ($SW = 0.995, df = 903, p = 0.007$); and MCRC for fall ($SW = 0.961, df = 893, p < 0.001$), winter ($SW = 0.929, df = 906, p < 0.001$), and spring ($SW = 0.931, df = 720, p < 0.001$) time points. Inspection of histograms and Q-Q plots suggested a somewhat leptokurtic distribution of OAKS gains, and negative skew in the distribution of MCRC scores particularly for winter and spring, however fall PRF appeared relatively normal. Box plots indicated no significant outliers across time points and assessments (see Appendix E).

OAKS gains did not appear to be linearly related with classroom practice scores for any of the four domains. The majority of average classroom practice scores were between 2 (Proficient) and 3 (Exemplary); yet, across all domains, no more than two classrooms had average OAKS gains greater than that of classrooms rated Proficient (Appendix F). Scatterplots of classroom practice scores on the four standards domains and student scores seemed to indicate a nonlinear relationship between MCRC scores and classroom domain scores over time. Similar scatterplots for PRF scores revealed patterns more akin to linear changes over time for most domains.

After specifying growth models in HLM, a test of homogeneity of Level 1 variance (Raudenbush & Bryk, 2002) indicated normally distributed errors across MCRC

time points $\chi^2 (675, N = 932) = 359.102, p > .500$. However, results for the two-level PRF model indicated heterogeneous level-1 variance $\chi^2 (817, N = 936) = 1338.068, p < .001$, suggesting that PRF variance differed between fall, winter, and spring. PRF analyses proceeded as planned since the aim of research questions four and five was to explicitly examine dispersion patterns.

Residuals for all levels of the HLM models were examined in SPSS for consistency with assumptions underlying multilevel models. Residuals for OAKS gains and MCRC scores were not distributed uniformly, instead reflecting a systematic pattern between residuals and predicted reading scores, indicating that important additional predictors may be missing from the analyses. PRF residuals appeared to indicate linearity with less indication of bias and appeared largely homoscedastic (See Appendix G). Additionally, a scatterplot of the Mahalanobis distances against the expected chi-square distribution values (Appendix H) indicated a linear relationship for MCRC ($R^2 = 0.941$), PRF ($R^2 = 0.904$), and OAKS gains ($R^2 = 0.967$). Collectively, results provide some evidence that OAKS gains and MCRC scores were less consistent with assumptions underlying multilevel models compared to PRF scores, posing potential limitations to the validity of OAKS and MCRC results.

Model Specification

The observed trends in PRF and MCRC data over time warranted examination of non-linear models. Appropriate functional form was examined using Mplus version 6.1 (Muthen & Muthen, 2010). Growth models were specified where fall was freely estimated, winter = -1, and spring = 0, resulting in a fall estimate of -2.34 for PRF. A MCRC model where fall = -2, winter was freely estimated, and spring = 0, resulted in an

estimated winter weight of 0.26. These results indicated that changes in MCRC and PRF scores over the year were not necessarily linear and consideration of a quadratic term was warranted.

Based on findings from the Mplus analyses, the author compared five quadratic, two-level HLM null models in which fractional weights were calculated by taking the linear weights (-2, -1, 0) to a power of 1.2, 1.4, 1.6, 1.8, or 2 (Royston, 1994). A linear PRF model with time coded as -2.34 for fall, -1 for winter, and 0 for spring was also examined and compared to the quadratic and fractional models. Of models with quadratic terms, the pure quadratic (weights of 4, 1, 0) accounted for the greatest amount of level one variance for both PRF (14.9%) and MCRC (23.3%). All deviance tests between quadratic models and those using fractional weights were not statistically significant ($p > 0.50$). However, the PRF linear model with fall coded as -2.34 was a significant improvement in fit over the pure quadratic model ($df = 5, p < 0.001$).

Results from four-level models, indicated that school membership was not significantly related to MCRC scores ($ICC = 0.001, p > 0.50$), PRF scores ($ICC = 0.010, p = 0.21$), or OAKS gains ($ICC = < 0.001, p > 0.50$). However, the proportion of classroom variance in MCRC ($ICC = 0.054, p < 0.001$) and PRF ($ICC = 0.102, p < 0.001$) scores and OAKS gains ($ICC = 0.105, p < 0.001$) was significant and meaningful, informing the final two level model for OAKS and three level models for MCRC and PRF.

Research Question One

Research Question One asked how much variance in reading, as measured by the Oregon Assessment of Knowledge and Skills (OAKS) and curriculum-based measures

(CBMs), was attributable to differences between classrooms. Table 12 reports the amount of variance in OAKS gain scores attributed to differences between students versus differences between classrooms, while Table 13 decomposes variance in CBM scores unique to time, students, and classrooms. Approximately 12.4 percent of variance in OAKS gains, 11.1 percent of variance in PRF scores, and 5.2 percent of variance in MCRC scores was between classrooms. These results suggest that most of the variance in reading scores was due to differences between students, or factors specific to the time of test administration. Nonetheless, some variance was due to differences between classrooms.

Table 12
Variance in OAKS Attributed to Differences Between Classrooms

Model Level	Variance
Level 1, Students	40.610
Level 2, Classes	5.032
Classroom ICC	0.124

Table 13
Variance in PRF and MCRC attributed to differences between classrooms

Model Level	PRF	MCRC
Level 1, Time	167.503	5.311
Level 2, Students	1682.141	6.924
Level 3, Classes	171.065	0.674
Classroom ICC	0.111	0.052

Research Question Two

Research Question Two investigated the extent to which average class performance in reading was similar when measured by OAKS and reading CBMs. To address this research question, intercepts and slopes from PRF and MCRC were correlated with OAKS gains. Spring intercepts for PRF and MCRC were included since OAKS is thought of as an end-of-year indicator. Nonetheless, OAKS as represented here captures change from one year to the next, thus making the relationship between growth and gains more interesting substantively. Classroom spring score estimates for PRF and MCRC were not related to classroom gain score estimates on OAKS. Though no statistically significant relationship was found between average classroom OAKS gains scores and MCRC growth, average PRF growth was significantly related ($r = 0.66$, $p < 0.001$) to OAKS gains. PRF spring scores were also related to MCRC gains as displayed in Table 14.

Table 14
Correlations Between Estimates of Average Classroom OAKS Gains, CBM Spring Scores, and CBM Growth (N = 35)

Measure	1	2	3	4	5
1. OAKS gain	--				
2. PRF spring	-.247	--			
3. PRF growth	.658**	-.271	--		
4. MCRC spring	-.196	.269	-.283	--	
5. MCRC growth	.242	-.759**	.179	.160	--

Note. Empirical Bayes estimates were used.

** $p < .001$, all else $p > 0.1$

Higher classroom averages for PRF and MCRC spring scores were not associated with greater OAKS gains, nor were average classroom growth in MCRC. However, a moderate, positive correlation between PRF growth and OAKS gains indicated that the greater the average class increase in students' passage reading fluency, the higher the average gain in OAKS. The magnitude of the correlation indicates that while CBM passage reading fluency measures similar aspects of reading as OAKS, each measure still provides some unique information on students' reading abilities. Moreover, a strong negative correlation between MCRC growth and PRF spring scores suggest that the greater the average classroom PRF score for spring, the lower the average rate of growth in MCRC. Average class performance in reading differed depending on the measure used, as well as whether the interest is in end of year performance or change in reading performance over the school year.

Research Question Three

Research Question Three examined the extent to which variance between classrooms on scores on four standards domains predicted classroom differences in students' reading performance on OAKS and CBMs. Only 0.04 percent of variance in OAKS gains and 0.02 percent of variance in PRF scores was accounted for by the addition of classroom practice scores in each of four domains (Planning and Preparation, Classroom Environment, Instruction, and Professional Responsibilities). Variance in MCRC scores increased after the addition of classroom practice scores, suggesting the resulting model did not follow the trend of the data (See Table 15). Results for these final models are provided in Tables 16 through 18.

Table 15
Proportion of Variance Explained After the Addition of Classroom Practice Scores

Variance Components	OAKS	PRF	MCRC
Unconditional s^2	40.610	167.503	5.311
Conditional s^2	40.595	167.469	5.312
Pseudo R^2	0.0004	0.0002	-0.0002

The average OAKS gain score across all classrooms was approximately 5 points. Classroom practices regarding Planning and Preparation, Classroom Environment, and Professional Responsibilities were not associated with students' gain scores. However, Instruction did predict classroom differences in OAKS performance. For a positive standard deviation difference in evaluation scores for Instruction ($SD = .43$), average OAKS gain scores were approximately 1.6 points higher than average ($t = 2.25, p = 0.03$). In terms of classroom practices, average OAKS gains would be 3.8 points higher in classrooms rated *Exemplary* in Instruction compared to classrooms rated *Proficient*.

Table 16
OAKS Gains and the Contribution of Classroom Practices

Fixed Effect	Coefficient	<i>SE</i>	<i>t</i>	<i>p</i> -value
Average gain, γ_{00}	5.102	0.400	12.744	<0.001
Planning & Preparation, γ_{01}	-2.427	1.637	-1.482	0.149
Classroom Environment, γ_{02}	-1.134	1.027	-1.105	0.278
Instruction, γ_{03}	3.792	1.696	2.236	0.033
Professional Responsibilities, γ_{04}	0.206	1.510	0.137	0.892

Note. Student $N = 879$.

The average spring MCRC score across all classrooms was approximately 14 points ($t = 68.93, p < 0.001$). Since time was reverse coded to facilitate interpretation of outcomes, the meaning of the coefficients is not transparent. The linear and quadratic slopes reported in Table 17 resulted in predicted growth in MCRC of 1.958 points between fall and winter, and a loss of 0.294 points from winter to spring. Differences in classroom practices were not associated with students' spring MCRC scores or MCRC growth (See Table 17).

Table 17
MCRC Outcomes, Growth, and the Contribution of Classroom Practices

Fixed Effect	Coefficient	SE	<i>t</i>	<i>p</i> -value
Spring Outcome, γ_{000}	13.990	0.203	68.928	<0.001
Planning & Preparation, γ_{001}	-0.896	0.938	-0.956	0.347
Classroom Environment, γ_{002}	0.256	0.423	0.605	0.550
Instruction, γ_{003}	-1.003	0.917	-1.094	0.283
Professional Responsibilities, γ_{004}	1.631	0.943	1.731	0.094
Linear Growth, γ_{100}	-1.420	0.451	-3.152	0.004
Planning & Preparation, γ_{101}	0.475	0.582	0.816	0.421
Classroom Environment, γ_{102}	-0.181	0.323	-0.560	0.580
Instruction, γ_{103}	-0.229	0.654	-0.351	0.728
Professional Responsibilities γ_{104}	0.734	0.601	1.220	0.232
Quadratic Growth, γ_{200}	-1.126	0.173	-6.498	<0.001

Note. Student $N = 908$. Time was reverse coded with spring as the intercept to facilitate interpretation of student outcomes.

As displayed in Table 18, the average spring PRF score was approximately 149 words read per minute across all classrooms ($t = 56.88, p < 0.001$). However, once again, the slopes are not easily interpreted here due to the way time was coded. Student PRF increased by an average of 17.17 words read per minute from fall to winter and of 12.82 words read per minute from winter to spring ($t = 21.75, p < 0.001$). Differences in classroom practices were not related to spring PRF scores, but differences in classroom instruction and professional responsibilities were related to growth in passage reading fluency. A positive standard deviation difference in classroom ratings for instruction ($SD = .43$) was associated with faster PRF growth by 2.8 words read per minute ($t = 3.851, p < 0.001$). This translates to a predicted average growth rate of 11.7 words read per minute for a class rated *Proficient* in Instruction compared to 18.4 words read per minute in classrooms rated *Exemplary*.

However, for a positive standard deviation difference in teachers' professional responsibilities ($SD = .45$), average classroom PRF growth was slower by 1.51 words per minute, per season ($t = -3.304, p = 0.002$). The predicted growth rate for classrooms rated *Proficient* for professional responsibilities was 16 words read correctly per minute, approximately 6 words more per minute than for classrooms rated as *Exemplary*. This suggests higher than average ratings for classroom instruction were associated with faster student PRF growth, while higher than average classroom ratings for professional responsibilities (i.e., professional growth, professionalism, communication, and commitment to instruction initiatives) were associated with reduced student growth.

