Evaluation Plan for Wisconsin’s Educator Effectiveness System

Prepared for Wisconsin Department of Public Instruction

By
Josef Dvorak
Daniel Marlin
Christa Pugh
Nels Rude
Mai Choua Thao

Workshop in Public Affairs
Robert M. La Follette School of Public Affairs
Spring 2014
# Table of Contents

List of Tables .......................................................................................................................... iv  
List of Figures .......................................................................................................................... iv  
Foreword .................................................................................................................................. v  
Acknowledgments ..................................................................................................................... vi  
Executive Summary ................................................................................................................... vii  
Introduction ............................................................................................................................... 1  
Background of Educator Evaluations ....................................................................................... 2  
  Traditional Educator Evaluation Systems .............................................................................. 3  
  Emerging Educator Evaluation Systems ............................................................................... 4  
    Multi-Dimensional Use of Evaluation Measures ............................................................... 4  
    Providing Support and Meaningful Feedback ................................................................. 6  
  Wisconsin’s Educator Effectiveness System ...................................................................... 7  
Methodology ............................................................................................................................. 11  
Logic Model ............................................................................................................................... 12  
Evaluation Reviews ................................................................................................................. 16  
  Research Questions ............................................................................................................. 16  
  Research Question Recommendations ............................................................................. 17  
  Data Collection .................................................................................................................. 18  
  Data Collection Recommendations .................................................................................. 19  
Metrics ....................................................................................................................................... 20  
  Metric Recommendations ................................................................................................. 23  
  Data Analysis ...................................................................................................................... 23  
  Data Analysis Recommendations ..................................................................................... 24  
Limitations in Evaluating Educator Evaluation Systems .................................................... 25  
Conclusion ................................................................................................................................. 27  
Appendix A: The RFP Process ................................................................................................. 29  
Appendix B: Sample Research Questions ............................................................................. 31  
Appendix C: Sample Survey and Focus Group Questions .................................................... 32  
Appendix D: Sample Metrics ................................................................................................. 36
List of Tables

Table 1. Wisconsin Educator Effectiveness System Framework ........................................9
Table 2. Wisconsin EE System Logic Model ......................................................................15
Table 3. Implementation Metrics .......................................................................................20
Table 4. Educator Outcome Metrics ....................................................................................22
Table 5. Student Outcome Metrics .....................................................................................23

List of Figures

Figure 1. Map of States With Written Evaluations .........................................................11
Foreword

This report is the result of a collaboration between the La Follette School of Public Affairs at the University of Wisconsin–Madison and the Wisconsin Department of Public Instruction. Our objective is to provide graduate students at La Follette the opportunity to improve their policy analysis skills while contributing to the capacity of the Department of Public Instruction to evaluate educator effectiveness.

The La Follette School offers a two-year graduate program leading to a master’s degree in public affairs. Students study policy analysis and public management, and they can choose to pursue a concentration in a policy focus area. They spend the first year and a half of the program taking courses in which they develop the expertise needed to analyze public policies.

The authors of this report are all in their final semester of their degree program and are enrolled in Public Affairs 869 Workshop in Public Affairs. Although acquiring a set of policy analysis skills is important, there is no substitute for doing policy analysis as a means of learning policy analysis. Public Affairs 869 gives graduate students that opportunity.

This year the workshop students were divided into six teams. Other teams completed projects for the U.S. Government Accountability Office, the City of Madison, the Wisconsin Department of Children and Families, the Wisconsin Children’s Trust Fund, and the Financial Clinic of New York City.

The state of Wisconsin, like many others, is exploring how to improve its education system. As part of that process, it has passed a statewide educator evaluation program that has just begun to be implemented. The state needs to understand if this new approach is working, and to do so needs a rigorous and careful evaluation. This report draws on insights from the field of program evaluation, as well as a review of what other states are doing, to offer guidance about how Wisconsin can best evaluate new education policies.

Professor Donald Moynihan
Professor of Public Affairs
May 2014
Madison, Wisconsin
Acknowledgments

We are grateful to the Educator Effectiveness team at the Wisconsin Department of Public Instruction, and in particular Katharine Rainey, for providing us with background information, feedback, and support throughout the development of this report. We appreciate the time that Ms. Rainey spent with us and the thoughtfulness with which she approached our questions. We are also grateful to Curtis Jones at the University of Wisconsin-Milwaukee for providing us with information about the evaluation of Wisconsin’s Educator Effectiveness System pilot program and general evaluation approaches. In addition, we appreciate the help of state education agencies and evaluation firms across the country who took the time to answer our questions about state approaches to evaluation. Finally, we thank Professor Donald Moynihan for his guidance and advice.
Executive Summary

