

Information Regarding SB 103 (Teacher Evaluations)

4/14/2015

Dawn Zimmer

President of Zimco & K-12 Evaluation Solutions (a division of Zimco)

Based out of Frankenmuth, Michigan

dzimmer@zimco.net

Cell: 989-529-7510

Office: 888-549-4626

Zimco Background

- In business since 1994, providing technology support to K-12 and government entities.
- 2008 - Zimco selected by Saginaw Valley State University to be the exclusive reseller of STAGES, teacher evaluation software developed at SVSU.
 - STAGES allows districts to automate their evaluation framework of *all* staff
- We worked with Michigan Association of School Administrators (MASA) in the development of their School ADvance principal evaluation framework.
- We worked with Silver Strong & Associates (developers of Thoughtful Classroom) to automate their teacher evaluation framework.
- Thoughtful Classroom was one of four evaluation frameworks in the MCEE Pilot. (All districts that piloted STAGES/Thoughtful Classroom renewed their subscriptions for the 13-14 and 14-15 school year.)
- Approximately 200 Michigan Districts, Public School Academies and ISDs currently use STAGES to automate their existing evaluation frameworks.
- Districts in 10 other states uses STAGES.

Establishing Credibility, Dawn Zimmer

- Worked extensively with teacher evaluation since 2008
- Communicate with districts every day regarding their evaluation process
- For the past 5 years – detailed review of how districts completed their evaluations – not in theory, but in reality

Evaluation Terminology

- Frameworks
- Tools
- Evaluation Tools

These 3 terms are used interchangeably throughout the legislation. A framework / tool / evaluation tool describes the evaluation *process*, including the rubric.

- Rubric – the main part of the framework / tool / evaluation tool:

Example of one piece of the RISE Teacher Effectiveness Rubric (out of the state of Indiana):

Competency	Highly Effective (4)	Effective (3)	Improvement Necessary (2)	Ineffective (1)
Engage students in academic content	<p>For Level 4, much of the Level 3 evidence is observed during the year, as well as some of the following:</p> <ul style="list-style-type: none"> - Teacher provides ways to engage with content that significantly promotes student mastery of the objective - Teacher provides differentiated ways of engaging with content specific to individual student needs - Teacher effectively integrates technology as a tool to engage students in academic content 	<ul style="list-style-type: none"> - 3/4 or more of students are actively engaged in content at all times and not off-task - Teacher provides multiple ways, as appropriate, of engaging with content, all aligned to the lesson objective - Ways of engaging with content reflect different learning modalities or intelligences - Teacher adjusts lesson accordingly to accommodate for student prerequisite skills and knowledge so that all students are engaged 	<ul style="list-style-type: none"> - Fewer than 3/4 of students are engaged in content and many are off-task - Teacher may provide multiple ways of engaging students, but perhaps not aligned to lesson objective or mastery of content - Teacher may miss opportunities to provide ways of differentiating content for student engagement - Students may appear to actively listen, but when it comes time for participation are disinterested in engaging 	<ul style="list-style-type: none"> - Fewer than 1/2 of students are engaged in content and many are off-task - Teacher may only provide one way of engaging with content OR teacher may provide multiple ways of engaging students that are not aligned to the lesson objective or mastery of content - Teacher does not differentiate instruction to target different learning modalities - Most students do not have the prerequisite skills necessary to fully engage in content and teacher makes no effort to adjust instruction for these students

Teacher Evaluation Legislation Timeline

- 2009 – 2011 Michigan’s Race to the Top Education Reform Legislation
 - Round I: Public Act 205 of 2009
 - Evaluate every teacher every year
 - Use 4 proficiency levels, not just “Satisfactory” & “Unsatisfactory”
 - Round II: Public Acts 100 – 103 of 2011
 - Tenure Act changes
 - Governor’s Council on Educator Effectiveness to be created, Leftone state evaluation tool, one student growth assessment tool
 - Round III: This is where we are today. Still waiting for “Round III” to be completed.
- Many districts acted on the above legislation (Round I & II) and developed an effective way to evaluate their teachers.
 - Many districts have put time and resources into developing evaluation frameworks.
 - Many districts have refined and improved the process, and it is working.
- Fall 2011 – Five members of the Governor’s Council on Educator Effectiveness are named
 - Council’s name was later changed to the Michigan Council on Educator Effectiveness (MCEE)
- November 1, 2011 – Deadline for districts to notify the MCEE that they wanted to “opt out” of the yet-to-be-identified state teacher and principal evaluation frameworks.
 - Districts had to mail in their board resolutions declaring an exemption.
 - I was aware of only one district that received a response based on this process.
- 2012-2013 school year: Pilot enacted by the MCEE
 - MCEE chooses 4 different teacher evaluation frameworks to pilot
 - 13 districts pilot 1 of 4 frameworks
- July 2013 – MCEE Recommendation
 - Recommend all four frameworks that were piloted – unless the not-yet-completed pilot *report* identifies issues with any of the frameworks.
 - "If final results from the pilot study produce evidence that suggests that any of these tools is less reliable or practical, this information should be taken into account."
- December 2013 – University of Michigan Institute for Social Research Report on the pilot is complete
 - Report is found at www.mcede.org/reports
 - The final recommendation (page 50) does suggest that some tools are less reliable or practical.
- January 15, 2014 – House Bill 5223 introduced regarding Teacher Evaluation
 - “....shall use 1 or more of the following tools:” – with the 4 piloted frameworks listed
- February 12, 2015 – Senate Bill 103 introduced
 - Does not list any specific frameworks

Feedback Regarding Senate Bill 103

- Positive:
 - a. Annual evaluation – with feedback being a focus
 - b. Use evaluations to inform decisions on: promotion, retention, professional development
 - c. Evaluation process has to be consistent within all schools in a district, Public School Academy, or ISD
 - d. If a teacher is rated Highly Effective for 3 years they can be evaluated every other year. (But what should be done on the “off year”? Will districts be penalized if they don’t report an effectiveness rating for that teacher to the state?)
 - e. Districts post information about their evaluation process and framework on their website
 - f. Student Growth percentages for each school year
 - Better (lower percentages) than the original legislation, or other recent bills:
 - 14-15 = “Significant Factor”
 - 15-16 = “Significant Factor”
 - 16-17 = “Significant Factor”
 - 17-18 = 25%
 - 18-19 = 45%

- Negative:
 - a. Student Growth
 - Incorporating student growth district-wide will not be fair without a state assessment tool that tests all subject areas *and* provides timely data on student growth.
 - Typical scenario with 25% Student Growth (in Michigan):
 - 25% Student Growth / 75% Final Rubric Scoring
 - Scenarios with no Student Growth:
 - 100% Rubric Scoring
 - 20% Goal Achievement / 80% Final Rubric Scoring
 - 15% Goal Achievement / 15% Assessment of available data / 70% Final Rubric Scoring
 - b. Requirement of evaluation tool – “Evidence of reliability, validity and efficacy” will probably be perceived as extremely difficult, and perhaps cost prohibitive for districts. This could be very limiting to districts.



**INSTITUTE FOR
SOCIAL RESEARCH**
Social Science in the Public Interest

PROMOTING HIGH QUALITY TEACHER EVALUATIONS IN MICHIGAN:

LESSONS FROM A PILOT OF EDUCATOR EFFECTIVENESS TOOLS

**BRIAN ROWAN
STEPHEN G. SCHILLING
ANGELINE SPAIN
PREM BHANDARI
DANIEL BERGER
JOHN GRAVES**

DECEMBER, 2013

Table of Contents

Table of Contents	i
Acknowledgements	ii
Abstract	iii
Chapter 1: Introduction	1
The Pilot of Educator Effectiveness	1
Tools Goals of the Pilot Research	2
Structure of this Report	2
Chapter 2: Key Findings on Pilot Activities	3
Key Findings on District Policy Development	3
Key Findings on Principal Workload	3
Key Findings on Classroom Observation Tools: Vendor Training	5
Key Findings on Classroom Observation Tools: Fidelity of Use	6
Key Findings on Student Growth Tools	8
Key Findings on Final Evaluation Ratings	10
Key Findings on Principals' and Teachers' Views of the Evaluation	12
Chapter 3: Improving the Teacher Observation Process	15
Improving the Use of Classroom Observation Data	15
Chapter 4: Are Value-Added Models an Option for Michigan?	22
VAM Pilot Data and Procedures	22
Statistical Models Used by VAM Vendors	23
Data Processing Issues Prior to VAM Analyses	24
The Pilot Roster Project	27
Results of VAM Analyses	28
Chapter 5: Setting Standards for Teacher Evaluation	33
Two Approaches to Performance Rating	33
Estimated Levels of Performance	34
Imprecision in Teacher Performance Estimates	37
Taking Imprecision into Account in Making "High Stakes" Personnel Decisions	39
The Problem of Joint Classification	41
Classification Without Confidence Intervals: Simple Ranking Systems	45
Chapter 6: Action Steps to Improve Teacher Evaluations in Michigan	47
Improving District Policy and Procedure Manuals	47
Improving Classroom Observation Procedures	47
Improving Measurement of Student Growth	48
Assignment of Effectiveness Ratings to Teachers	49
Timing of Improvement Steps	50
Costs	52

Acknowledgments

Work on this report was conducted for the Michigan Council for Educator Effectiveness by the University of Michigan's Institute for Social Research under an intergovernmental services agreement between the State of Michigan's Department of Technology, Management, and Budget and the Regents of the University of Michigan.

The authors want to express deep appreciation to staff of the Survey Operations Unit of the Survey Research Center of the University of Michigan's Institute for Social Research. Without this unit's operational support, this project could not have been conducted. Stephanie Chardoul and Meredith A. House, senior staff of the unit, were especially critical to the success of the project. Catherine Thibault, Assistant Director of the Survey Research Center also provided key support for the project.