Table 18
PRF Outcomes, Growth, and the Contribution of Classroom Practices

Fixed Effect	Coefficient	SE	<i>t</i>	<i>p</i> -value
Spring PRF scores γ_{000}	149.250	2.624	56.883	<0.001
Planning & Preparation, γ_{001}	-23.355	11.766	-1.985	0.056
Classroom Environment, γ_{002}	8.724	6.968	1.252	0.220
Instruction, γ_{003}	2.984	10.951	0.272	0.787
Professional Responsibilities, γ_{004}	-3.505	9.615	-0.365	0.718
Linear Growth, γ_{100}	12.816	0.589	21.747	<0.001
Planning & Preparation, γ_{101}	-1.908	1.822	-1.047	0.303
Classroom Environment, γ_{102}	-1.673	1.295	-1.292	0.206
Instruction, γ_{103}	6.596	1.713	3.851	<0.001
Professional Responsibilities γ_{104}	-3.346	1.013	-3.304	0.002

Note. Student $N = 919$ Time was reverse coded with spring as the intercept to facilitate interpretation of student outcomes.

Research Question Four

Research Question Four examined the extent to which classrooms differed in dispersion of CBM scores over the school year. Approximately 78% of variance in dispersion patterns of PRF scores and 60% of variance in dispersion patterns of MCRC scores was attributable to differences between classrooms. This indicates that dispersion patterns differ markedly across classrooms.

Table 19

Variance in Dispersion Patterns of Reading Scores Over Time, Between Classrooms

Variance	PRF	MCRC
Level 1, s^2	71540.166	13.929
Level 2, t_p	251022.642	20.513
ICC	0.778	0.596

Research Question Five

Research Question Five explored whether classroom dispersion patterns differed and the extent to which classroom practice scores on four standards domains predicted changes in within-classroom variability on reading fluency and reading comprehension CBMs. Approximately 1.5% of variance in MCRC dispersion patterns was accounted for by the addition of classroom practice domain scores, as displayed in Table 19. However, classroom practices did not account for variance in PRF dispersion patterns between classrooms.

Table 20

Proportion of Dispersion Pattern Variance Explained After the Addition of Classroom Practice Scores

Variance Components	PRF	MCRC
Unconditional s^2	71540.166	13.929
Conditional s^2	71594.545	13.719
Pseudo R^2	-0.001	0.015

Classrooms dispersion in fall PRF varied significantly between classrooms. ($t = 17.927$, $SE = 86.526$, $p < 0.001$), with smaller differences between classrooms rated higher in instructional practice ($t = -2.735$, $SE = 285.183$, $p = 0.010$). Change in classroom dispersion over the school year was positive and significant ($t = 2.550$, $SE = 53.636$, $p = 0.016$). In other words, in classrooms with average ratings for instruction, the gap between the lowest and highest performers grew. Differences in classroom ratings were not significantly related to these changes in PRF dispersion over the school year (see Table 20).

Table 21

PRF Fall Dispersion, Dispersion Patterns Over the Year, and the Contribution of Classroom Practices, (N = 35)

Fixed Effect	Coefficient	SE	<i>t</i>	<i>p</i> -value
Fall Variance, β_{00}	1551.160	86.526	17.927	<0.001
Planning & Preparation, β_{01}	170.949	292.048	0.585	0.563
Classroom Environment, β_{02}	76.296	193.840	0.34	0.697
Instruction, β_{03}	-779.910	285.183	-2.735	0.010
Professional Responsibilities, β_{04}	188.067	237.616	0.791	0.435
Changes in Dispersion, β_{10}	136.784	53.636	2.550	0.060
Planning & Preparation β_{11}	12.099	214.603	0.06	0.955
Classroom Environment, β_{12}	50.780	125.087	0.406	0.688
Instruction, β_{13}	336.068	27.143	1.622	0.115
Professional Responsibilities, β_{14}	-304.052	228.645	-1.330	0.194

As indicated by Table 22, dispersion of students' fall MCRC scores and dispersion patterns over the school year were not significantly different between classrooms. Also, classroom practices were not related to dispersion of MCRC scores at fall or changes in dispersion patterns over time.

Table 22

MCRC Fall Dispersion, Dispersion Patterns Over the Year, and Contribution of Classroom Practices, (N = 35)

Fixed Effect	Coefficient	SE	<i>t</i>	<i>p</i> -value
Fall Variance, β_{00}	3.302	4.322	0.764	0.451
Planning & Preparation, β_{01}	3.943	4.285	0.920	0.365
Classroom Environment, β_{02}	-1.163	2.421	-0.480	0.634
Instruction, β_{03}	0.351	4.682	0.075	0.941
Professional Responsibilities, β_{04}	2.153	4.423	0.487	0.630
Dispersion Patterns β_{10}	3.821	3.112	1.228	0.229
Planning & Preparation β_{11}	0.454	2.620	0.173	0.864
Classroom Environment, β_{12}	-1.776	1.382	-1.284	0.209
Instruction, β_{13}	0.357	2.469	0.144	0.886
Professional Responsibilities, β_{14}	-1.248	2.246	-0.556	0.583

CHAPTER IV

DISCUSSION

This study examined average classroom reading performance, change in performance over time, and changes in classroom dispersion, given the reading and classroom practice measures used by one school district. The purpose was to analyze existing district data to provide information that the partner district could use to consider and enhance their accountability approach. In this real-world context, accountability systems are driven by data collection, data analysis, and data use in formative ways, which poses a variety of challenges to districts (Means, Padilla, and Gallagher, 2010). The research questions that guided this study addressed three broader topics: a) the extent of uniqueness and similarity across multiple reading measures; b) the degree to which classroom practices are related to student reading outcomes and development over a school year; and c) the challenges and potential of explicitly modeling changes in classroom dispersion to examine the Matthew effect. However, these topics should be considered within the scope and limitations of this study.

Limitations

Results were based on a sample of fourth and fifth grade data from one school district. As noted in the Results section, compared to the district population, the analytic sample included a greater proportion of students typically thought of as less advantaged including students with limited English proficiency, students of Hispanic ethnicity, and students eligible to receive free or reduced price lunch. Studies have reported that students from low-income households often start the academic year behind their peers in reading (e.g., Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996; Kim & Quinn,

2013). The pattern of missing data identified in this study could reflect decisions on the part of teachers to focus on teaching prerequisite skills or decisions to use other measures in lieu of these specific assessments, based on students' fall test results. Similarly, a greater proportion of students receiving special education services were missing scores at OAKS time 1 and OAKS time 2, compared to peers who were not designated for these services. Students receiving special education services would include those who take the alternative assessment in place of the standard OAKS assessment. A reversed pattern, where a significant proportion of LEP students were missing MCRC time 1 scores only, might reflect the use of other fall screening measures designed for students whose primary language is other than English. Differences between the analytic sample and the district population, and between students in the sample with and without reading data, severely limits the extent to which results can be generalized even to other 4th and 5th grade students.

Beyond the restricted generalizability of findings due to the specificity of the sample and statistically significant demographic differences between a less advantaged analytic sample compared to the district population, methodological choices also limit the validity of reported results and findings in and of themselves. PRF and MCRC data were analyzed using full information maximum likelihood (FIML) estimation to account for missingness. However, estimates are based on collective patterns of variability among available scores. Listwise deletion was used for OAKS data, reducing the analytic sample by 7%, from 945 down to 879 students. Listwise deletion reduces statistical power, reducing the likelihood of detecting effects or relationships should they exist, and producing biased estimates and standard errors (Enders, 2001). FIML estimation is a

stronger method for dealing with missing data, but many people believe multiple imputation is an even stronger approach. This is a particularly important limitation since examination of data for this study suggested non-random patterns of missingness for OAKS, PRF, and MCRC data. Given the limited number of classrooms representing each school, it remains unclear whether missing data reflected teacher decisions or a broader school-wide administrative policy in select schools.

The choice of analytic method affects the ability to detect existing relationships and the accuracy and credibility of reported estimates and relationships. Thus, not only are findings not generalizable to other four and fifth grade students and teachers, other grades, or scores from other assessments within the district or beyond, they also might not accurately represent student learning and relations to classroom practices for participants in this study. However, the challenges encountered and the limitations of this study, including patterns of missing data, do raise broader considerations aligned with existing research.

Uniqueness and Similarity Among Multiple Reading Measures

The amount of variance in reading scores attributed to differences between classrooms was limited for all three reading measures. This suggests that student specific factors (e.g., demographic characteristics) and contextual factors specific to the time of test administration, contributed sizably to differences in reading performance. The contribution of student and time specific factors also differed by assessment, with twice as much variance in PRF or OAKS scores attributed to differences between classrooms compared to variance in MCRC scores. However, the treatment of missing OAKS scores could have influenced the amount of variance in OAKS gains and potentially the

distribution of variance between students and classrooms. The extensive amount of missing MCRC data, particularly in spring, serves as an indicator of time specific difference in testing. Understanding the context behind large amounts of missing data that do not appear to be random in nature, as well as other key situational or student specific variables is essential for constructing models that account for the unique characteristics of the districts' students and in situ teaching, learning, and assessment decisions.

Collectively the lack of association between classroom OAKS gains with MCRC growth and CBM spring scores raises implications for choices that the district faces when considering the use of student reading data as part of a teacher evaluation system. Different measures and different indices (e.g., spring scores versus growth) could produce drastically different indications of students' reading development and success. In other words, the extent to which a class of students is viewed as performing below, above, or at an average level may differ depending on the district's choice of the reading measure. These differences could be due to ceiling effects--specifically for MCRC, given the negative skew in the distribution of scores--or other test properties (Kodel & Betts, 2009) or the nature of OAKS as a criterion-referenced test versus norm referenced CBMs (Stuit, Berends, Austin, & Gerdeman, 2014). Additionally, model specification choices--such as those made while conducting this study can influence results, limiting power and internal validity (Lockwood and McCaffrey, 2007).

Choices made in the broader context of the assessments may be equally important to consider. Beyond the monetary expense to the state and districts, assessments also cost valuable teaching and learning time. Descriptive examinations of data prior to analysis

revealed classrooms where a substantial amount of data were missing, and a few whole classrooms of students without data for CBM reading comprehension. Clear patterns of missingness limit the validity of conclusions drawn for this study, but could provide meaningful insight into classroom specific choices when it comes to assessment practices. Patterns of missingness to this extent suggest classroom choices to prioritize students' time toward other activities perceived to be of greater benefit or utility. The return on investment (time for learning versus the value of information gained from assessments) may be a consideration for teachers and worth consideration at the district level.

Contribution of Classroom Practices to Student Outcomes and Growth

Certain types of classroom practices, those relevant to instruction (i.e., lesson delivery, feedback to students, assessment for learning) in particular, were differentially related to the average classroom performance or change in performance (gains or growth) across the three assessments. However, there was limited variability in classroom practice data⁸, and across all reading assessments the addition of classroom practice scores minimally reduced variance at the classroom level. While a relationship was found between some classroom practices and students' reading development (e.g., quality of instruction with PRF growth and OAKS gains), these aspects of classroom practice as measured by the school district did not account for much of the differences that exist between classrooms.

⁸ The majority of domain averages were between two (Proficient) and three (Exemplary). Additionally, overall teaching designation was not used for this study, given that nearly all teachers received the same designation.

Selecting measures of student learning and measures of classroom practice are key decisions in and of themselves, particularly given the high-stakes nature of teacher evaluation systems in which these types of data could be used. Oregon's requirement to consider both student learning and classroom practice within Educator Evaluation and Support Systems provides flexibility for districts, but excludes specific guidelines and research-based recommendations to facilitate rigorous design and implementation. The American Educational Research Association (AERA) Council (2015) has raised critical factors regarding value-added models⁹ that can affect the precision and validity of results, many of which were related to the limitations of this study (e.g., sample size, quality of data, and how models are specified). Regarding selection of student assessments, the Council emphasized use of standardized assessments capable of measuring the full spectrum of student achievement within the classroom and those specific to the content of teaching and learning, while encouraging consideration of other data including those derived from classroom observations. Similar types of criteria are needed to enhance the design, selection, and implementation of classroom observation tools.

The legacy of process-outcome research is perhaps the development of classroom observation tools as a mechanism for gathering information about classroom environment and practices. Yet, limited attention has been paid to the validity and reliability of these tools, let alone the effects of district implementation. Classroom practice data for this study were previously collected by the district using a modified version of the Danielson's (2007) Framework for Teaching. The Danielson Framework was originally developed based on findings from literature reviews and has not undergone rigorous

⁹ Value-added models are an input-output approach to accountability which uses test data to determine the contribution of teachers or administrators for evaluation purposes

examinations of reliability and validity. Moreover, adaptations of the framework are encouraged so that included components are applicable for a given setting (Danielson, 2007). This flexibility has the potential to facilitate rich conversations among educators' and reflection on professional practices, but poses sizable measurement and assessment challenges in that every modified version is essentially a new tool in need of validation.