The authors also want to thank the students, faculty, administrators, and staff of the districts that participated in the pilot of educator effectiveness tools for cooperation in various research and development activities. We also would like to thank the many state employees without whose work the project could not have been completed. Finally, we thank the many vendors who provided support, training, data, and data analyses for the project.

Special thanks are due to the members and staff of the Michigan Council for Educator Effectiveness for important assistance and feedback at all stages of the work. Special thanks also are due to Dennis Schornack, Senior Strategy Advisor in the Executive Office of the Governor, for managing the intergovernmental services agreement and for showing keen interest in the project. Thanks finally to Deborah Ball, Dean of the School of Education and Chair of the Michigan Council for Educator Effectiveness for support and advice over the course of the project.

Abstract

This is a preliminary report by the University of Michigan's Institute for Social Research (ISR) on the pilot of educator effectiveness tools commissioned by the Michigan Council for Educator Effectiveness and conducted during the 2012-2013 school year in 13 public school districts in Michigan. Chapter 1 of this report briefly introduces the main goals of the pilot initiative and describes the research and development activities conducted in schools as part of the pilot. Chapter 2 describes ISR's main findings about how pilot activities were carried out in local schools. The next three chapters discuss some approaches to improving teacher evaluation practices in local schools. Chapter 3 discusses various approaches to improving classroom observations conducted as part of the teacher evaluation process; Chapter 4 discusses the extent to which value-added measures of teaching effectiveness might represent a viable approach to measuring teachers' contributions to students' academic growth in the teacher evaluation process; and Chapter 5 discusses some approaches to assigning final effectiveness ratings to teachers as part of the evaluation process. Chapter 6 describes some action steps that might be taken by state and local education agencies in Michigan to improve teacher evaluation activities in local schools.

The reader will note that the absence of an executive summary of this report. Instead, the text of the report has been organized to help any reader obtain an overview of the report's central details. *Important points in the report are highlighted in bold text with italics, and many tables are provided to give the reader a good sense of the data on which the report's findings are based.* Therefore, a quick scan of the highlighted text and tables should give any reader an initial sense of the report's major findings. The reader is also advised that several of the analyses reported here are preliminary and subject to change with additional analyses. While such changes are unlikely to alter the main conclusions of the report, the reader is nevertheless advised that the data collected and analyzed during the pilot project were complex and that a final technical report on the project will not be released until March 31, 2014.

This page intentionally left blank

Chapter 1: Introduction

Michigan’s Public Act (PA) 102 of 2011 fundamentally redefined the nature of teacher evaluation in the state’s public schools. The new law required public education agencies to evaluate teachers using multiple criteria—including classroom observations and evidence of student learning—and to assign a final effectiveness rating to teachers as a result of an annual evaluation process. The law also established the Michigan Council for Educator Effectiveness (MCEE) as a temporary state commission to advise the Governor, State Board of Education, and State Legislature on a number of issues related to the implementation of PA 102 of 2011. To inform the Council’s deliberations, the University of Michigan’s Institute for Social Research (ISR) was engaged to conduct a pilot of educator effectiveness tools. The pilot was funded through an intergovernmental services agreement between the State of Michigan’s Department of Technology, Management, and Budget and the Regents of the University of Michigan.

The Pilot of Educator Effectiveness Tools

The pilot of educator effectiveness tools was conducted in 13 Michigan school districts during the 2012-2013 school year. During the pilot year, participating school districts: (a) piloted one of four classroom observation tools being considered for possible adoption as the state tool for classroom observations in Michigan; (b) piloted a set of student assessments that closely resembled (but were not identical to) the student growth tools the Council recommended in its June, 2013 final report; and (c) allowed researchers to administer surveys to principals and teachers, to conduct classroom observations alongside district personnel, and to collect documents related to the conduct of teacher evaluations.

At a Glance: The Pilot Project	
•	13 school districts in lower Michigan participated:
○	Big Rapids
○	Cassopolis
○	Clare
○	Farmington
○	Garden City
○	Gibraltar
○	Harper Creek
○	Leslie
○	Marshall
○	Montrose
○	Mt. Morris
○	North Branch
○	Port Huron
•	Four observation tools were piloted:
○	Danielson’s Framework for Teaching
○	5 Dimensions of Teaching & Learning
○	Marzano Teacher Evaluation Model
○	Thoughtful Classroom Framework
•	Four student growth tools were piloted:
○	NWEA MAP Series (Grades K-6)
○	ACT Explore (Grades 7-8)
○	ACT Plan (Grades 9-10)
○	ACT (Grade 12)
•	Seven research activities were conducted:
○	Interviews with district administrators
○	Evaluation policy documents collected
○	Teacher survey administered (n = 1182)
○	Principal survey administered (n = 99)
○	Independent classroom observations
○	Value-added scores calculated
○	Final effectiveness ratings collected

Goals of the Pilot Research

Research on the pilot initiative conducted by ISR had the following goals:

- to gather a wide variety of interview, survey, and observational data on the ways teacher evaluations were conducted in pilot schools;
- to examine various approaches to improving the teacher evaluation process by modifying classroom observation procedures, devising rigorous, fair, and useful procedures for measuring student growth and estimating teachers' contributions to their students' achievement, and developing rigorous and fair approaches to assigning teachers to effectiveness ratings; and
- to solicit the opinions of teacher and administrators in pilot schools about the teacher evaluation process and how it might be improved.

Structure of this Report

These issues are discussed in five subsequent chapters. Chapter 2 of this report describes key findings about how teacher evaluations were conducted in schools during the pilot year. Chapter 3 closely examines the data from classroom observations conducted during the pilot year and explores some approaches to improving this process. Chapter 4 examines the value-added statistical modelling conducted by vendors and discusses the steps needed for Michigan to use "value-added" modeling (VAM) in teacher evaluations. Chapter 5 discusses various approaches to assigning teachers to effectiveness ratings using data from classroom observations and value-added measures. Chapter 6 suggests some action steps for the development and implementation of high quality teacher evaluations in Michigan.

Chapter 2: Key Findings on Pilot Activities

This chapter reports some key findings from the pilot initiative. The results are reported in the following areas: (a) district policy development; (b) the workload of educators who conducted teacher evaluations in pilot schools; (c) how classroom observations were conducted in pilot schools; (d) how student growth was measured for the purposes of teacher evaluation in pilot schools; (e) how data from classroom observations and measures of student growth were combined in order to assign final evaluation ratings to teachers; and (f) the reports of teachers and principals on the quality and consequences of teacher evaluation practices enacted during the pilot year.

Key Findings on District Policy Development

When the pilot of educator effectiveness tools was launched in the summer of 2012, many pilot districts were in the beginning stages of implementing PA 102 of 2011. The law required certain evaluation activities to be implemented in schools but still gave districts wide discretion to develop and conduct teacher evaluations according to local preferences. Like the law, ISR imposed few requirements on districts. ISR asked only that participating districts use their assigned teacher observation tool according to vendor guidelines, implement the testing regimes associated with piloted student growth tools, and allow ISR to conduct research activities in their schools. Apart from these requirements, districts were responsible for developing teacher evaluation policies and practices that complied with the provisions of PA 102 of 2011 according to local preferences.

At the beginning of the pilot year, most pilot districts lacked fully-developed policies to guide the teacher evaluation practices required under PA 102 of 2011. As a result, during the pilot year, participating districts worked diligently to develop such policies. Two key findings emerged from ISR's study of district policy development:

Policies were developed by teams. All districts in the pilot used a team approach to developing new district

policies about teacher evaluation. A few of these districts used teams composed *only* of central office administrators and principals. However, most districts also included teachers in the planning process. In these latter districts, however, the size of the planning teams varied, as did the role of teachers. In one large district, planning was done through a number of task forces, while in most other districts, planning teams were smaller.

Procedural documents were often under-developed.

At the beginning of the pilot year, districts generally lacked well-structured and detailed documents describing policies and procedures for conducting the new teacher evaluations. By the end of the year, however, most districts had produced such documents. Still, the detail included in such documents varied considerably. Three pilot districts produced thorough and well-articulated statements about teacher evaluation policies and procedures. These documents described the classroom observation process, how student growth would be measured for the purposes of teacher evaluation, and the criteria and procedures that would be used to assign final effectiveness ratings to teachers. However, many other pilot districts had only fragmentary documentation of their evaluation procedures and were just beginning to weave these fragments into a well-designed manual of policies and procedures.

Key Findings on Principal Workload

In all districts, teacher evaluations were largely the responsibility of administrators, although teachers played a critical role in the generation of student growth data.

Responsibility for completing various tasks was distributed across principals, teachers, and central administrators. All education professionals in a district were involved in the work of evaluating teachers, but the work that specific groups undertook varied by role. In all districts, administrators (not teachers) were trained and made responsible for

conducting classroom observations, and in these districts, classroom observations for the purposes of teacher evaluations were conducted mostly by principals, occasionally by central administrators, and *never* by teachers. In contrast, the measurement of student growth was typically organized as a shared responsibility in which the data to be used for the “student growth” portion of teacher evaluations was developed *jointly* by teachers and principals. In this process, the tools that would be used to measure student learning were often chosen by teachers (in consultation with their principals) from a list of approved assessment data. Finally, the compilation and analysis of observation and student growth data, and the assignment of final effectiveness ratings to teachers, was typically given over to principals. In all but two districts, however, the assignment of these final ratings was governed by a formula that specified the “weight” to be given to district-specified performance criteria and “cut points” for the assignment of teachers to final effectiveness ratings based on summary rating scores.

Several key findings about workloads emerged from ISR’s research:

Evaluation entailed completing numerous tasks. In all pilot schools, the evaluation process included numerous steps. For example, the median teacher in a pilot school was observed on 4 occasions (but not always for a “full” class period); 99% of principals also reported conducting pre- and/or post-observation conferences with teachers, and 82% of teachers reported participating in such conferences. In addition, 99% of principals and 90% of teachers reported assembling and using student growth measures in their annual evaluation. Finally, nearly all teachers reported receiving an annual effectiveness rating, and about half of all teachers reported receiving a mid-year or year-end evaluation report (not required in AY 2012-2013). Evaluation therefore included: classroom observations, conferencing, measuring student growth, and reporting.

At a Glance: Principals’ Evaluation Workload

Evaluation Workload	
Median Number of Teachers Evaluated	25
Median Number of Observation per Teacher	4
% Principals Reporting Conferences	99%
% Teachers Reporting Conferences	82%
% Principals Using Growth Tools in Evaluations	100%
% Teachers Receiving Mid-Year Report	37%
% Teachers Receiving End-of-Year Report	47%
% Teachers Assigned Effectiveness Rating	97%
Time Spent on Teacher Evaluations	
Median Hours Spent on Training	24
Median Hours Observing Probationary Teacher	4.25
Median Hours Observing Tenured Teacher	4.25
Median Hours Rating/Report Probationary Teacher	5.75
Median Hours Rating/Report Tenured Teacher	3.75
Median Days/Year Spent on Teacher Evaluation	31

Principals had very large spans of control. An important feature of the teacher evaluation process was that the median principal in pilot schools was responsible for the annual evaluation of 25 teachers (23 tenured teachers and two probationary teachers). In most organizations, the span of control (i.e., ratio of supervisors to employees is 1 to 7, so principals had a very large span of control).¹

The teacher evaluation process consumed a great deal of principals’ time. Because the annual evaluation process required principals to complete multiple steps for a large number of teachers, it consumed a large amount of time. In the pilot year, the median principal reported spending about 248 hours (or 31 full work days) on teacher evaluation activities. There is reason to expect, however, that this expenditure of time will decline over the next several years. In the pilot year, the evaluation process had to be completed for every teacher. In future years, however, at least some percentage of teachers will have been rated as “highly effective” in consecutive years, and this will reduce principals’ evaluation workloads somewhat in out years.

¹ Brian Rowan and Stephen W. Raudenbush, “Teacher Evaluation in American Schools”, in *Handbook of Research on Teaching*, American Educational Research Association, in press.

Key Findings on Classroom Observation Tools: Vendor Training

We turn now to a key component of principals' evaluation workload: the teacher observation process. The reader will recall that PA 102 of 2011 requires public education agencies in Michigan to conduct classroom observations as part of the teacher evaluation process, and that if a state tool is used in teacher evaluations, PA 102 further requires that districts conduct observations "in a manner consistent" with guidelines of the observation tool vendor.

MCEE decided to field test four classroom observation tools during the pilot. These were: Danielson's Framework for Teaching (FFT), Five Dimensions of Teaching and Learning (5D), the Marzano teacher effectiveness model (M), and the Thoughtful Classroom teacher evaluation model (TC).

MCEE contracted with each of these tool vendors to provide four days of training in tool use to participating school districts during late summer of 2012. All tool vendors provided roughly similar training. Over four days, tool vendors: (1) explained the conceptual framework underlying their observation tool; (2) discussed the tool's scoring rubric; (3) helped trainees understand the evidence that should be used to assign evaluative scores on classroom observations; (4) discussed how to conduct observation conferences with teachers; and (5) explained how to use the software associated with each observation tool (to record and score notes, communicate about observation data with teachers, schedule observations and conferences, and perform analytic tasks [like summarizing observation scores]).

Using survey data from principals, ISR researchers came to several conclusions about the trainings:

Vendor training was only partially successful. On the principal survey, ISR researchers asked principals about the quality of this initial training. On this survey, a large majority of principals agreed or strongly agreed that trainers did a good job explaining the underlying conceptual framework of the observation protocol. However, principals were less inclined to agree or strongly agree that trainers did a good job in other areas of the training, including: explaining the scoring rubric, explaining the evidence that should be used in scoring; explaining how to conduct teacher

At a Glance: Observation Tool Training	
Training for Use of Observation Tool	
Percent of Principals Reporting:	
• No training	7%
• Initial training only	7%
• Initial training plus team meetings in district	20%
• Initial training, team meetings, follow-up training	34%
• Initial training, team meetings, follow-up training, individual support	20%
• Missing information	12%
Quality of Initial Vendor Training	
Percent of Principals Who Agreed or Strongly Agreed that Vendors Did a Good Job Explaining:	
• Protocol framework	80%
• Scoring rubric	66%
• Evidence to be used in scoring	60%
• How to conduct conferences	40%
• How to use vendor software	33%
Percent of Principals Who Agreed or Strongly Agreed that, at the End of Initial Training, they were :	
• Confident to conduct teacher observations	60%
• Prepared to conduct teacher conferences	52%
• Confident my scoring was in line with others	39%

conferences; and explaining how to use the software associated with the observation instrument.

At the end of training, many principals lacked confidence in their ability to use classroom observation tools with fidelity. Sixty percent of principals reported that, at the end of initial training, they felt confident to conduct teacher observations, and 52% felt confident to conduct pre- and/or post-observation conferences with teachers. However, only 39% were confident that their scoring of lessons was in line with the scoring of others.

Many principals engaged in additional training. Given these findings, it is important to explore what (if any) additional training principals received in observation procedures. Overall, patterns of training varied among principals. About 7% of principals in the pilot study reported receiving *no* training in the use of an observation tool, whereas the remaining 93% attended initial trainings. But patterns of training departed from there. About 7% of principals reported receiving only the initial vendor training; 20% reported receiving initial training and then discussing how to use the assigned observation tool in district meet-

ings; another 34% added some follow-up training to these experiences; and another 20% of principals had all of these previous experiences plus some individual follow-up training.

**Key Findings on Classroom Observation Tools:
Fidelity of Use**

Given the variation in initial and on-going training, an important question is how well principals were prepared to conduct classroom observations. To investigate this issue, ISR researchers examined vendor databases, which included all scores from principals who conducted classroom observations when they used one of the vendor’s tools. In addition, ISR hired a cadre of former educators to conduct independent classroom observations, sometimes alongside principals and sometimes alongside each other. Records from these observations also were available in vendor databases.

Principals spread observations across the year. The analysis of vendor databases showed that principals tended to spread their classroom observation workload equally across the school year. In addition, when teachers were observed on more than one occasion, the elapsed time between consecutive observations was typically from 10-90 days. Both of these practices make for good sampling of teaching practice, avoiding the observation of a given teacher within a single period of the school year and, instead, sampling across school days to capture the variety of lessons a teacher might conduct.

There was low fidelity in item scoring. In other areas of observation practice, however, principals did not perform as well. To begin, the observation tools in use in the pilot varied as to whether it was mandatory for items on the protocol to be scored on every observation occasion or whether items were to be scored only when lesson activities were judged as relevant. The 5D tool and the FFT tool assumed that all items could be scored across any type of lesson. Thus, all items were mandatory. On the other hand, the Marzano (M) and Thoughtful Classroom (TC) tools assumed that at least some items could be scored only when certain lesson activities were being observed—although TC assumed that four items (measuring what it called the “four corners” of instruction) would always be scored.

At a Glance: Fidelity of Observation Tool Use				
	5D	FFT	M ¹	TC ²
Item-Level Scoring				
Median % of times item was scored by:				
• Principal (mandatory item)	73%	90%	NA	77%
• ISR (mandatory item)	94%	100%	NA	100%
• Principal (any item)	73%	90%	9%	25%
• ISR (any item)	94%	100%	52%	51%
Median % of time raters agreed any item should be scored:				
• Principal with ISR observer	77%	97%	10%	38%
• ISR observer with ISR obs.	95%	100%	72%	46%
Median % exact agreement on score (when items are scored)				
• Principal with ISR obs.	51%	50%	40%	42%
• ISR observer with ISR obs.	50%	46%	56%	61%
Median ICC for scored items				
• Principal – ISR pair	.16	.07	.08	.09
• ISR observer pair	.31	.32	.43	.56
Scale Scores				
Estimated Correlation of Scale Scores				
• Prin.-ISR observer pair	.22	.60	NA ³	.50
Percentage of variance in scale score due to rater effects⁴				
• Principal observations	11%	15%	NA ³	38%
Average difference in scale scores (leniency)				
• Principal–ISR observer	.28	.40	NA ³	.58
¹ Marzano data do not include one district that was piloting only one section of the protocol. ² Tool includes both mandatory items (scored on all observation occasions) and non-mandatory items (scored only when appropriate to lesson activities). ³ ISR researchers were unable to calculate scale scores for Marzano data because items were scored on too few occasions by principals. ⁴ This is the total variance in scale scores accounted for by rater fixed effects. Given the structure of the data, the model confounds rater and school/district effects and thus should be viewed with caution.				

When ISR researchers examined the observation data, a striking pattern emerged. Many principals failed to score items during an observation—even when vendors advised them that scoring of an item was mandatory (this pattern was least prevalent for the FFT protocol). Moreover, the observation data showed that non-mandatory items on both the TC and Marzano protocols were scored at very low frequencies. Importantly, this pattern of (non)scoring was less prevalent among ISR observers using these same tools—probably because ISR observers received

six additional, one-hour trainings from observation tool vendors in order to improve their scoring. Clearly, in the absence of such additional training, many principals did *not* score observation protocols in the “manner prescribed” by vendors (as required by PA 102 of 2011).

Inter-rater reliability was low. Another key finding from the analysis of vendor databases was that when two raters scored the same lesson, there were seldom high levels of agreement about whether an item should be scored, or if an item was scored, about the actual score assigned to teachers.

At the item level, agreement among ISR observers and principals about when to score mandatory and non-mandatory items was low (the exception was FFT, where the median rate of agreement between ISR observers and principals about when to score an item was 97%). On the two other tools with mandatory items (5D and TC) the median agreement rate was about 75% for mandatory items. There was even less agreement among raters about when to score non-mandatory items. On these items, the median rate of agreement among ISR observers and principals about when to score an item varied from 40-50%, with about 10% higher agreement rates among ISR observers. Clearly, different raters had different opinions about when the characteristics of a lesson warranted scoring of non-mandatory items.