Building on the history of observation tools established through process-outcome research, more attention is being paid to the quality of observation tools, their focus, and the implementation procedures needed to make more valid claims regarding classroom practice (Goldring et. al, 2015). The AERA Council's (2015) recommendations that student assessments are equipped to measure the full spectrum of student achievement in the classroom and are specific to the content of teaching and learning have implications for the design of classroom observation tools. To the extent that students within a class differ in achievement, their teacher is tasked with implementing approaches that address differentiated student needs. Moreover, given differences in the content, procedures, and the specific skills demanded by different content areas, quality instructional approaches may differ between disciplines. Process-outcome researchers could consider teaching practices unique to each discipline, those that might generalize across disciplines, and those essential to supporting differentiated student needs within a classroom.

Recent research provides some insights on the need to consider the interaction between classroom observation measures and student assessments. In their value-added analysis using observation data guided by the Protocol for Language Arts Teaching Observation (PLATO), researchers reported PLATO ratings for Cognitive and Disciplinary Demand were more strongly related to students' scores from the Stanford

Achievement Test (SAT-9) than state test scores (Grossman, Cohen, Ronfeldt, & Brown, 2014). The difference could be due to better alignment between what the observation protocol targets and the student assessment outcomes. As Grossman and colleagues noted, the PLATO's focus on ambitious instructional practices aligns with the SAT-9 focus on more ambitious outcomes. In other words, the degree to which indicators of classroom practices relate to student outcomes could be a function of the combination of student assessments and the practices targeted by the classroom observation measures used for accountability models. While the current study examined accountability implications given different student measures of reading, the partner district might also consider how accountability measures align as a larger system.

Researchers have identified a variety of factors that could influence the results of this study and others like it. These factors include, but are not limited to a) the design of the teacher evaluation tool (e.g., Allen, et al., n.d; Hamre, Pinanta, Mashburn, & Downer, 2007); b) variation in how evaluators interpret each standard (Sartain, Stoelinga, & Brown, 2011); c) differences in the composition of students in each class (Newton, Darling-Hammond, Haertel, and Ewart, 2010); or d) specific characteristics of the measurement occasion (Concoran, 2011; Papay, 2011) as in this case where whole classrooms of missing data at time points. Many of these factors may have impacted the results of the current study. As previously noted, the districts evaluation tool was based on the Danielson Framework for Teaching, which has not yet undergone rigorous examinations of reliability and validity. Even if such information were available, district modifications necessitate vetting of their unique tool. Moreover, the rating options were limited to four levels (Unsatisfactory, Basic, Proficient, or Exemplary), restricting the

ability of evaluators to differentiate the degree to which each standard was met. Ratings of Proficient or Excellent were especially common, indicating room for additional levels of performance between those two categories, accompanied by more fine-grained descriptions of evidence for each category. Lack of differentiation in ratings reduced variance in classroom practice data. This, along with the inclusion of constructs such as Commitment to Instructional Initiatives that might not directly inform the quality of instruction, may have been a factor in the limited relationships reported in this study between classroom practices and student learning.

Additionally, results from the current study may reflect policy decisions that could contribute to or impeded successful implementation of the teacher evaluation tool (Sartain, Stoelinga, & Brown, 2011). The design of classroom observer training is one policy aspect related to successful implementation of the accountability model. As part of an initial training, the district could embed opportunities for observers to score the same classrooms, similar to the process described by Kane et al., 2011. This would allow the district to determine inter-rater reliability, diagnose domains and standards where evaluators differ the most, and ensure raters reach a certain threshold for reliability. However, initial training alone might not be sufficient. In a study of classroom observer trends, Casabianca, Lockwood, and McCaffrey (2015) found the severity or leniency of individual observers' scores changed slightly over time. As a group, observers did not reflect a shared level of rating severity during initial observations, and differences persisted and increased over time. Follow up professional development opportunities for evaluators could be used to diagnose observer drift in terms of different interpretations of constructs as well as differences in the forms and extent of evidence raters look for.

Dispersion Patterns and the Matthew Effect

Since the Coleman report (1966), equitable opportunities for all students to learn has been a priority of public education and remains a focus today as reflected in the Every Student Succeeds Act (United States Executive Office of the President, 2015). However, analytic techniques such as value-added modeling often focus on average scores and average growth (e.g., Ballou, Sanders, & Wright, 2004), whether comparing average growth between classrooms or average growth between students from different demographic populations. In this study, the amount of variance in reading scores over time, attributed to differences between classrooms, was substantially less than the variance in dispersion patterns of reading scores. This contrast suggests that dispersion models could provide an additional avenue to examine discrepancies in equitable learning outcomes and opportunities, by analyzing the diversity of performance and changes in the spread of scores over time.

The significant relationship between classroom instruction and dispersion of fall PRF scores identified in this study likely reflects factors related to classroom composition more than quality of instruction, given that student would not have been exposed to much instruction by the fall administration of CBM assessments. Findings by Newton, Darling-Hammond, Haertel, and Ewart (2010) noted the value of accounting for the proportion of students in a class who represent demographics of interest, above and beyond inclusion of demographic indicators at the student level. Hill et al. (2011) reported correlations between value-added scores and the population of student that teachers serve, reinforcing the argument to consider and account for classroom composition.

However, just as the evaluation of average growth requires contextual knowledge (Sartain, Stoelinga, & Brown, 2011), so does the accurate interpretation and evaluation of dispersion patterns. A narrowing dispersion pattern where students become more similar in their performance over time reflects the common notion of closing the achievement gap. Yet this pattern could be reflected in more than one situation. In the ideal case, all students remained challenged and continue to grow, with some students demonstrating greater growth as they catch up to their peers. Alternatively, a narrowing pattern could result from situations where the classroom focus is primarily on students who need more support, at the expense of continued growth for accelerated learners. A third situation reflecting the same pattern, could be the result of a ceiling effect, in which the narrowing pattern reflects students who perform at the highest level allowed by the assessment, or the lack of utility in performing any better (e.g., the point at which reading more words correctly per minute is no longer supporting the ability to make sense of text).

Future Studies

There are many alternative ways in which data for this study and others like it could be analyzed to benefit the partner school district as well as the broader field of educational research. Pursuing such alternatives could elucidate the impact of analytic choices, not only regarding how missing data are handled, but to further examine the impact of including student characteristics, and proportions of students in each classroom that represent those demographic factors. Research that could inform districts' teacher evaluation and support models is especially pertinent given the limited guidance currently provided by the ODE regarding rigorous approaches and model design implications.

Additionally, results from the dispersion models used in this study suggest the potential for this analytic technique. Score based models provide a sense of student and classroom progress in terms of averages, but do not provide indications of how a class as a whole becomes more similar or different in performance overtime. Dispersion patterns combined with average classroom growth could serve as a *low stakes* and *informal* indicator. An increase in the spread of scores over time combined with declining or no change in average growth, could signal a need to learn more about a classroom. Future studies could explore how differences in the proportion of students representing demographic characteristics of interest relate to classroom dispersion patterns. For example, such an approach could be used to examine the extent to which classrooms with higher proportions of Hispanic students are showing a growing gap in student performance, a shrinking gap, or other patterns over the course of the school year. A growing gap associated with higher proportions of Hispanic students district wide may warrant exploration of classroom practices, whether practices are related to slower growth rates for Hispanic students, and if so, how to enhance support for teachers in meeting the needs of Hispanic students. However, dispersion models are challenging to interpret without understanding what is occurring in classrooms, and must be interpreted using contextual knowledge of observed classroom practices.

CHAPTER V

CONCLUSION

Similar to prior validity studies of classroom practices and student learning (e.g. Sartain et al. 2011; Kane et al., 2011), findings from this study support Messick's (1989) argument that scores, ratings, and other judgments are based on what is presented (e.g., aspects of the assessment), who it is presented to (e.g., unique qualities of the student), and the context in which presentation occurs (e.g., testing, classroom, and district context). Classrooms and school districts are dynamic places. Decisions are made at every level by district leadership, school leadership, teachers, students, and family and community members that affect students on a regular basis. In the case of this school district's accountability system, the choice of reading measure matters. Different reading measures provide different information on student reading gains, growth, or end of year performance, and vary in their utility in informing classroom practices.

Choices when implementing a classroom evaluation system also matter. Districts choose the aspects of classroom practice to prioritize in their evaluation system. Districts design the process for training evaluators, which includes how they build evaluator understanding of the constructs, what evidence evaluators would/could look for, and the scope and/or quality of evidence needed to support different levels of ratings. These among other choices frame what practices are examined from those that are not, as well as the extent to which valid arguments can be drawn about the relation between these practices and student learning. At the classroom level, decisions are made when weighing the assessment demands of an accountability system with the best use of time for

students' development. All of these decisions influence the data collected and analyzed by researchers.

Researchers' analytic choices, including methods for addressing missing data, affect the results of this study and others like it. Models of dispersion have the potential to enhance understanding of gaps in achievement and growth and how they differ based on the unique practices in each classroom. However, researchers rely on district knowledge--including the choices made at each level of the accountability system--to construct more valid interpretations and arguments. Likewise, administrators rely on valid research findings to inform changes to their accountability system. Applied research can play a crucial role as states and districts grapple with designing accountability systems that balances the unique needs of local communities with ensuring all student have the opportunities needed for post-secondary success. Yet, reflecting the theme of this study, the extent to which applied researchers can have an effect will depend on their choices in cultivating relationships with practitioners.

APPENDIX A

DISTRICT'S STANDARDS FOR TEACHING PRACTICES

Teacher Evaluation Standards

Performance targets are defined by A, B, or C, and correspond to the following years of a teacher's career in Springfield:

- A = Probationary Year One
- A, B = Probationary Year Two
- A, B, C = Probationary Year Three and Contract Teachers

Domain I: Planning and Preparation
Standard 1: Knowledge of Content
A. Shows an effective command of the subject to guide student learning.
B. Uses effective instructional resources, including technology, to communicate content knowledge.
C. Takes an active role in adapting new content standards and frameworks to their teaching.
Standard 2: Knowledge of Students
A. Builds upon students' knowledge and experiences.
B. Uses school and district resources to support and advocate for student needs.
C. Recognizes and addresses students' learning styles.
Standard 3: Instructional Goal Setting
A. Selects appropriate instructional goals based upon national, state, and local standards.
B. Selects goals that are measurable and states them in terms of student learning.
C. Selects goals appropriate for students with different learning styles and cultural backgrounds.
Standard 4: Curriculum Design
B. Designs coherent instruction that reflects researched-based best practices.
B. Designs instruction that promotes critical thinking and problem-solving.
C. Ensures that the curriculum is relevant to student needs.
Standard 5: Assessment Planning
A. Is familiar with content area, school, district, and state assessment methods and options.
B. Uses assessments that are congruent with instructional goals.
C. Develops and uses formative and summative assessment tools and information for planning, instruction, feedback, and reflection.
Domain II: Classroom Environment
Standard 6: Climate of Respect and Learning
A. Creates an environment that promotes equity, respect, and positive interpersonal interactions.
B. Interactions are appropriate to developmental and cultural norms.
C. High expectations for student success, quality work, and student achievement.
C. Active participation of students is evident.
Standard 7: Classroom Procedures and Physical Environment
A. Develops and employs classroom procedures that promote student learning and facilitates positive classroom interactions.
A. Designs a safe and accessible classroom environment for all students.
B. Facilitates smooth transitions with little loss of instructional time.
C. Ensures all students have access to materials, technology, and necessary resources.

Standard 8: Managing Student Behavior
A. Clearly communicates and enforces classroom expectations.
A. Addresses inappropriate behavior consistently, appropriately, and predictably.
B. Monitors inappropriate behavior in a preventive way.

Domain III: Instruction

Standard 9: Lesson Delivery
A. Exhibits lesson delivery that is clear, reflects appropriate pacing, and uses a variety of strategies.
B. Activates students' prior knowledge.
C. Differentiates instruction to meet the needs of diverse learners.
C. Uses a variety of questioning and discussion techniques that elicit student reflection and higher order thinking.

Standard 10: Feedback to Students
A. Demonstrates an ability to listen to students so that feedback is more effective and received in a positive way.
B. Provides feedback that facilitates learning and academic growth.
C. Provides feedback that is consistent, ongoing, timely, and in a variety of forms.

Standard 11: Assessment for Learning
A. Uses assessment data to prepare for individual and group instruction.
B. Uses formative assessment during classroom instruction to facilitate student learning.
C. Demonstrates the ability to use summative assessment to guide and inform instruction through the collection, maintenance, and analysis of classroom, district, and state assessment data.

Domain IV: Professional Responsibilities

Standard 12: Professional Growth
A. Is aware of and pursues professional development including teacher leadership opportunities.
B. Actively engages in meaningful goal setting.
C. Pursues professional growth through reflection, self-assessment, lifelong learning, and being knowledgeable about best practice.

Standard 13: Professionalism
A. Carries out duties as assigned.
B. Maintains accurate records.
C. Is available to others and provides support when necessary.

Standard 14: Communication
A. Communicates effectively and respectfully with all stakeholders: students, parents, colleagues, and supervisor.
A. Facilitates meetings effectively if required.
B. Collaborates with colleagues and other professionals.

Standard 15: Commitment to Instructional Initiatives
A. Is aware of and supports building and district instructional priorities.
B. Knows and applies strategies that facilitate continuous progress on building and district instructional initiatives.

APPENDIX B

DISTRICT'S SUMMATIVE TEACHER EVALUATION FORM

Original to Human Resources
Copy to Supervisor
Copy to Teacher

Summative Evaluation Form²

Teacher Name _____ Probationary: A: B: C: Contract:

Supervisor: _____ School: _____ Assignment: _____

Domain I: Planning & Preparation

	U	B	P	E
Standard 1: Knowledge of Content				
Standard 2: Knowledge of Students				
Standard 3: Instructional Goal Setting				
Standard 4: Curriculum Design				
Standard 5: Assessment Planning				
Comments:				

Domain II: Classroom Environment

	U	B	P	E
Standard 6: Climate of Respect and Learning				
Standard 7: Classroom Procedures and Physical Environment				
Standard 8: Managing Student Behavior				
Comments:				

² The following definitions are used as a guide to evaluate a teacher's performance. **Unsatisfactory (U)**: The performance is unacceptable and must improve significantly. **Basic (B)**: The performance is satisfactory, but there are specific areas that can be improved. **Proficient (P)**: The performance is strong, and there are minimal weaknesses. **Exemplary (E)**: The teacher's skills in this content area are in the top 10% of their field and serve as a model and example to other teachers and administrators. There must be significant evidence for a teacher to be ranked Exemplary on a teaching standard.

Summative Evaluation Form

Domain III: Instruction

	U	B	P	E
Standard 9: Lesson Delivery				
Standard 10: Feedback to Students				
Standard 11: Assessment for Learning				
Comments:				

Domain IV: Professional Responsibilities

	U	B	P	E
Standard 12: Professional Growth				
Standard 13: Professionalism				
Standard 14: Communication				
Standard 15: Commitment to Instructional Initiatives				
Comments:				

Overall Comments and Recommendation:

Satisfactory: Not Satisfactory: Directed Performance Goals:

Proceed to:
 Probationary B: C: Professional Growth Cycle: (Contract Teachers Only)

This evaluation has been discussed between the supervisor and teacher.

The teacher has attached comments to this evaluation: Yes No

Teacher _____ Supervisor _____ Date _____

APPENDIX C

DESCRIPTION OF CBM VOCABULARY ASSESSMENT AND MISSINGNESS

The easyCBM™ vocabulary measures use items designed around the Oregon State Standards for Vocabulary and various word lists (e.g., Fry) to determine words (Alonzo, Anderson, Park, & Tindal, 2012; Park, Irvin et al., 2011). Each fourth grade test form contains 20 items, each written as a sentence with the target vocabulary word in bold. Students are asked to select the meaning of the bolded word out of three possible options. Students can earn up to 20 points total, one for each item answered correctly.

Items for easyCBM™ vocabulary were selected for use within progress monitoring and benchmark forms based on distractor analysis results and Mean Square Outfit, Standard Error of Measure, and Measure indices from a 1PL Rasch model (Alonzo et al., 2012). Mean measure across ten vocabulary progress monitoring forms was -0.07, ranging from -0.09 to -0.05. Mean measure across the benchmark forms was -0.06, ranging from -0.06 to -0.05. Initial criterion validity analyses reported a moderate correlation ($r_s = .63$) between the easyCBM™ vocabulary measure and the Gates-MacGinitie Word Knowledge subtest. Given that the Gates-MacGinitie measures idioms, parts of speech, and word meaning (Lai, et al., 2013), this correlation indicates a moderate degree of concurrent validity, while also reflecting measurement of distinct information on students' word knowledge. Fourth grade easyCBM™ cut-points of 14 and 18 for fall and spring, respectively, were established to identify students at-risk of not meeting OAKS cut-scores (Park, Anderson et al., 2011). Sensitivity and specificity were 0.87 and 0.79 for the fall cut score, and 0.83 and 0.83 for spring. Additional analyses using two randomly selected groups of fourth grade students suggested that the established cut-points are relatively stable (Park, Irvin et al., 2011).

Due to the relatively newer release of the easyCBM™ Vocabulary measure during the years of focus for this study, only fall and spring benchmark scores were available for students within the sample. As a result, the winter time point was missing for all students in the sample, precluding use of growth modeling. Approximately 45% of students were missing spring data (see Table 23), and missingness at spring was associated with school (see Table 24). However, due to the limited number of classrooms associated with each school, it is unclear whether missingness is due to school or classroom level decisions.

Table 23. Proportion of Students Missing Vocabulary Data, by Data Pattern, N = 945

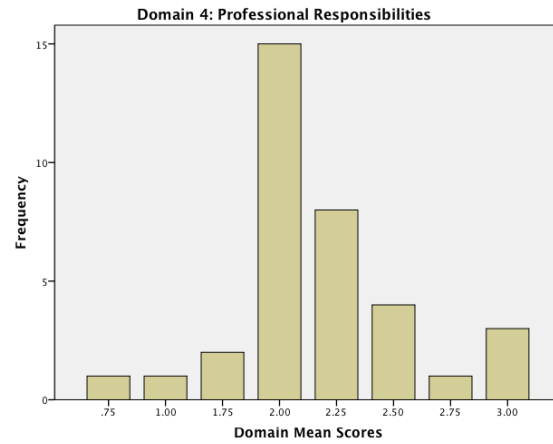
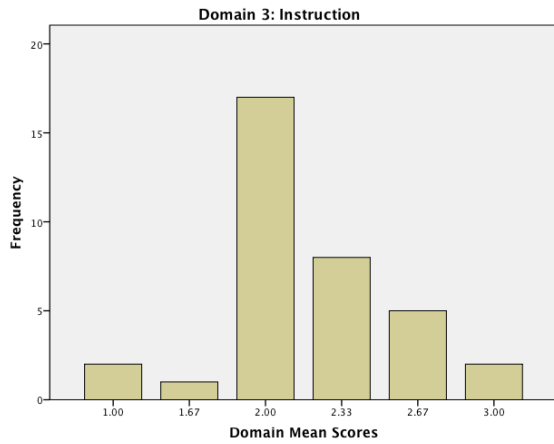
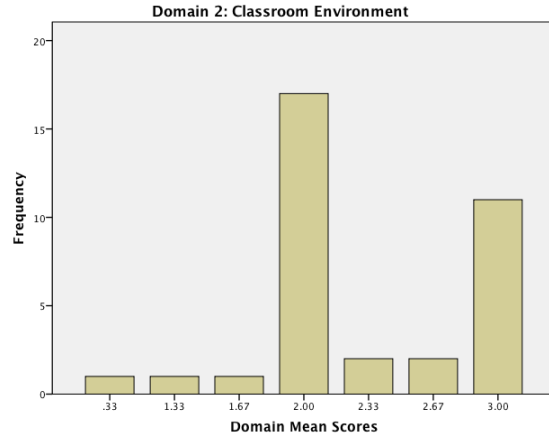
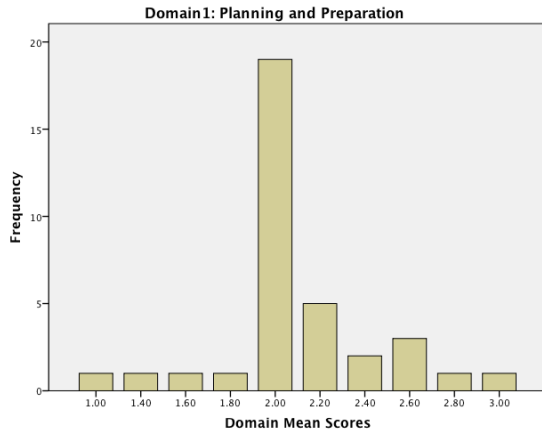
Missing Data Pattern	n	%
No missing data	472	49.9
Missing Fall	20	2.1
Missing Spring	424	44.9
Missing Fall and Spring	29	3.1

Table 24. Association Between School and Missing Vocabulary Data, by Assessment Occasion

Assessment	Not Missing		Missing		$\chi^2(12)$	<i>p</i>
	<i>n</i>	%	<i>n</i>	%		
Fall Vocabulary	896	95	49	5	18.32	.106
Spring Vocabulary	492	52	45	48	321.69	< .001

APPENDIX D

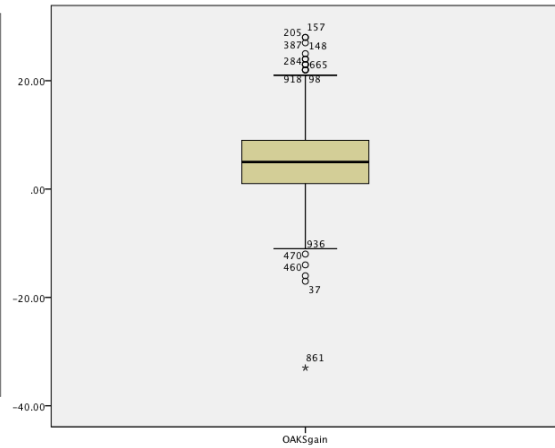
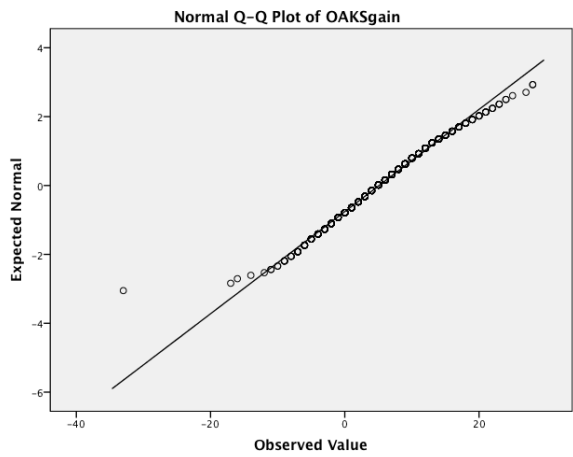
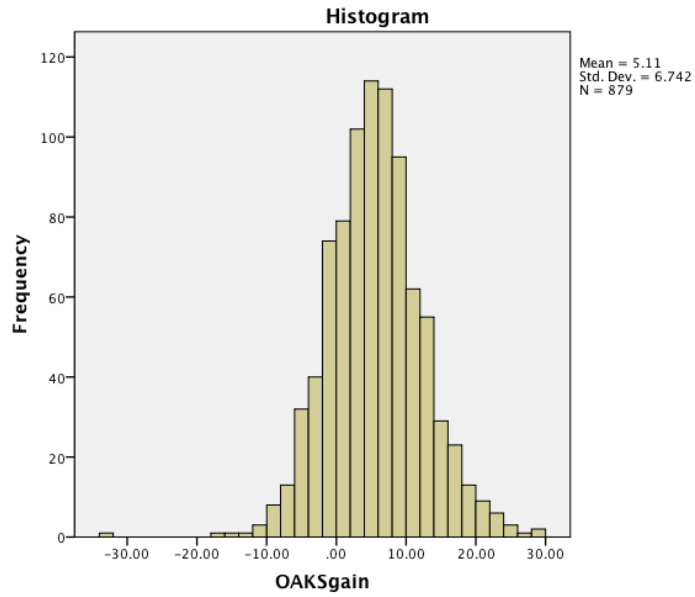
HISTOGRAMS OF CLASSROOM PRACTICE DOMAIN AVERAGES