When items *were* scored, there were low rates of agreement about the exact score to be assigned to a teacher for the lesson being observed. Once again, the median rate of agreement about the score to be given to a teacher on an item during the same lesson varied from 40-50% exact agreement among principals and ISR observers, and from 50%-60% among ISR observers. Apparently, the additional training given to ISR observers increased agreement rates by about 10%.²

² A statistical measure of inter-rater agreement is the “intra-class correlation” (or ICC), which in the present case is a one-way ANOVA model, with rater scores nested within observation occasions. The ICC can be interpreted as a classical reliability coefficient that varies from a low of 0 to a high of 1. In the pilot observation data, item-level ICCs were quite low, showing (once again) that there was low inter-rater agreement in item scoring. The ICCs, however, are lower for principal-ISR observer pairs than for ISR observer pairs. One reason for this finding is that when ISR observers were paired together, their item scores had a wider spread among teachers than was found for principal-ISR observer pairs. Also, there was more agreement on item scores among ISR pairs than among principal-ISR observer pairs.

Low inter-rater reliability at the item level also carries through to measurement at the scale score level. For example, ISR researchers combined item scores for teachers into multi-item summary scores (using a one-parameter, multi-level, IRT measurement model). When summary scales were created by this process using data from the vendor databases, the percentage of total variance in scale scores that was accounted for by “rater effects” varied from a low of around 11% for 5D to a high of 38% for TC. These rater effects are substantial. For example, if the same teacher was observed by two administrators a standard deviation apart in the distribution of rater effects, the scores received by that teacher could differ by as much as .40 of a standard deviation across raters for TC, and around .20 of a standard deviation across these raters for FFT and 5D. Another way to see the effects of rater error on teacher scores is to note that the correlation among scale scores assigned to a given teacher by two different raters on a single occasion (one an administrator, the other an ISR observer) was quite low—ranging from .22 for 5D to .60 for FFT.^{3,4}

Principals tended to be more lenient in their scoring than ISR observers. Finally, disagreements among principals and ISR observers tended to run in a certain direction – with principals scoring items for the same lesson higher than ISR observers. For example, across principal-ISR observer pairs scoring the same lesson, principals tended to score a teacher .28 points higher than ISR observers on the 5D tool, about .40 points higher on the FFT tool, and .58 points higher on the TC tool. Moreover, the distribution of scores assigned to teachers by principals tended to be skewed positively (more scores above the mean than below). The findings that a teacher’s immediate supervisor is more lenient in scoring than an independent observer and that supervisor scores are positively skewed are quite common, not only in research on

³ Readers with a technical background will want to know that the “rater effects” estimated here come from a model that had fixed effects for raters and random effects for items, occasions and teachers. Given the structure of the data, these rater effects are conflated with “school effects” and are far from an ideal estimate of rater errors in measurement.

⁴ The reader will note that ISR researchers did not calculate scale scores for the Marzano tool. That is because there was simply too much missing item data, a result of the fact that principals rarely coded any single item in the Marzano tool. As we discuss at a later point in this report, this is a major drawback of the Marzano tool in teacher evaluations.

teacher evaluation, but also in research on personnel evaluation more generally.⁵

Key Findings on Student Growth Tools

In addition to requiring schools to conduct classroom observations as a component of teacher evaluations, PA 102 of 2011 requires that public education agencies in Michigan: (a) establish clear approaches to measuring student growth as measured by national, state, or local assessments (or other objective criteria); (b) provide teachers and school administrators with relevant data on student growth; and (c) use student growth as a “significant factor” in evaluating a teacher’s job performance. Interviews conducted by ISR researchers with district officials suggested that districts struggled to fulfill these requirements for a number of reasons. Key findings were:

Michigan’s state testing system does not provide sufficient, timely data for use of MEAP, MME, or other state tests in teacher evaluations. One problem with Michigan’s current state testing system is that MEAP tests are given in October and, as a result, can only be used to examine student growth over a Fall-to-Fall period that is one year behind the current school year. For this reason, student growth measures based on MEAP scores will also be at least one year behind the current annual teacher evaluation cycle. Another problem is that MEAP tests do not produce vertically equated scale scores. As a result, simple gain scores in MEAP cannot be calculated and analysts must instead resort to estimation of more complex statistical models of learning gains (that involve examining the difference between a student’s current Fall test score and that student’s predicted score, where the current score is predicted from prior test scores and, perhaps, social background factors). Such models *can* be estimated with MEAP data (but, again, only for the year prior to the current teacher evaluation cycle). MME data present a different set of problems. It is theoretically possible to calculate a gain score using the MME (by using its ACT component), but in the current Michigan testing system, there is no available pre-test since an ACT test on the same scale is not administered in Spring of 10th grade. A final problem with the current state testing system is that only about 33% of teachers in the state teach grade levels and subject

⁵ For documentation of these patterns in research, see Rowan and Raudenbush, *op cit*.

At a Glance: Use of Student Growth Tools			
Type of Growth Tool Used in Teacher Evaluations ¹			
	Elem. Schools	Middle Schools	High Schools
% schools using teacher made tests	54%	69%	64%
% schools using locally-developed common tests	25%	23%	22%
% of schools using MEAP/MME/ or other state assessment	19%	30%	18%
% of schools using other commercial standardized test	44%	23%	15%
Metric Used for Measuring Student Growth ¹			
% using pre/post-test score	80%	87%	85%
% using students meeting learning objectives	60%	13%	27%
% using months of growth	43%	33%	12%
% using change in MEAP proficiency	60%	13%	27%
% using regression-adjusted score (i.e., value-added model)	0%	0%	0%
¹ Includes only principals who reported that annual teacher evaluations included evidence of student growth.			

areas that can be included in value-added modeling. Thus, for the most part, the state testing system cannot provide sufficient or timely data for the measurement of student growth in teacher evaluations.

The most commonly used measures of student growth in pilot schools came from teacher-made and other locally-developed tests. Because of the problems associated with using state test data in teacher evaluation, it is not surprising that other assessments were used. For example, 54% of elementary schools, 69% of middle schools, and 64% of high schools used teacher-made tests as measures of student growth in teacher evaluations, and 25% of elementary schools, 23% of middle schools, and 22% of high schools used locally-developed, common (benchmark or end-of-course) assessments in teacher evaluations. The Michigan Council for Educator Effectiveness argued in its final report that such tests lack desirable psychometric properties and are thus not the best measures for use in teacher evaluations.

A substantial percentage of elementary schools also used commercially-produced standardized tests as student growth tools, but this was not the case in middle and high schools. The use of standardized tests in elementary schools is perhaps best explained by the fact that many elementary schools in the pilot were already using standardized tests for monitoring

instruction, instructional grouping, and referral to compensatory and special education programming. Among the tests commonly used for these purposes were DIBELS (for early grades reading assessment) and AIMS Web and Star (for reading and mathematics). However, while these and other standardized tests can be used fairly easily to measure student growth within the year of an annual evaluation cycle (if administered on two or more occasions), they are not necessarily well-aligned to state curricular standards.

The student growth tools provided to schools as part of the pilot were not widely used as measures of student growth in teacher evaluations. A basic goal of the pilot program was to provide schools with better tools for evaluating student growth. For this reason, elementary schools were provided with paid licenses to use Northwest Evaluation Associates Measures of Academic Progress (NWEA MAP) in grades K-6 and were asked to administer various ACT tests at grades 7-12. However, only 20% of elementary school teachers in the pilot reported using NWEA MAP scores as a measure of student growth in their annual evaluation, and even fewer middle and high school teachers reported using one of the ACT tests paid for by the pilot as a growth measure in their teacher evaluation. The lack of use of ACT tests is understandable. To measure student growth with these tools required school systems to use a “value-added” approach to measuring student growth. On the other hand, simple gain scores or other growth measures could be calculated easily from NWEA MAP scores, so it is surprising that teachers (or administrators) did not use these measures more frequently.

The most common approaches to measuring student growth in teacher evaluations were not sophisticated from a psychometric standpoint and could have been applied inappropriately in local settings. The most common approach to measuring student growth was to develop a simple gain score on a locally-developed test. If these tests were “equated” (i.e., were tests of the same content), such measures make sense. But ISR researchers did not have the resources to examine the myriad of local tests used in teacher evaluations or the ways in which gain scores were calculated. Thus, the extent to which these scores were actually meaningful measures of growth is somewhat uncertain.

The other common ways of calculating “growth” from test scores in pilot schools seem even more problematic. Consider, as one example, the finding that 60% of elementary schools used “percent of students meeting learning objectives” as a measure of student growth in teacher evaluations. *Prima facie*, this is not a measure of student growth, and indeed only becomes a measure of growth if one calculates changes in the percentages of students meeting particular learning objectives. The extent to which this latter calculation was employed in teacher evaluations remains uncertain.

Similarly, 60% of elementary schools reported using “changes in MEAP proficiency” as a method of measuring learning gains. This method—described in various Michigan Department of Education documents—is a feasible way of measuring student growth using consecutive MEAP test data on students, but ISR researchers were unable to tell from surveys or district documents whether the data provided to teachers were appropriate to this task. In particular, if a teacher examined the difference in proficiency rates that resulted between last October’s MEAP administration and the current October administration of MEAP for her current students, that teacher would, in fact, be measuring changes in MEAP proficiency that occurred during a time when her students were mostly under the supervision of *other* teachers. To use MEAP scores appropriately as measures of student learning requires quite a bit of data processing and, as discussed above, currently involves measuring student growth for the group of pupils the teacher taught in a prior year.