APPENDIX E

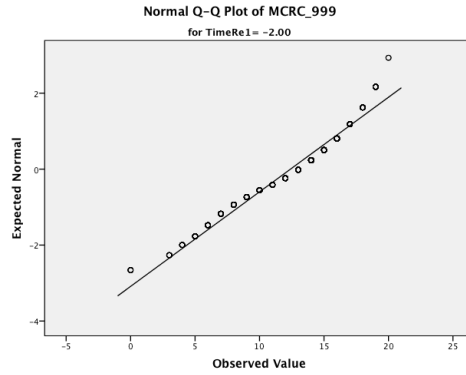
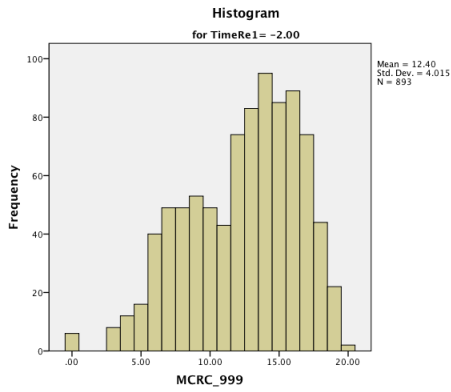
HISTOGRAMS, Q-Q PLOTS, AND BOXPLOTS OF OAKS GAINS AND MCRC AND PRF SCORES BY SEASON

OAKS GAINS

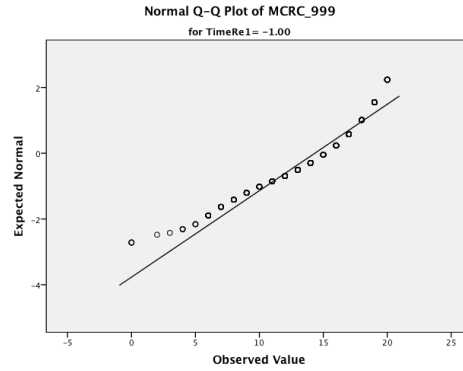
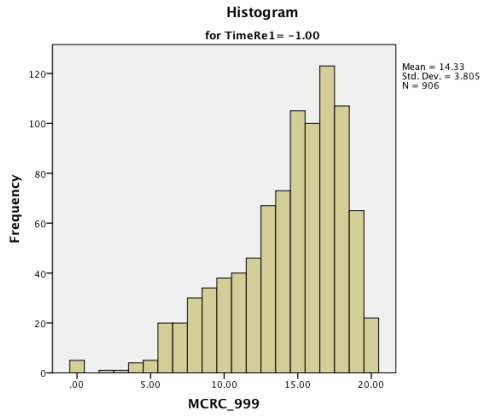


MCRC SCORES

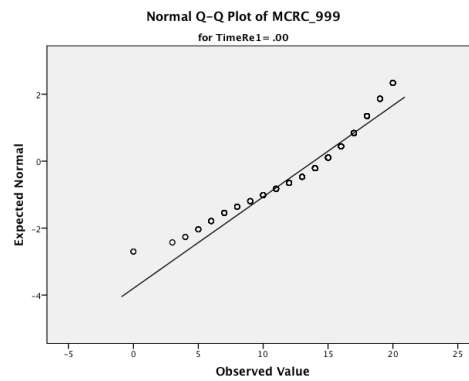
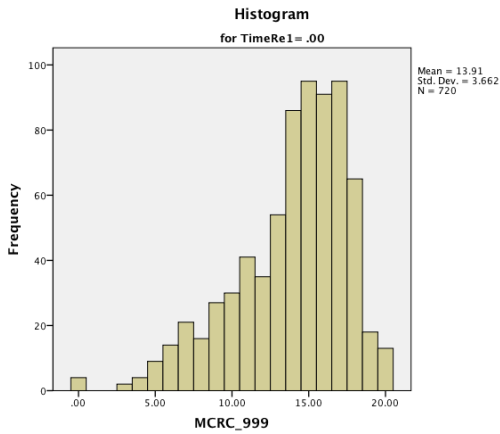
MCRC Fall



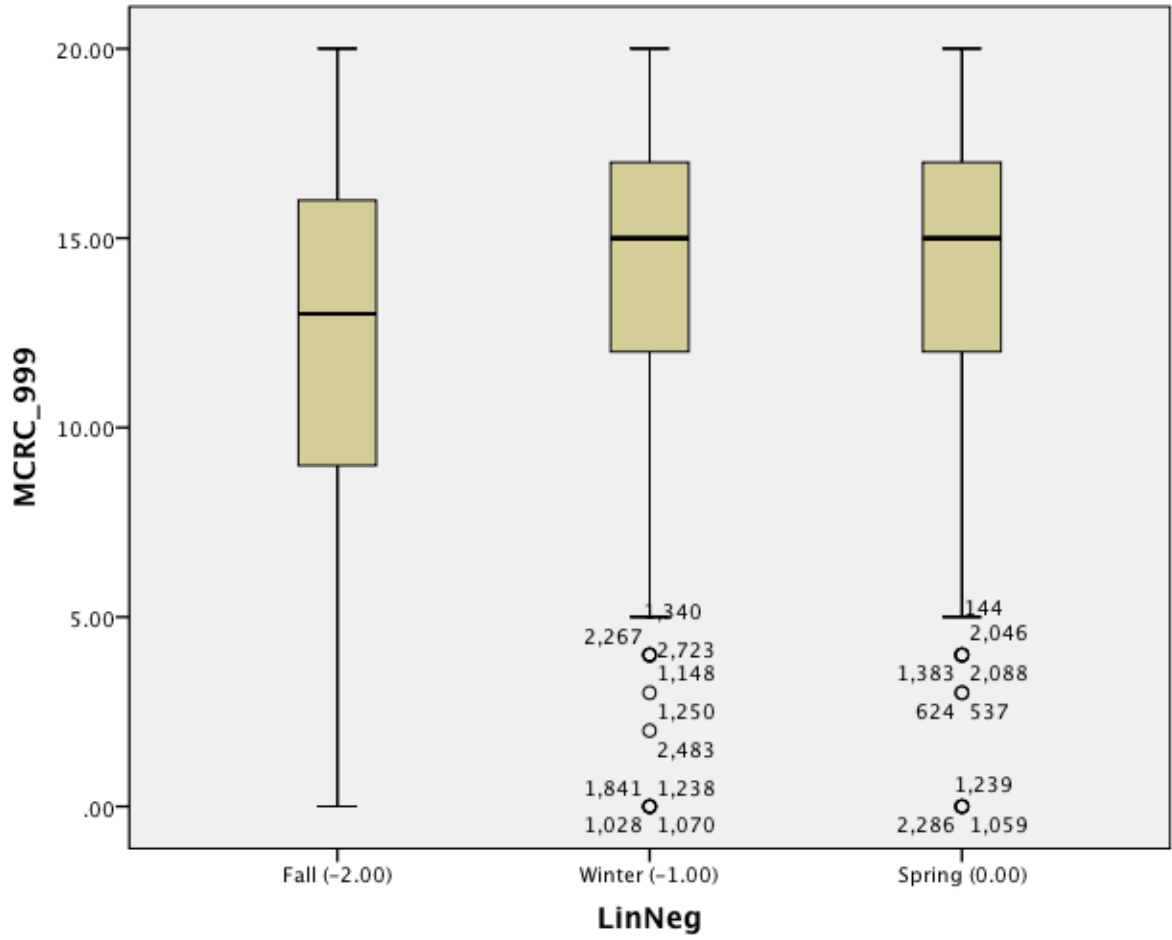
MCRC Winter



MCRC Spring

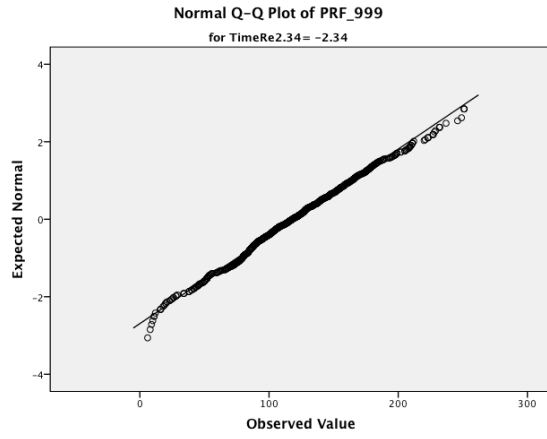
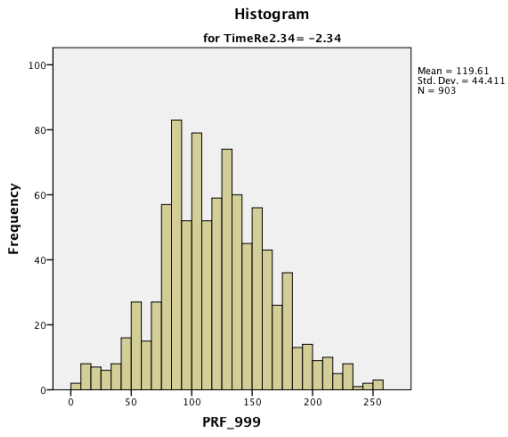


MCRC: Fall, Winter, Spring

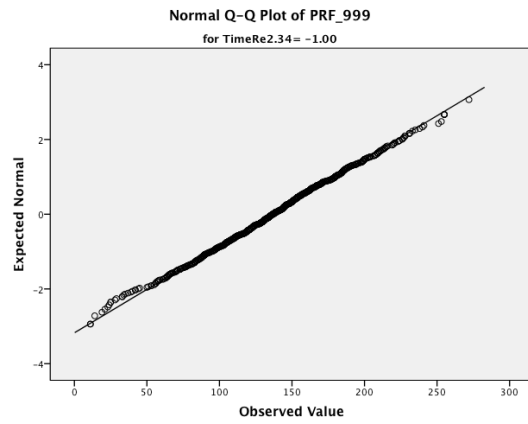
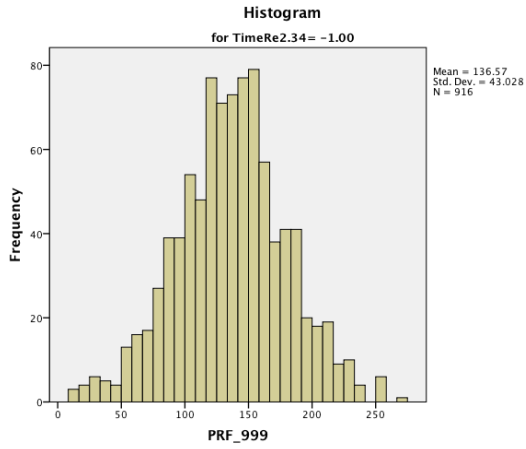


PRF SCORES

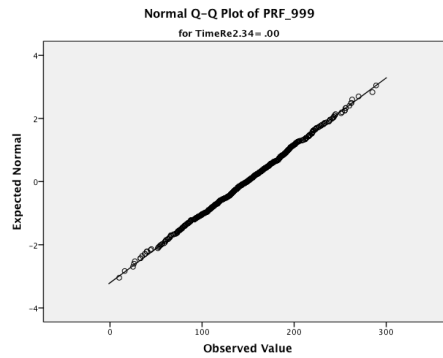
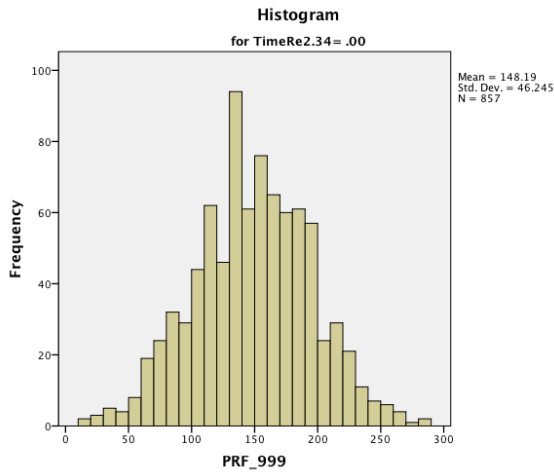
PRF Fall



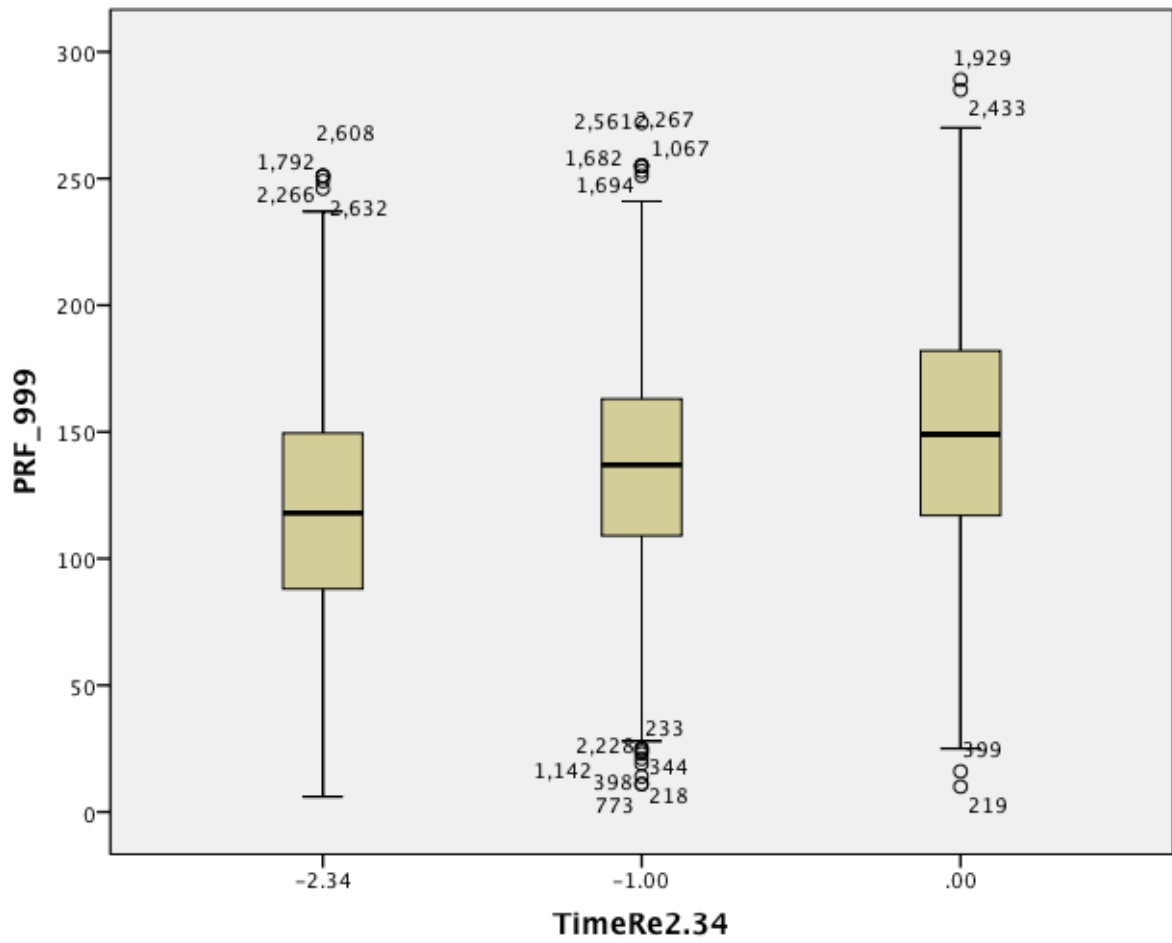
PRF Winter



PRF Spring



PRF: Fall, Winter, Spring



APPENDIX F
 SCATTER PLOTS OF CLASSROOM READING MEANS AGAINST DOMAIN
 SCORES ACROSS SEASONS

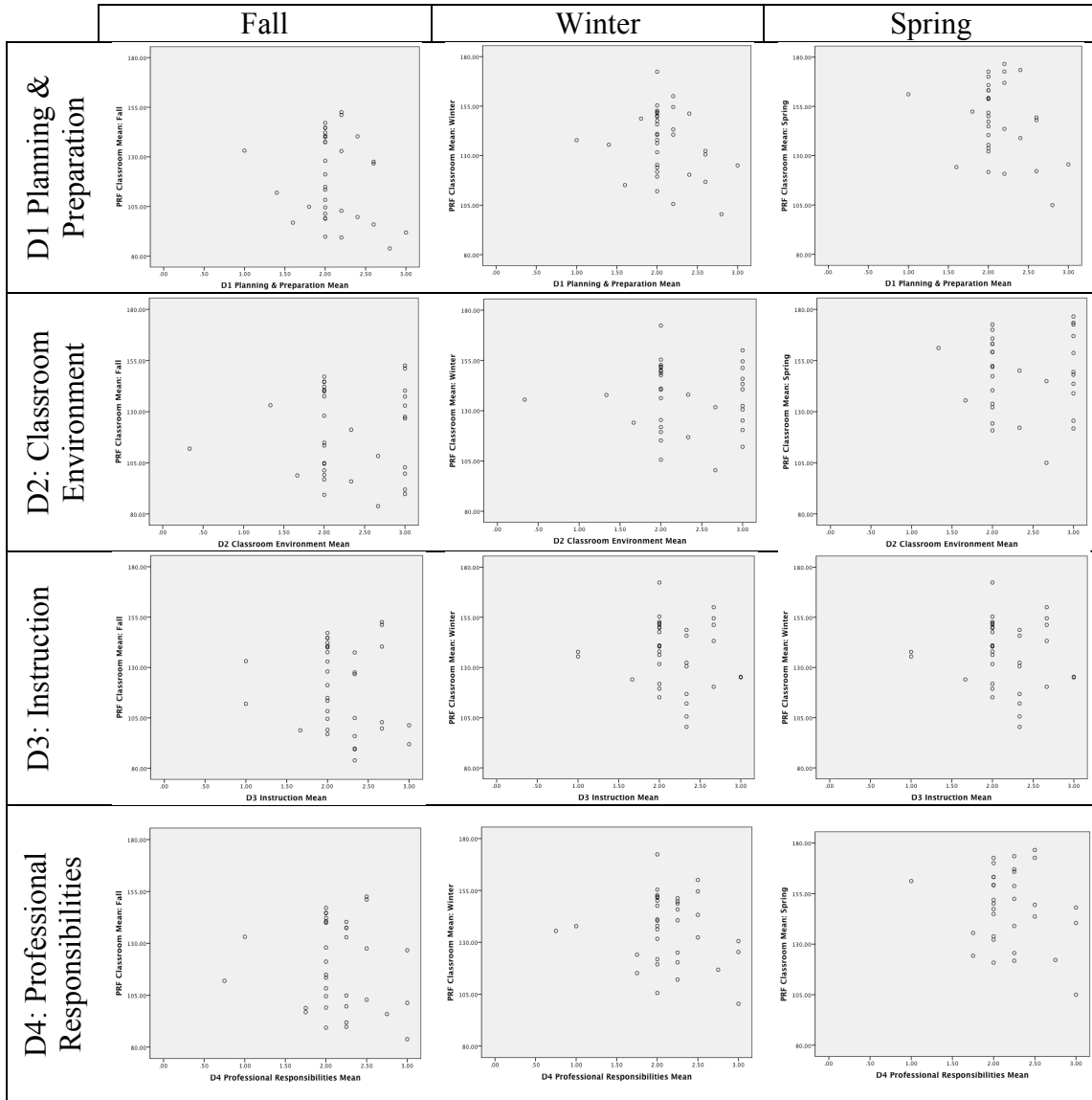


Figure 1. Scatter plots of average classroom PRF means by domain standards, for fall, winter, and spring.

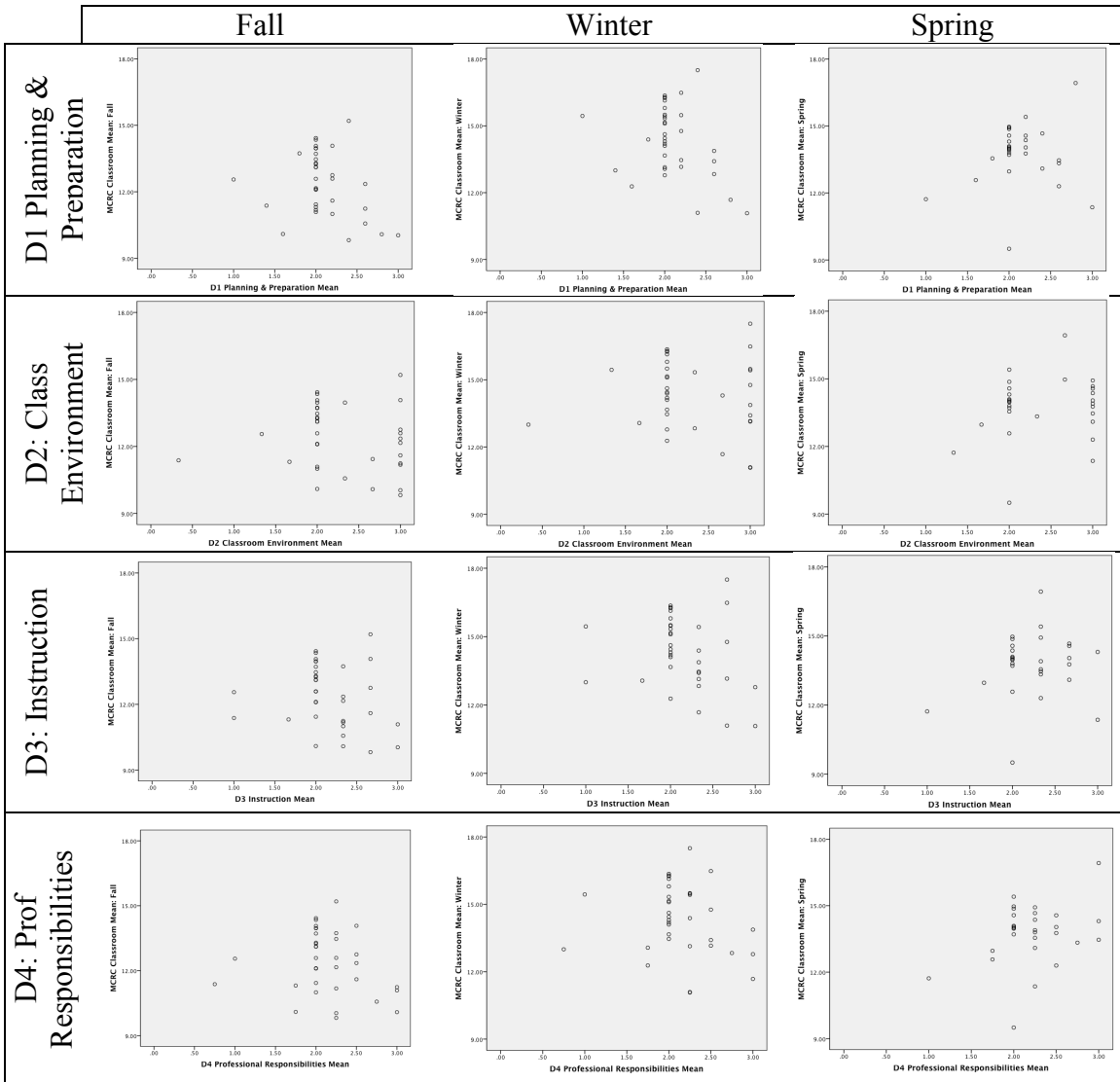


Figure 2. Scatter plots of average classroom MCRC means by domain standards, for fall, winter, and spring.