The combination of multiple tests with multiple methods of assessing growth produced a staggeringly complex array of non-uniform measures of student growth in annual teacher evaluations. In particular, when ISR researchers examined the data closely, they found that 266 distinct combinations of assessments were used by pilot school teachers in annual evaluations. Moreover, the data showed the tests used to demonstrate student growth almost always varied among teachers at the same grade, in the same school, teaching the same subjects. Such diversity in assessment makes the application of a “uniform” standard for judging teachers’ success in promoting students’ academic growth nearly impossible.

Key Findings on Final Evaluation Ratings

The final step in the teacher evaluation process involved classification of teachers into one of four effectiveness ratings defined in section 2(e) of PA 102 of 2011 (i.e., ineffective, minimally effective, effective, or highly effective). The law allows school systems to use multiple performance criteria to assign these ratings, including student growth, pedagogical skill, classroom management, attendance and disciplinary record, and other accomplishments. Moreover, in AY 2012-2013 (the pilot year), the law did not specify the percentage “weight” to be given to student growth or other performance criteria in final evaluations.

Districts in the pilot study used a variety of performance criteria and gave different weights to the same performance criteria as they assigned final effectiveness ratings to teachers. In general, three classes of performance criteria were used in performance ratings: student growth, classroom instructional practice (including classroom management), and other professional criteria. We have already seen that many different assessments were used to measure student growth, not only across districts, and but also among schools within the same district. However, measures of classroom practice were generally standardized within a district and usually based on the classroom observation tool in use in the district. Measures of other professional criteria also were standardized within districts, but across districts, the measures came from one of two sources—either data from the classroom observation rubrics (which sometimes measured planning and preparation for teaching and professional behaviors) and/or from locally-determined data on a teachers’ professional contributions.

Most districts used a simple, additive formula to arrive at a teacher’s total performance score and then established “cut scores” on this metric to assign teachers to the various effectiveness ratings required by law. The procedures by which a teacher received a final effectiveness rating generally proceeded through a series of steps. First, principals took the various pieces of data (observation data, test score data, other data) collected as part of the evaluation process and assigned scores to each piece of data using “scoring rubrics” standardized by the district. Thus, a principal would look at data on student growth from a teacher and then use a scoring rubric to assign an overall score for this component of the evaluation. This would then be repeated for all other

At a Glance: District Effectiveness Ratings¹

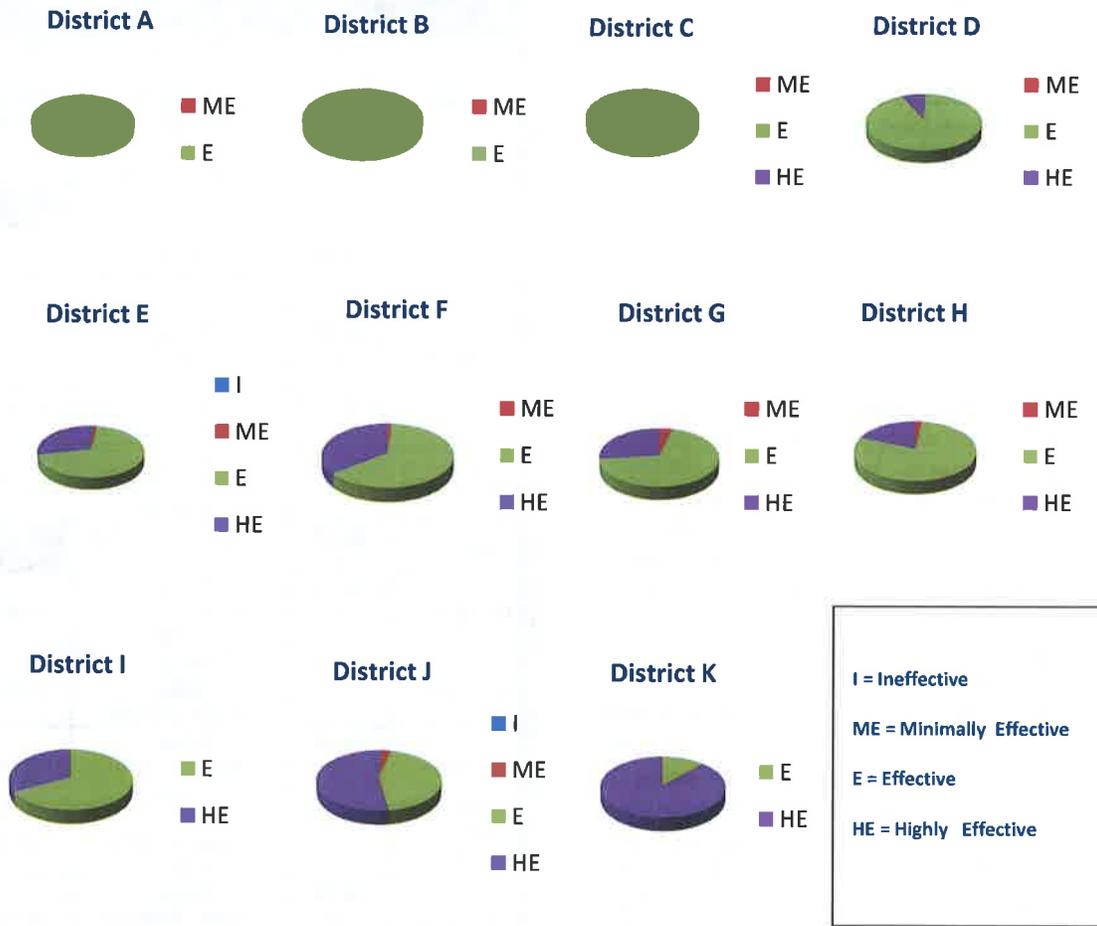
District	Criteria ²	Weighting	Decision Logic
Big Rapids	SG, CP, O	Judgment (no formula)	
Cassopolis	SG, CP	Rating = .75(CP) + .25 (SG)	Cut scores
Clare	SG, CP, O	CBA (no ratings)	
Farmington	SG, CP, O	Rating = .375(CP) + .375(O) + .25(SG)	Cut scores
Garden City	SG, CP, O	Rating = .25(SG)+.345(CP) +.405(O)	Cut scores
Gibraltar	CP, O	Rating = .75(CP)+.25(O)	Cut scores
Harper Creek	SG, CP, O	Rating = .15 (SG) + .55(CP) + .30(O)	Cut scores
Leslie	SG, CP, O	No formula provided	
Marshall	SG, CP, O	Rating = .20(SG) + .40(CP) + .40(O)	Cut scores
Montrose	SG, CP, O	Rating = .12(SG) +.64(CP)+.24(O)	Cut scores
Mt. Morris	SG, CP, O	Rating = .20(SG) + .67(CP) +.13(O)	Cut scores
North Branch	SG, CP, O	Rating = .25(SG)+.50(CP) +.25(O)	Cut scores
Port Huron	SG, CP, O	Rating = .25(SG)+.345(CP)+.405(O)	Cut scores

¹ Data are from district documents sent to ISR. Not all districts replied to the request for documentation, and documentation was uneven.

² SG = student growth, CP = classroom practice, O= other criteria (such as attendance, professional accomplishments, planning and preparation).

pieces of data. Once these component scores were assigned, the principal would then use a standard (district-wide) formula to compute an overall score. In all of the pilot districts, this overall score was based on what was essentially a simple, additive formula that assigned weights to scores on a given performance measure. Once total scores were computed, certain “cut points” were established to map total performance scores onto final effectiveness ratings.

**At A Glance:
Percentage of Teachers Assigned to Effectiveness Ratings in Pilot Districts**



Because performance criteria and “cut scores” varied across districts, the effectiveness ratings established by section 2(e) of PA 102 of 2011 did not have a consistent meaning across districts. Put differently, a teacher with the same scores on measures of student growth and classroom observations would not necessarily receive the same final effectiveness rating. Instead, two otherwise similar teachers would receive a final effectiveness rating that was based on the weights assigned to their scores and the cut points established by their districts for assigning particular effectiveness ratings. Thus, each district, in effect, had its own standard.

Partly because each district had its own standards, the percentage of teachers placed into the various effectiveness ratings mandated by law varied considerably

across pilot districts. In the pilot sample as a whole, .2% of teachers (n=5) were classified as ineffective, 1.5% of teachers were classified as minimally effective, 63% of teachers were classified as effective, and 35% of teachers were classified as highly effective. But sample-wide results are not the ones worth noting. What is striking about the final assignment of ratings to teachers is just how much district-to-district variation is present in the percentage of teachers classified as “effective” vs. “highly effective.” Looking at the data above, for example, it can be seen that the percentage of teachers in each school district that received various effectiveness ratings differed strikingly across districts. In two districts (A and B), no teachers were classified as “highly effective,” whereas in two other districts (J and K), a large majority of teachers were labeled as “highly effective.”

Only two districts (E and J) identified any teachers as “ineffective,” and in most districts, very few teachers were rated as “minimally effective.”

To corroborate that such district-to-district variation was an artifact of classification formulae and cut points in districts, not real differences in teacher effectiveness, ISR researchers engaged in two types of analyses. In the first, ISR researchers examined the distribution of value-added measures for teachers in districts (as provided by SAS, one of the VAM vendors working with the pilot research program). This analysis showed that the distributions of measured teacher effects on students’ MEAP scores (as estimated for AY2011-2012 by the SAS MRM statistical model) were very similar across districts—suggesting that “real” differences on this dimension of teaching quality could not account for the differences in ratings distributions observed across districts. In fact, one district with the highest average teacher value-added scores in the sample (District B) classified no teachers as highly effective, while another district with teacher value-added scores below the sample average (District K) classified the vast majority of its teachers as highly effective. ISR researchers also examined whether teacher scores taken from classroom observation tools varied sufficiently across districts to account for the large differences in ratings distributions. Here, too, the explanation that real differences in the quality of classroom instruction were at the root of district-to-district differences in ratings distributions was implausible. Therefore, the most plausible explanation for the differences in ratings distributions shown on the previous page is that districts assigned different weights to performance criteria and set different “cut scores” for placing teachers into effectiveness ratings, and these processes, rather than real differences in the distribution of teaching quality were accounting for the observed differences in the distribution of teachers to effectiveness ratings observed among pilot districts.

Key Findings on Principals’ and Teachers’ Views of the Evaluation Process

The goal of PA 102 of 2011 was to create “a rigorous, transparent, and fair performance evaluation system” for teachers. One goal of the pilot of educator effectiveness tools was to gather data on teachers’ and principals’ views about these aspects of the teacher

evaluation process as enacted during the pilot year. Key findings in this area are now described.

Principals and teachers differed in how positively they viewed the observation tools used in the pilot.

In general, principals viewed the observation tools piloted by MCEE very favorably. To begin, the panel on the left-hand top of the next page shows that 76% of principals thought the tool they piloted was easy to understand and a similar percent thought the piloted tool was better than what they had used in the past. In addition, 89% of principals felt the observation tool they piloted was focused on important aspects of teaching that contribute to student learning (with just 20% reporting that the protocol omitted key aspects of teachers’ instructional practice). Perhaps for these reasons, 77% of principals felt the observation protocol they piloted provided a thorough picture of teachers’ instructional practice, and 64% felt the protocol was a good indicator of a teacher’s impact on student learning. Importantly, 50% of principals agreed or strongly agreed that they needed more information about the protocol they used.

The same table shows that teachers were less enthusiastic than principals about the piloted observation tools. Fifty three percent of teachers felt that the observation tools used in the pilot focused on key aspects of teaching that contribute to student learning. Moreover, while 62% of teachers thought the ratings assigned to them from classroom observations were accurate, 40% were worried that use of observation protocols would lead to unfair comparisons of teachers. In addition, only 47% of teachers thought the observation protocol their principal used was easy to understand, and 50% felt they needed more information about the protocol.

Both principals and teachers had favorable views of teacher conferencing activities associated with classroom observations.

The table on the right-hand top of the next page shows that the majority of principals and the majority of teachers agreed or strongly agreed that pre-/post-observation conferences were focused on targeted and specific feedback goals. The majority of principals and teachers also agreed or strongly agreed that teachers were putting ideas discussed in conferences into practice. Teachers reported that conferences most often focused on issues of student engagement and instructional strategies, and less often on issues of classroom management and subject matter content.

**At a Glance:
Principal & Teacher Views of Observation Tools**

Percent Agree or Strongly Agree	Principals	Teachers
Observation protocol focused on important aspects of teaching and learning	89%	54%
Observation protocol focused on activities that contribute to student learning	89%	53%
Observation protocol was easy to understand	76%	47%
Observation protocol can be used with just about any kind of lesson plan	72%	45%
Percent Agree or Strongly Agree	Principals	Teachers
Observation protocol provides a thorough picture of teachers' instructional practice	77%	
Observation protocol better than what I used in the past	75%	
Observation protocol is a good indicator of a teacher's impact on student learning	64%	
Observation protocol focuses on too many dimensions of instruction	40%	
Observation protocol omits key aspects of teachers' instructional practice	20%	
Percent Agree or Strongly Agree	Principals	Teachers
Ratings assigned to me during the observation were accurate		62%
Observation protocol says more about quality of teaching than student growth data		52%
Observation protocol can be a good tool in professional development		51%
Observation protocol will lead to unfair comparisons among teachers		40%
I could use more information about observation protocol used in annual evaluation	50%	50%

**At a Glance:
Principal and Teacher Views on Conferencing**

Percent of Teachers Reporting that Specific Topics Were Always or Often Discussed in Conference		
Student Engagement	75%	
Instructional Strategies	71%	
Evaluation of Student Learning	62%	
Classroom Management	49%	
Subject Matter Content	48%	
Percent of Principals and Teachers Who Agreed or Strongly Agreed with the Following Statements		
	Principals	Teachers
Conferences provided specific and targeted feedback	90%	71%
Feedback in conferences was geared to specific teacher goals	75%	66%
Conferences were characterized by lively give and take	76%	60%
Conferences were stressful	11%	26%
Teachers are putting ideas from conferences into practice	79%	68%
At a Glance: Principal and Teacher Views on Student Growth Tools		
Percent of Teachers Who Agreed or Strongly Agreed with the Following Statements		
Student learning objectives developed by me a good way to judge the academic growth of students	71%	
Locally-developed common assessments a good way to judge the academic growth of students	54%	
Standardized tests a good way to judge the academic growth of students	9%	
Procedures used to measure student growth at this school easy to understand	40%	
Student growth measures at this school take adequate account of student background and prior achievement	27%	
Percent of Principals Who Said the Following Should Be a Major Focus of Teacher Evaluations		
Student performance on standardized tests	47%	
Student performance on locally-developed assessments	75%	

At a Glance: Principal & Teacher Views of the Evaluation Process

Percent Agree or Strongly Agree	Principals	Teachers
Individuals who conducted annual evaluation had subject matter expertise needed	69%	44%
Teacher evaluations provided thorough assessment of teaching performance	83%	41%
Teacher(s) improved teaching as a result of evaluations	84%	28%
Annual evaluation an important basis for setting professional development goals	87%	42%
Teacher evaluations more about personnel decision making than improvement	13%	35%
Teacher evaluations simply a matter of 'going through the motions'	3%	31%
I spent too much time this year on the teacher evaluation process	47%	46%

The majority of teachers and principals agreed or strongly agreed that conferences were characterized by a “lively give and take” and that teachers were attempting to put ideas raised in conferences to practice. As with views of teacher observation tools, however, principals tended to be more positive about these issues than teachers.

Principals and teachers strongly favored the use of locally-developed assessments over standardized tests as a means of assessing student growth in teacher evaluations. Indeed, as the table on the previous page shows, 71% of teachers agreed or strongly agreed that student learning objectives they developed were a good way to judge the academic growth of their students, and 75% thought student performance on locally-developed tests should count as a major factor in teacher evaluations. By contrast, only 9% of teachers agreed or strongly agreed that standardized tests were a good way to judge the academic performance of their students, and just 47% of principals thought standardized test results should count as a major factor in teachers’ evaluations.

Despite the widespread use of teacher-made and locally-developed tests, many teachers had concerns about the student growth measures used in their annual evaluations. For example, as the table on the right-hand side of the previous page shows, only 40% of teachers agreed or strongly agreed that they understood the procedures used to measure student growth in their annual evaluations, and 27% felt that such

measures did not take adequate account of students’ home background and prior achievement. *Moreover, teachers were far less positive about the quality of the teacher evaluation process as a whole than were principals.* This can be seen in the table at the bottom of the previous page. It shows that 69% of principals (but just 44% of teachers) thought the individuals conducting teacher evaluations in their school had the necessary subject matter expertise. The table also shows that 83% of principals (but just 41% of teachers) thought the teacher evaluations conducted in the pilot year provided a thorough assessment of teaching performance. In addition, the table shows that 84% of principals (but only 28% of teachers) thought the evaluation process was leading to improvements in teaching performance. Finally, 87% of principals (but just 42% of teachers) thought that annual evaluations conducted in the pilot year could be used to set professional development goals.

Still, principals and teachers did not view teacher evaluation practices as mere exercises in bureaucratic procedure. For example, just 13% of principals and 35% of teachers agreed or strongly agreed that teacher evaluations in their school were more about making personnel decisions than promoting teachers’ professional growth, and only 3% of principals and 31% of teachers agreed or strongly agreed that teacher evaluations were simply a matter of going through the motions. However, nearly half of all principals *and* teachers agreed or strongly agreed that they spent too much time on the teacher evaluation process.

Chapter 3: Improving the Teacher Observation Process

In addition to asking ISR to examine how teacher evaluations were conducted in pilot schools, MCEE asked ISR researchers to explore how teacher evaluation practices might be improved in Michigan. This chapter addresses a central question raised by MCEE: Do the data from the pilot project suggest better ways to conduct classroom observations and use data from this process in annual teacher evaluations?

Improving the Use of Classroom Observation Data

The findings reported thus far suggest that the classroom observation data gathered during the pilot were subject to three problems: (1) principals (and other administrators) were *not* scoring all items on an observation tool, even when the vendor advised that scoring of an item was expected; (2) there were low levels of inter-rater reliability; and (3) principals (and other administrators) expressed concerns about the amount of time they were devoting to the teacher evaluation process—much of which was devoted to conducting classroom observations.

These issues raise three questions:

- How many classroom observations is it reasonable to expect administrators to conduct in order to obtain a good picture of a teacher’s instructional practice?
- Is it really necessary to score all items on a protocol, or can we just score a select few? and
- Is there any way to adjust the scores that teachers receive on classroom observations for lack of inter-rater reliability?

The first way ISR researchers addressed these questions was through an application of Generalizability (or G) theory. In essence, G studies examine how different errors in measurement affect measurement reliability. Using G theory, ISR researchers investigated three potential sources of error: (1) items not being rated by principals; (2) principals not conducting enough classroom observations; and (3) raters disagreeing about the scores to assign to the same lesson.

At a Glance: G Study Results by Observation Instrument*			
	Percentage of Variance in Multi-item Scale Score Due to Different Facets of Measurement		
	5D	FFT	TC
Items	60%	49%	34%
Occasions	16%	10%	9%
Raters	10%	16%	20%
Teacher	13%	13%	39%
Reliability of IRT Scale Scores As Implemented in Pilot			
Overall Reliability**	.69	.81	.82

*Data are for variation in scale scores estimated via a one-parameter, multi-level IRT measurement model that combine all 32 items in 5D, all 10 items in FFT, and only the “four corners” items from domains one to four of TC. The model nests items within occasions and occasions within teachers, and has rater fixed effects. Variance for rater effects was estimated by calculating the reduction in total variance that occurred after the introduction of rater fixed effects into the model. ISR researchers did not create a scale for the Marzano tool because of the high frequency of missing item data.