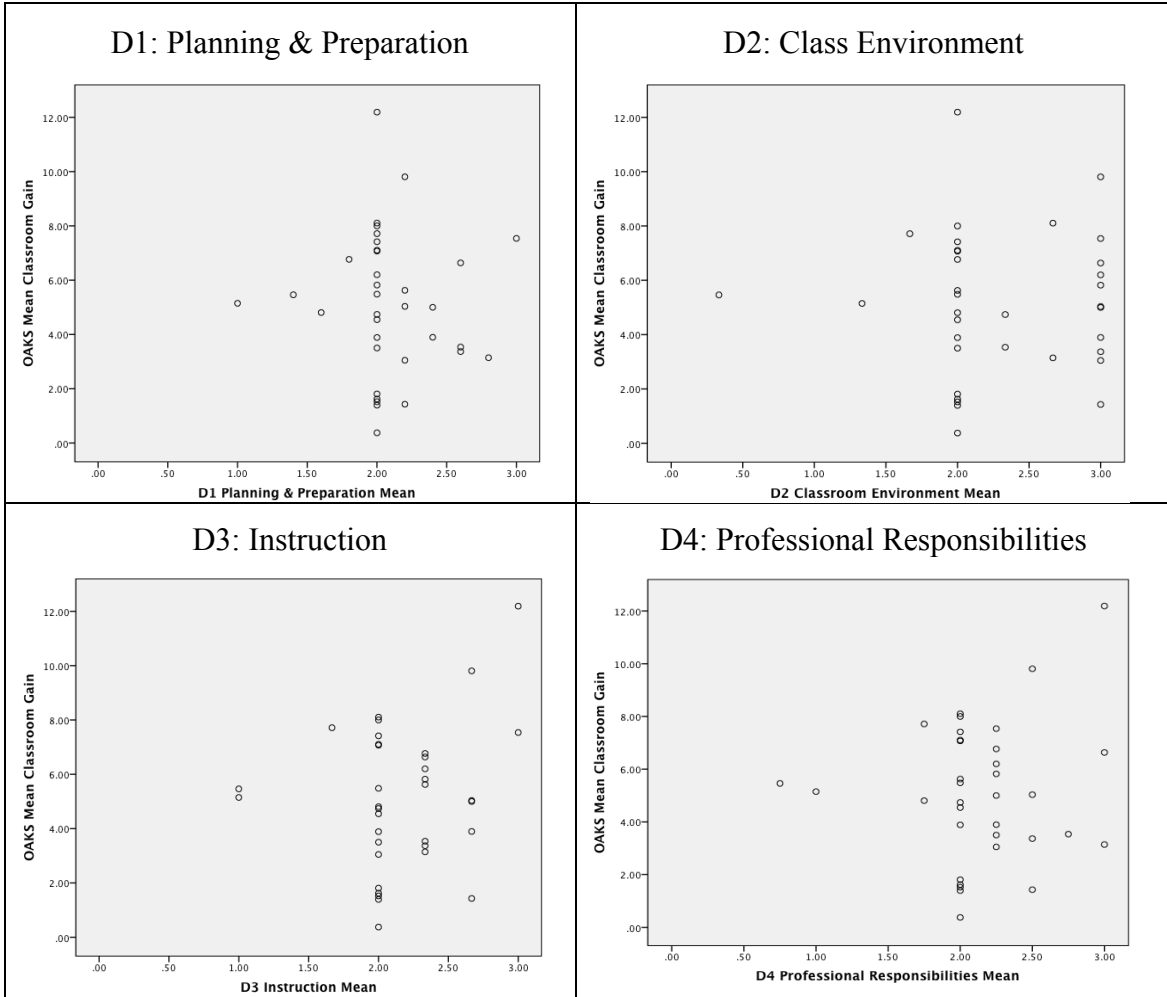
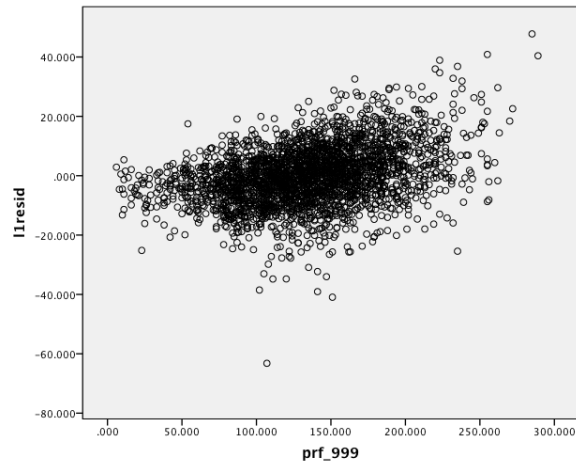
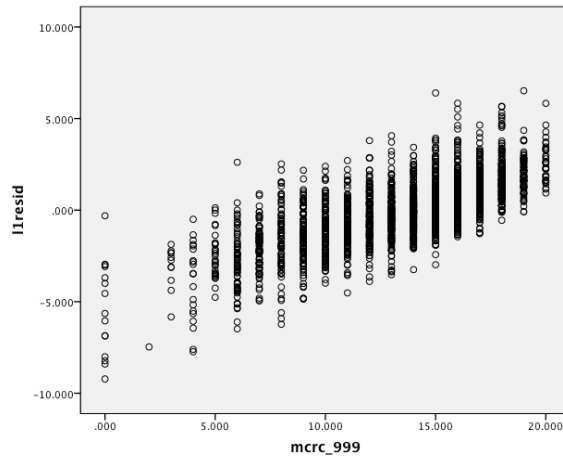
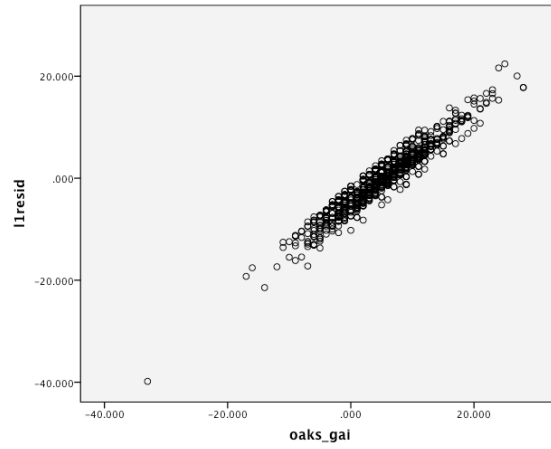


Figure 3. Scatter plots of average classroom OAKS gains by domain standards, for fall, winter, and spring.

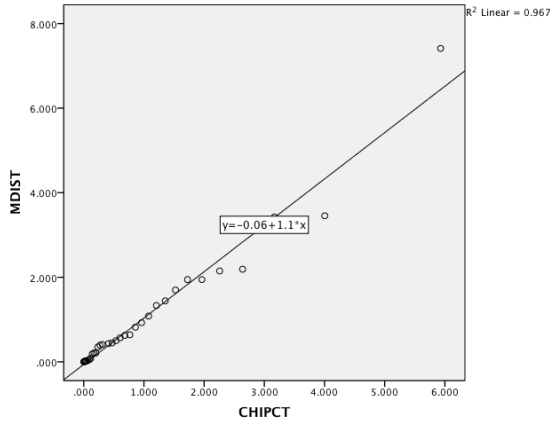
APPENDIX G
BIVARIATE SCATTER PLOTS:
PREDICTED READING SCORES AGAINST RESIDUALS



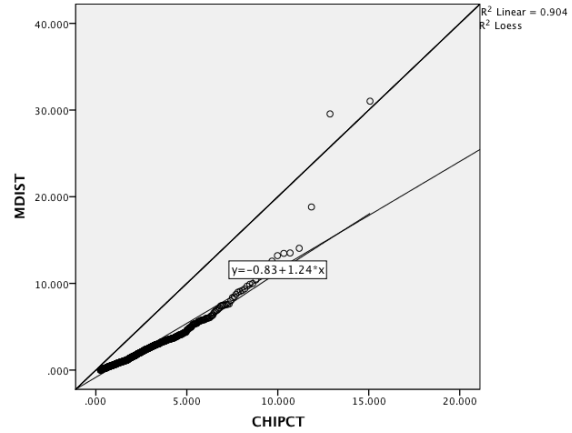
APPENDIX H

SCATTERPLOTS OF MAHALANOBIS DISTANCES AGAINST EXPECTED CHI-SQUARE DISTRIBUTION FOR OAKS GAIN, MCRC, AND PRF

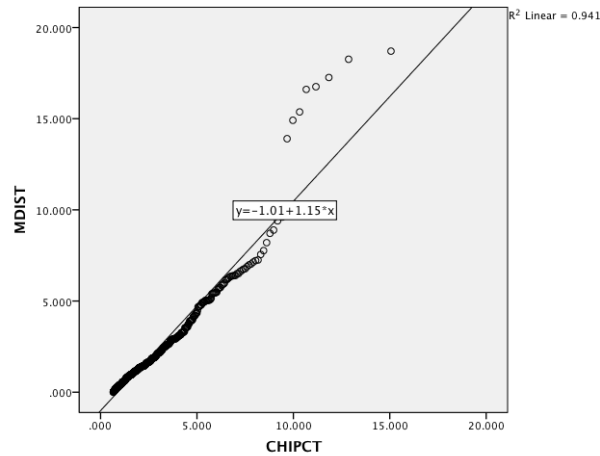
OAKS Mahalanobis Distances Against Expected Chi-Square Distribution Values



PRF Mahalanobis Distances And Expected Chi-Square Distribution Values



MCRC Mahalanobis Distance And Expected Chi-Square Distribution Values



REFERENCES CITED

- AERA Council. (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. Retrieved from <http://edr.sagepub.com/content/early/2015/11/10/0013189X15618385.full.pdf+html>
- Allen, J. P., Gregory, A., Mikami, A., Lun, J., Hamre, B., Pianta, R. C. (n.d.). *Predicting adolescent achievement with the CLASS-S Observation Tool*. Charlottesville, VA: CASTL, University of Virginia, Curry School of Education.
- Alonzo, J., Anderson, D., Park, B. J., & Tindal, G. (2012). The Development of CBM Vocabulary Measures: Grade 4 (Technical Report No. 1211). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Alonzo, J., Lai, C. F., Anderson, D., Park, B. J., & Tindal, G. (2012). An Examination of Test-Retest, Alternate Form Reliability, and Generalizability Theory Study of the easyCBM Reading Assessments: Grade 4 (Technical Report No. 1219). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Alonzo, J., Liu, K., & Tindal, G. (2007). *Examining the technical adequacy of reading comprehension measures in a progress monitoring assessment system* (Technical Report No. 41). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Alonzo, J., & Tindal, G. (2007). *The development of word and passage reading fluency measures in a progress monitoring assessment system* (Technical Report No. 40). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Alonzo, J., & Tindal, G., (2012). *Teachers' manual for Regular easyCBMTM: Getting the most out of the system*. Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- 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), p.37-65. Retrieved from <http://www.jstor.org/stable/3701306>
- Bill & Melinda Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf
- Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J., (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, 31(4), 416-440. doi: 10.3102/0162373709353129

- Brophy, J. Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Third handbook of research on teaching* (pp. 328-375). New York: Macmillan.
- Chingos, M. M. & Peterson, P. E. (2010). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review*, 30, 449-465.
doi:10.1016/j.econedurev.2010.12.010
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Department of Health, Education, and Welfare.
- Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research*, 66(3), 227–268.
- Corcoran, S. P., Jennings, J. L., & Beveridge, A. A. (2011). *Teacher effectiveness on high- and low-stakes tests* (Paper). Retrieved from https://files.nyu.edu/sc129/public/papers/corcoran_jennings_beveridge_2011_wkg_teacher_effects.pdf
- Council of Chief State School Officers (2013). Interstate Teacher Assessment and Support Consortium InTASC Model Core Teaching Standards and Learning Progressions for Teachers 1.0: A Resource for Ongoing Teacher Development. Washington, DC: Author.
- Croninger, R. G., King Rice, J., Rathbun, A., & Nishio, M. (2007). Teacher qualifications and early learning: Effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review*, 26, 312-324.
doi:10.1016/j.econedurev.2005.05.008
- Danielson, C. (2007). *Enhancing Professional Practice: A Framework for Teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.
- Darling-Hammond, L. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching*. Stanford Center for Opportunity Policy in Education.
- Doherty, K. M., & Jacobs, S. (2013). State of the states 2013 Connecting the Dots: Using evaluations of teacher effectiveness to inform policy and practice. Washington, DC: National Council on Teacher Quality.
- Dwyer, C. A. (1994). Development of the knowledge base for the PRAXIS III: Classroom performance assessments assessment criteria. Princeton, NJ: Educational Testing Service.

- Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling* 8(1), 128-141.
- Fowler, F. C. (2009). *Policy studies of educational leaders: An introduction* (3rd ed.) Boston, MA: Pearson Education Inc.
- Gage, N. L., & Needels, M. C. (1989). Process-product research on teaching: A review of criticisms. *The Elementary School Journal*, 89(3), 253-300. Retrieved from <http://www.jstor.org/stable/1001805>.
- Goe, L. Bell, C., Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.
- Goldring, E., Garissom, J. A., Rubin, M., Meumerski, C. M., Cannata, M., Drake, T., and Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, 44(2), 96-104.
- Greenwald, R., Hedges, L. V., & Laine, R. D. (1996). The effects of school resources on student achievement. *Review of Educational Research*, 66(3), 361-396.
- Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Research*, 43(6), 293-303
- Hamre, B. K., Pianta, R. C., Mashburn, A.J., Downer, J.T. (2007). Building a science of classrooms: Application of the CLASS Framework in over 4,000 U.S. early childhood and elementary classrooms. Retrieved from [http://fcdus.org/sites/default/files/ BuildingAScienceOfClassroomsPiantaHamre.pdf](http://fcdus.org/sites/default/files/BuildingAScienceOfClassroomsPiantaHamre.pdf)
- Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public school, *Journal of Economic Literature*, 24(3), 1141-1177. Retrieved from <http://www.jstor.org/stable/2725865>
- Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, 100(1), 84-117. Retrieved from <http://www.jstor.org/stable/2138807>
- Hanushek, E. A. (2010). Education Production Functions: Developed Country Evidence. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (Vol. 2, pp. 407-4011). Oxford: Elsevier.
- Hill, H. C., Kapitula, L., Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, 48(3), 794-831. doi: 10.3102/0002831210387916

- Ho, A. D., & Kane, T. J. (2013, January). *The reliability of classroom observations by school personnel*. Retrieved from <http://files.eric.ed.gov/fulltext/ED540957.pdf>
- Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly*, 23,27-50.
- Hung, F. L., & Moon, T. R. (2009). Is experience the best teacher? A multilevel analysis of teacher characteristics and student achievement in low performing schools. *Educational Assessment, Evaluation, and Accountability*, 1(3), 209-234.
- IBM Corp. Released 2013. IBM SPSS Statistics for Macintosh, Version 22.0. Armonk, NY: IBM Corp.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527-535.
- Kane, T. J., Taylor, E. S., Tyler, J. H., Wooten, A. L., (2011). Identifying effective classroom practices using student achievement data. *The Journal of Human Resources*, 46(3), p. 587-613.
- Kim, J. S., & Quinn, D., M. (2010). The effects of summer reading on low-income children's literacy achievement from kindergarten to grade 8: A meta-analysis of classroom and home interventions. *Review of Educational Research*, 83(3), 386-431.
- Kimball, S. M., White, B., Milanowski, A. T., Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in *Washoe County*, *Peabody Journal of Education*, 79(4), 54-78.
- Kodel, C., & Betts, J. (2009). Value-added to what? How a ceiling in the testing instrument influences value-added estimation. *Education Finance and Policy*, 5(1), 54-81. doi: 10.1162/edfp.2009.5.1.5104
- La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The classroom assessment scoring system: Findings from the prekindergarten year. *The Elementary School Journal*, 104(5), 409-426.
- Labree, D. F. (2008). An uneasy relationship: The history of teacher education in the university. In M. Cochran-Smith, S. Feiman-Nemser, D. J. McIntyre, & K. E. Demers (Eds.), *Handbook of Research on Teacher Education* (290 - 306). NY: Routledge.
- Lockwood, J. R., & McCaffrey, D. F., (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics*, 1, 223-252. doi: 10.1214/07-EJS057

- Mashburn, A. J., Pianta, R. C., Hamre, B.K., Downer, J.T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, & D. M., Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*(3), 732-749.
- Means, B., Padilla, C., & Gallagher, L. (2010). *Use of educational data at the local level: From accountability to instructional improvement*. Retrieved from <http://files.eric.ed.gov/fulltext/ED511656.pdf>
- Medley, D. M. (1977). *Teacher competence and teacher effectiveness: A review of process-product research*. Washington, DC: The American Association of Colleges for Teacher Education.
- Merton, R. K. (1968). The Mathew effect in science: The reward and communication systems of science are considered, *Science, 159*(3810), 56-63.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp.13-103). Washington, DC: American Council on Education/Macmillan.
- Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R., (2013). *A composite estimator of effective teaching*. Retrieved from http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf
- Milanowski, A. T. (March 18, 2011). Validity research on teacher evaluation systems based on the Framework for Teaching. Paper presented at the American Education Research Association annual meeting, New Orleans, LA. Retrieved from <http://files.eric.ed.gov/fulltext/ED520519.pdf>
- Monk, D. H. (1992). Education productivity research: An update and assessment of its role in education finance reform. *Educational Evaluation and Policy Analysis, 14*(4), 307-332.
- Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus User's Guide*. Sixth Edition. Los Angeles, CA: Muthén & Muthén.
- National Council on Teacher Quality. (2009). 2009 state teacher policy yearbook: National Summary. Washington, DC: Author.
- Newton, X. A., Darling-Hammond, L., Haertel, E., & Ewart, T. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*(23). Retrieved from <http://epaa.asu.edu/ojs/article/view/810>
- No Child Left Behind Act of 2001, Pub. L. No. 107–110 (2002).

- Nye, B., Konstantopoulos, S., & Hedges, L.V. (2004). How large are teacher effects? *Education Evaluation and Policy Analysis*, 26(3), 237-257.
- Oregon Department of Education, Office of Assessment and Information Services. (2011a). *Oregon statewide assessment: Reading/literature test specifications and blueprints, Grade 3 2011-2012*. Retrieved from http://www.ode.state.or.us/wma/teachlearn/testing/dev/testspecs/asmtrdtestspecsg3_2011-12.pdf
- Oregon Department of Education, Office of Assessment and Information Services. (2011b). *Oregon statewide assessment: Reading/literature test specifications and blueprints, Grade 4 2011-2012*. Retrieved from http://www.ode.state.or.us/wma/teachlearn/testing/dev/testspecs/asmtrdtestspecsg4_2011-12.pdf
- Oregon Department of Education, Office of Assessment and Information Services. (2012a). *Oregon statewide assessment: Reading/literature test specifications and blueprints, Grade 3 2012-2014*. Retrieved from http://www.ode.state.or.us/wma/teachlearn/testing/dev/testspecs/asmtrdtestspecsg3_2012-14.pdf
- Oregon Department of Education, Office of Assessment and Information Services. (2012b). *Oregon statewide assessment: Reading/literature test specifications and blueprints, Grade 4 2012-2014*. Retrieved from http://www.ode.state.or.us/wma/teachlearn/testing/dev/testspecs/asmtrdtestspecsg4_2012-2014.pdf
- Oregon Department of Education. (2007). *2006-2007 Technical Report: Oregon's Statewide Assessment System: Reliability and Validity, Volume 4*. Retrieved from http://www.ode.state.or.us/teachlearn/testing/manuals/2007/asmtechmanualvol4_validity.pdf
- Oregon Department of Education. (2009a). *2007-2008 Technical Report: Oregon's Statewide Assessment System: Score interpretation guide, volume 6*. Retrieved from http://www.ode.state.or.us/teachlearn/testing/manuals/2008/asmtechmanualvol6_interpguide.pdf
- Oregon Department of Education. (2009b). *2009-2010 Technical Report: Oregon's Statewide Assessment System: Annual Report Volume 1*. Retrieved from http://www.ode.state.or.us/wma/teachlearn/testing/manuals/2011/asmtechmanualvol1_annualreport.pdf
- Oregon Department of Education. (2009c). *Achievement Standards 2009-2010*. Retrieved from <http://www.ode.state.or.us/search/page/?id=3176>
- Oregon Department of Education. (2010). *Achievement Standards 2010-2011*. Retrieved from <http://www.ode.state.or.us/search/page/?id=3176>

- Oregon Department of Education. (2011a). *2011-2012 Technical report Oregon's Alternate Assessment System: Peer review documentation: Sections 1-7*. Retrieved from http://www.ode.state.or.us/wma/teachlearn/testing/manuals/2012/asmttechmanualvol7_alternateasmt.pdf
- Oregon Department of Education. (2011b). *Achievement Standards 2011-2012*. Retrieved from <http://www.ode.state.or.us/search/page/?id=3176>
- Oregon Department of Education. (2011c). *OAKS Percentile Conversion Tables 2009-2010*. Retrieved from http://www.ode.state.or.us/wma/data/schoolanddistrict/testresults/2011/asmtconvpctiles_0910.xls
- Oregon Department of Education. (2011d). *OAKS Percentile Conversion Tables 2010-2011*. Retrieved from http://www.ode.state.or.us/wma/data/schoolanddistrict/testresults/2011/asmtconvpctiles_1011.xls
- Oregon Department of Education. (2012a). *Achievement Standards 2012-2013*. Retrieved from <http://www.ode.state.or.us/search/page/?id=3176>
- Oregon Department of Education. (2012b). *OAKS Percentile Conversion Tables 2011-2012*. Retrieved from http://www.ode.state.or.us/wma/data/schoolanddistrict/testresults/2012/asmtconvpctiles_1112.xls
- Oregon Department of Education. (2013a). *OAKS Percentile Conversion Tables 2011-2012*. Retrieved from http://www.ode.state.or.us/wma/data/schoolanddistrict/testresults/2012/asmtconvpctiles_1213.xls
- Oregon Department of Education. (2013b). *Oregon framework for teacher and administrator evaluation and support systems*. Retrieved from <http://www.ode.state.or.us/wma/teachlearn/educatoreffectiveness/oregon-framework--for-eval-and-support-systems.pdf>
- Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications. *Educational Evaluation and Policy Analysis, 30*(2), 111-140.
- Pakarinen, E., Lerkkanen, M., Poikkeus, A., Kiuru, N., Siekkinen, M., Rasku-Puttonen, H., Nurmi, J. (2010). A validation of the Classroom Assessment Scoring System in Finnish kindergartens. *Early Education and Development, 21*(1), 95-124 doi: 10.1080/10409280902858764
- Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193. doi: 10.3102/0002831210362589

- Park, B. J., Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2011). *Diagnostic Efficiency of easyCBMTM Reading: Oregon* (Technical Report No. 1106). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Park, B. J., Irvin, P. S., Alonzo, J., Lai, C. F., & Tindal, G. (2012). *Analyzing the Reliability of the easyCBMTM Reading Comprehension Measures: Grade 4* (Technical Report No. 1203). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Park, B. J., Irvin, P. S., Anderson, D., Alonzo, J., & Tindal, G. (2011). *A Cross-validation of easyCBMTM Reading Cut Scores in Oregon: 2009-2010* (Technical Report No. 1108). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.
- Reynolds, A. (1992). What is competent beginning teaching? A review of the literature. *Review of Educational Research*, 62(1), 1-35. Retrieved from: <http://www.jstor.org/stable/1170714>
- Rivkin, S. G., Hanushek, E. A., Kaine, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417-458.
- Rosenshine, B. (1971). *Teaching behaviours and student achievement*. London: National Foundation for Educational Research.
- Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modeling. *Journal of the Royal Statistical Society*, 43(3), 429-467.
- Sáez, L., Park, B. J., Nese, J. F. T., Jamgochian, E. M., Lai, C. F., Anderson, D., Kamata, A., Alonzo, J., & Tindal, G. (2010). *Technical Adequacy of the easyCBMTM Reading Measures (Grades 3-7), 2009-2010 Version* (Technical Report No. 1005). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education*, 12(3), 247-256.
- Sartain, L., Stoelinga, S. R., Brown, E. R., (2011). Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation. Chicago, IL: Consortium on Chicago School Research at the Unverisit of Chicago Urban Education Institute.

- Sartain, L., Stoelinga, S. R., Brown, E. R., (2009). Evaluation of the Excellence in Teaching Pilot Year 1 Report to the Joyce Foundation. Chicago, IL: Consortium on Chicago School Research at the University of Chicago Urban Education Institute. Retrieved from https://consortium.uchicago.edu/sites/default/files/publications/Joyce_TE_yr1_finaldoc.pdf
- Simon, A., & Boyer, E. (Eds). (1967). *Mirrors for behavior: An anthology of classroom observation instruments*. Philadelphia: Research for Better Schools.
- Stuit, D., Berends, M., Austin, M. J., & Gerdeman, R. D. (2014). *Comparing estimates of teacher value-added based on criterion- and norm-referenced tests* (REL 2014–004). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Midwest. Retrieved from <http://ies.ed.gov/ncee/edlabs>.
- Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. New York: Falmer Press.
- Tindal, G., Nese, J. T., & Alonzo, J. (2009). Criterion-related evidence using easyCBM™ reading measures and student demographics to predict state test performance in grades 3-8 (Technical Report No. 0910). Eugene, OR: Behavioral Research and Teaching, University of Oregon.
- United States. Executive Office of the President (2015). *Every Student Succeeds Act: A progress report on elementary and secondary education*. Retrieved from https://www.whitehouse.gov/sites/whitehouse.gov/files/documents/ESSA_Progress_Report.pdf
- U.S. Department of Education. (2012). *ESEA flexibility*. Retrieved from <http://www.ed.gov/esea/flexibility/documents/esea-flexibility-acc.doc>
- U.S. Department of Education. (n.d.). *Elementary and Secondary Education ESEA Flexibility* (website). Retrieved 21 February, 2016, from <http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html>
- Youngs, P. (2011). *InTASC Model Core Teaching Standards: Research synthesis*. Washington, DC: Council of Chief State School Officers.