**Overall reliability is from administrator data and is calculated for IRT scales under the observation conditions obtained in the pilot implementation.

The G study conducted by ISR researchers showed that items, occasions, and raters were significant sources of error in the pilot classroom observation measures. Items can be a significant source of error variance in ratings because each item adds additional information about a teacher’s instructional practice. Thus, a teacher’s score for a given observation might depend critically on the items used to rate the teacher. In addition, occasions can be a significant source of measurement error because the instructional practices used by teachers might vary as a matter of deliberate choice (as, for example, when teachers change their practice at different points in an instructional unit, pursue different instructional goals and objectives, etc.). Teaching practices also change as a simple matter of random variation. Finally, as discussed earlier, raters can be a significant source of error because not all raters will score a lesson in the same way.

Overall, the G study results (shown on the current page) suggested that each of the observation tools

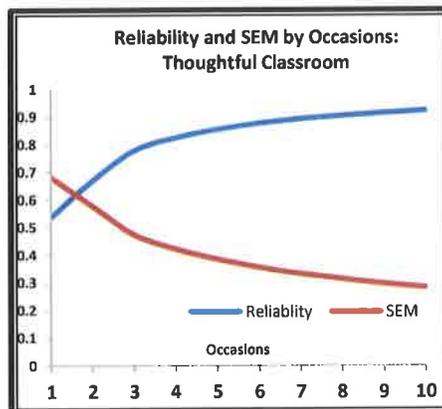
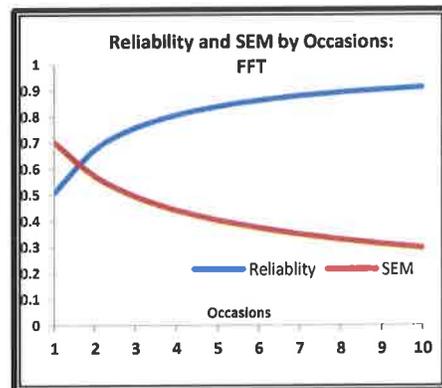
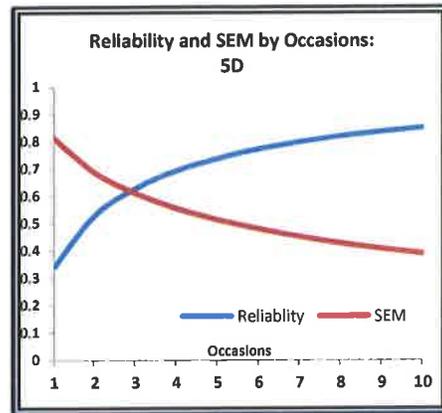
that ISR researchers examined had reasonable levels of reliability under the conditions of use in the pilot research. But the data also showed that each tool varied in the magnitude of particular errors. For example, item variance was larger for 5D and FFT than for TC; but rater variance was more pronounced for TC than for FFT and (especially) 5D. Overall, TC appears to contain more stable teacher variance than the other two tools, but it buys this scaling property by measuring very general aspects of teaching. As a result, it might not be as useful as the other protocols for measuring variation in instructional practice at a fine-grained level of detail.

The G study provides some guidance about the number of observations that need to be conducted to obtain reliable measures from classroom observation tools. The question of how many observations to conduct as part of the teacher evaluation process is especially important because we know from the survey data presented earlier that conducting observations was time-consuming and that many principals felt they were spending too much time on the evaluation process. The question we now address is the extent to which the reliability of classroom observation measures is affected by the number of times a teacher is observed. Relevant data are shown in the graphs immediately to the right. The G study results imply that:

- *The greater the number of observations conducted on a teacher, the more reliable (and precise) will be any measure of that teacher's instructional quality.*
- *However, the biggest gains in measurement reliability come when moving from one observation to about four observations. Improvements in reliability and precision will be slower as observations beyond four are conducted.*

The reader can see this by looking at the graphs to the immediate right. These graphs show changes in reliability (and the standard error of measurement [SEM]) for each scale as the number of observations increases. For analytic purposes, the graphs assume the number of items on each of the scales was 32 for 5D, 10 for FFT, and 4 for TC. The graphs assume a

At a Glance: Measurement Reliability as a Function of Number of Observations Conducted



single rater conducted the observations. The graphs use the variance components from the G study to chart how measurement reliability (and precision of measurement, denoted by the SEM) will vary as the number of observations increases. Clearly, conducting more observations improves measurement reliability, but as the graphs show, improvement slows after about four observations.

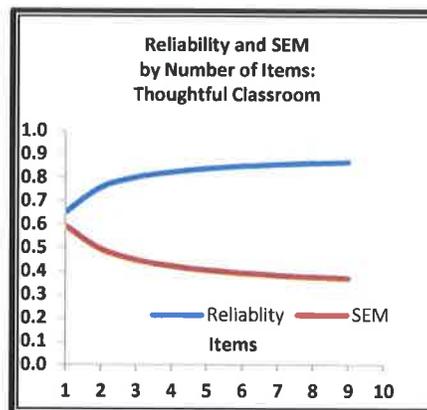
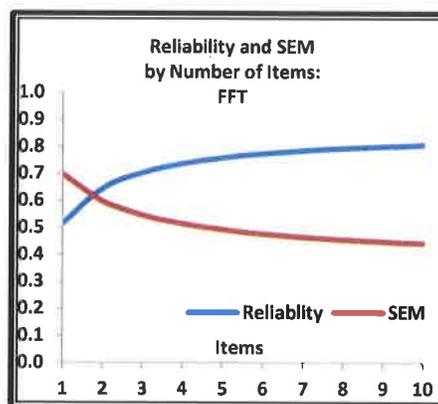
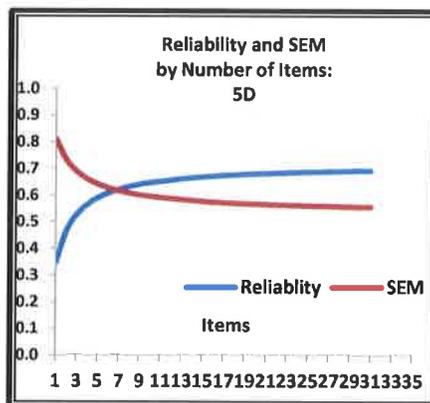
The G study also provides guidance about the number of items that need to be scored to obtain a reliable measure of teaching quality from an observational tool. This question is important, for as we discussed earlier in this report, many principals in the pilot did not score *all* of the mandatory items on a scale. We now use the G study results to see how scoring of different numbers of items affects reliability of measurement. In this example, we assume that a single rater observes a teacher on four occasions, and our analysis examines the reliability of one-parameter, multi-level IRT scale scores that would result if the number of items on that scale varied.

The results of this analysis are shown in the graphs on the right-hand side of this page. These graphs show that:

- *In general, the more items scored on an observation protocol, the more reliable (and precise) the resulting measure of a teacher's instructional quality will be.*
- *However, improvements in reliability and precision occur more slowly as the number of items scored increases beyond about 5 or 6 items.*

Importantly, these findings imply that through careful psychometric analysis, it might be possible to choose a subset of items from the larger item pool on a protocol and still be able to *reliably* discriminate among teachers' scored levels of instructional quality. As examples, the data show that if we cut the 5D protocol from 32 scored items to about 10 items, reliability of measurement would fall from about .66 to about .6, and that cutting the number of scored items on FFT from 10 to 6 would not change reliability at all. The trick to limiting the number of items in a protocol, however, is to be sure to include items in the reduced

At a Glance: Measurement Reliability as a Function of Number of Items Scored



form that come from the full range of item difficulties.⁶

There are potential advantages and disadvantages to using well-designed “short forms” of an observation instrument versus the vendor-recommended long form. For example, an advantage of a short form could be that by limiting the number items to be scored, principals would find it easier to learn how to use an observation tool (and would end up using it with more fidelity). The disadvantage of a short form is loss of qualitative information. As discussed earlier, the items included in the observation tools piloted in Michigan tended to provide information about particular aspects of teaching, and in the tools with more items, this allowed measurement of teaching practice at a more “fine-grained” level. It is possible that such fine-grained information is important to teacher learning and improvement, and this would argue against cutting the number of items included on a given observation protocol, even if a short form had reasonable reliability and precision.

The G study also provides guidance about the number of raters needed to obtain a reliable measure of teaching quality from an observational tool. Relevant data are presented in the graphs on the left-hand side of the next page. An answer to the question of how many different raters are needed to conduct observations is especially important given our earlier discussion of rater error in observation data. When rater error is present, it is desirable to have more than one rater observe in a classroom for two reasons. The first is that it might be possible to increase measurement reliability (holding constant the number of occasions) simply by adding raters. The second is that we can use data from observations by multiple raters to correct observed scores for rater error. The first approach handles rater error by averaging across raters; the second approach handles rater error by statistical control. Importantly, using multiple raters to conduct observations can be costly *and* present logistical problems, especially in the circumstance where all observations are conducted in real time by one or more

⁶ In fact, ISR researchers produced a set of analyses of this process for the FFT and 5D observation tools. The analysis (which will appear in our technical report) showed that, for both tools, it was possible to create a 5-item scale that represented an array of instructional dimensions. A 5-item scale for 5D had a reliability of .6 versus .66 using all items, and the 5-item FFT scale had a reliability of .795 versus .797 using all items. We did not reduce the items from the TC tool, since our scales to this point have only used the “four corners” items. Moreover, we did not conduct this analysis for the Marzano tool since there was too much missing item data to permit disciplined scaling.

administrators. Nevertheless, it is worth investigating the effects of using more than one rater during classroom observations if for no other reason than to understand the consequences of using the prevailing pattern of having only a single rater conduct classroom observations on a teacher.

We begin this discussion by examining how measurement reliability would change if a teacher was observed on a fixed number of occasions (set here at $n = 4$) using the standard observation tool, but on each occasion, we added another observer. The results of this examination are shown on the graphs on the left-hand side of the next page. Looking closely at these graphs, one can see that:

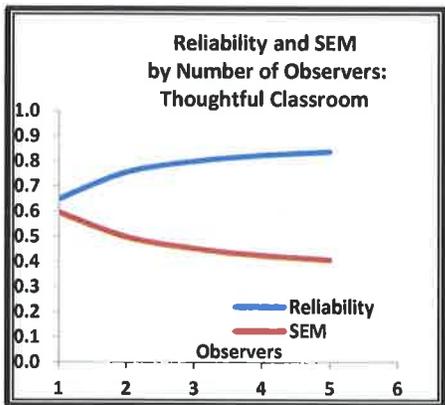
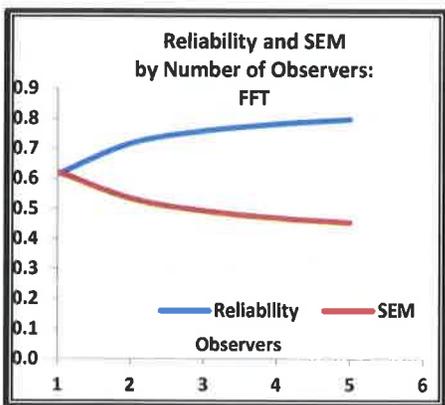
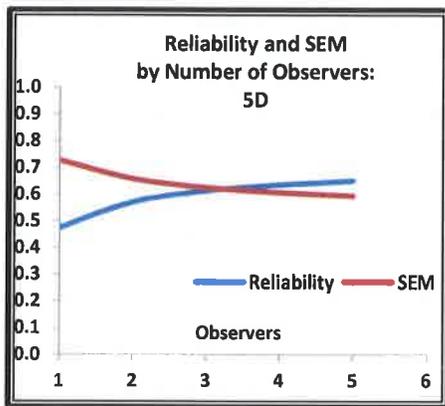
- *Adding more observers on each of four observation occasions increases measurement reliability.*
- *However, the biggest increase comes from adding a single observer on all occasions. After that, improvements in reliability come more slowly as more observers are added.*

The reader should note that this additional observer need not be the same person on *every* observation occasion. Indeed, if a school added one additional observer on every occasion, but that person was different each time, there would be positive benefits to measurement beyond improved reliability. In fact, this also would correct for rater bias, especially if the extra individuals were assigned at random from the pool of available raters (e.g., principals and administrators) in a district.⁷ The problem, of course, is that using multiple raters to increase measurement reliability (and correct scale scores for rater error) does not seem feasible in the usual school district context, where administrators already report being overburdened by conducting observations on the teachers they *must* evaluate.

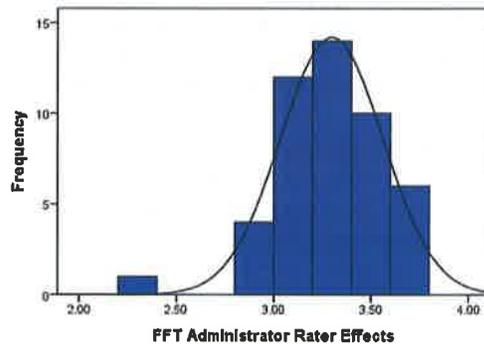
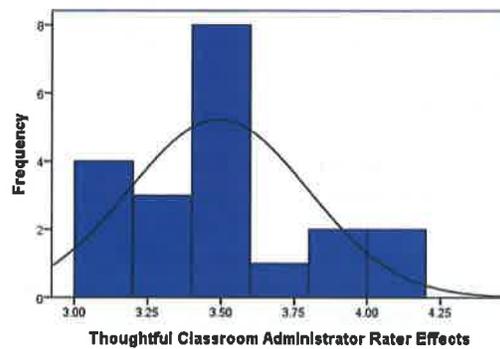
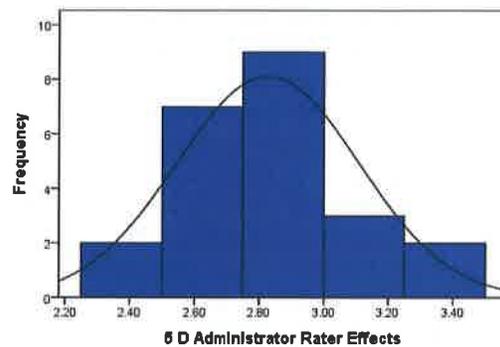
In the absence of using multiple raters in classroom observations, it will be especially important to address the potential bias in classroom observation data that results from rater errors in scoring. The graphs on the right-hand side of the next page illustrate the magnitude of these errors. For example,

⁷ By assigning additional raters at random, it would be possible to statistically separate rater from school effects in measurement models, producing not only improved precision, but also improved accuracy in scale scores.

At a Glance: Measurement Reliability as a Function of Number of Observers on Each Occasion



At a Glance: The Magnitude of Rater Effects Among Principals*



*These graphs show the distribution of scores that a hypothetically "average" teacher would receive if observed on the same occasions by all of the principals in the pilot sample. The graphs are based on estimates of rater fixed effects in the one-parameter, IRT model described previously. One can see from the graphs that the same teacher would receive different ratings depending on the principal conducting the observations.

they show the distribution of scores that a hypothetically “average” teacher would receive if he or she was observed on the exact same days by each of the principals in the pilot study. In the data for 5D, the average principal would assign that teacher a score of about 2.8, but as the histogram on the graph shows, many other principals would assign a higher or lower score to that teacher. Those scores would range from a low of about 2.3 to a high of about 3.4, with 75% of scores being clustered in the range of 2.6 to 3.0. For FFT, rater effects are shown as distributed around an average of about 3.3, but vary from a low of about 2.4 to a high of about 3.75, with the majority of scores ranging from about 3 to 3.5. For TC, the rater effects are distributed around an average scale score of about 3.5, and are distributed fairly evenly across a range of about 3 to 4.

The reader might wonder if the rater effects shown in the graphs are substantial enough to warrant concern. If considered as standardized “effect sizes” (a common metric in research), the rater errors observed in the pilot data are not particularly large. For example, two principals who are a standard deviation apart in the distribution of rater effects will assign the “average” teacher in the pilot sample a score that differs by .40 of a standard deviation on the TC scale and around .20 of a standard deviation for the FFT and 5D scales. These are generally considered to be medium to small effect sizes in educational research.

However, even small differences in scoring can have important consequences for a teacher’s assignment to a particular effectiveness rating. For example, imagine an average teacher being rated by two principals using the 5D protocol, both of whom score that teacher’s instruction with error. Suppose further that this imaginary teacher’s “true” score on the 5D protocol is at the mean of the score distribution (i.e., 2.8) but that one principal is “lenient” and assigns a score of 3.2 to her teaching, while another principal is “severe” and assigns a score of 2.4 to the observed teaching. Suppose further that the teacher works in a district where she must score above 3.0 on the observation protocol in order to be classified as “effective” in her annual rating. Using the scores assigned by the one rater who scored her teaching as a 3.2, this teacher is classified as effective, even though her “true” score is 2.8. Therefore, the illustration shows that even seemingly small rater errors of the sort shown in the figures can have strong implications for teacher evaluations.

One way to correct for rater errors in the teacher evaluation process is through statistical adjustments to observation scores. The most principled adjustment for rater error would be to randomly assign a district’s pool of administrators to conduct classroom observations in all of the schools in the district. Under randomization, lenient and severe raters would be randomly distributed across teachers, and while rater errors would decrease precision of measurement, assigned scores would be largely unbiased (due to random assignment). This is how most research projects operate. However, most districts in the pilot assigned a single (fallible) observer—typically the school principal—to conduct all of the observations on a given teacher.

When every teacher has been observed by only a single rater, districts can attempt to correct for rater error in one of two ways. One way would be to center teacher observation scores in each school around the mean score assigned by that school’s principal. Using this approach, one simply subtracts the school mean observation score from a teacher’s assigned observation score and then compares teacher scores within schools. Alternatively, one could adjust a teacher’s observation score for rater error by reference to the grand mean of ratings in a district (or the state). Suppose, for example, that there were 5 principals in a district, and each principal rated 25 teachers. One could calculate the mean of all ratings in the district (the so-called “grand mean”) and then deviate each principal’s mean rating from the grand mean. Lenient principals would have average scores above this mean (e.g., +.2), whereas severe principals would have scores below this mean (e.g., -.3). One could then use this information to adjust teachers’ scores. Using this latter approach, adjusted scores would remain in the original scoring metric and comparisons in adjusted scores could be made across schools.⁸

As a supplement to statistical adjustment, any administrator who conducts classroom observations should be required to engage in “calibration” training for use of an adopted observation tool. The goal of this training is to minimize rater errors by having principals learn how to assign scores that are close to the ones that would be assigned by an expert rater conducting the same classroom observation. In practice, calibration training typically involves having principals observe and score videos of classroom

⁸ In the first method, a teacher’s adjusted score = (teacher’s score – school mean). In the second method, a teacher’s adjusted score = (teacher’s score – principal’s rater effect), where the principal’s rater effect = (grand mean of ratings – principal’s mean rating).

teaching until the scores they assign show only very small departures from scores assigned to the same videos by “expert” raters. Such training, it should be noted, is offered by all of the observation tool vendors working in the pilot program and should be considered as a mandatory aspect of training for any state-approved observation tool. However, calibration training will not usually eliminate rater error, and absent random assignment of multiple raters to classroom observations, rater error will almost certainly be present in the observation scores of teachers. As a result, education authorities responsible for conducting classroom observations should attempt to introduce simple, statistical corrections for rater error when they use observation scores for evaluation purposes.