Over the past several years, an increasing number of states, including Wisconsin, have implemented programs to evaluate educator performance. In 2011, the Wisconsin Legislature passed Wisconsin Act 166, which established requirements for a statewide educator evaluation program. As a result, Wisconsin’s Educator Effectiveness (EE) System began as a pilot program in the 2012-2013 academic year and districts statewide must implement the EE System beginning in fall 2014. This program will provide educators and principals with ratings of their performance based on formal observations of their professional practice and evidence of student outcomes in order to improve educator performance, with the ultimate goal of supporting improved student achievement. To assess whether the program meets this goal, Wisconsin’s Department of Public Instruction (DPI) plans to carry out an evaluation study of the statewide System designed to answer questions about the program’s implementation as well as its impact on educators, principals, and students. An external consultant, to whom DPI will award a contract through the state’s competitive Request for Proposal process, will conduct this study.

This report supports DPI’s program evaluation by recommending appropriate research questions, data collection methods, metrics for measuring implementation and outcomes, and data analysis methods. The report provides a logic model that lays out how the program should work and can serve as a starting point to measure whether the program operates as intended. In addition, the recommendations draw from research centered on states that have conducted similar studies of their educator evaluation systems. The information gathered from these other states provides numerous examples that can inform DPI’s study. Overall, the report provides a picture of what other states are doing to evaluate their educator evaluation programs, as well as tailored guidance to Wisconsin to support its evaluation.

The report concludes with several specific recommendations for DPI. First, DPI should design research questions that address broad categories including program implementation, educator perceptions of program experiences, the burdens of reform, changes to teacher and educator behavior, long-term student outcomes, and metric and process fidelity. Next, DPI’s data collection methods should include interviews, focus groups, and surveys to collect a range of quantitative and qualitative data. Implementation metrics should include educator perceptions of and comfort with the EE System, as well as data about the number and percentage of evaluations completed and the amount of time spent conducting evaluations. Educator outcome metrics should include educators’ perceptions of changes to their practices, while student outcome metrics should utilize value-added modeling. Finally, the report recommends publishing information gleaned from surveys, focus groups, and interviews, as well as conducting data analysis using statistical models and methods.
Introduction

Over the past five years, state policies requiring measurement of educator effectiveness have gained traction across the United States. Competition for federal Race to the Top grants and U.S. Department of Education requirements for states wishing to pursue waivers for some provisions of the Elementary and Secondary Education Act have led to an increase in states developing consistent statewide processes for rigorous, data-based evaluations of teachers and administrators (Doherty and Jacobs 2013, 2). Many states required regular evaluations prior to the introduction of these federal requirements, but allowed for local flexibility with regard to the evaluation process. Recently, however, the number of states requiring evaluations has increased and states have instituted requirements to ensure evaluation processes that are consistent statewide. From 2009 to 2013, the number of states requiring annual evaluations of educators increased from 15 to 27. Including additional states that require less frequent evaluations, 40 states mandate regular educator evaluations that include objective evidence of student learning (Doherty and Jacobs 2013, 5).

Although an increasing number of states now have consistent, formal educator evaluation programs, few states have researched the statewide implementation of these systems or the outcomes for educators and students. Because many of these programs have been developed recently, many states are conducting pilot programs or have insufficient data for a wide-scale evaluation. While some states have evaluated implementation or the attitudes and perceptions of system participants, few studies have measured the outcomes of educator evaluation systems for educators, and fewer still have investigated system impact on student outcomes.

In 2010, Wisconsin State Superintendent Tony Evers established a design team at the Wisconsin Department of Public Instruction (DPI) to develop an Educator Effectiveness System (EE System) for Wisconsin to evaluate Wisconsin educators and principals using one consistent, statewide process.1 The program DPI designed provides educators and principals with ratings and substantive feedback on their performance based on observations of professional practice and evidence of student outcomes. In 2011, legislators passed Wisconsin Act 166, which formally established state requirements for educator evaluation (Department of Public Instruction 2014b). In the 2012-13 and 2013-14 school years, schools across the state participated in a pilot program testing components of the EE System. During this time, DPI has collected feedback and implementation data related to the pilot with the assistance of researchers at the University of Wisconsin-Milwaukee.

---

1 Throughout this report, the term “EE System” will refer to Wisconsin’s newly developed system for educator evaluation, while the term “educator evaluation” will refer more generally to systems or programs designed to evaluate educators.
Beginning in 2014, DPI will begin statewide implementation of the EE System, following which it hopes to evaluate system implementation and outcomes with the help of an external consultant. This type of evaluation is important to demonstrate that the EE System is meeting its goals of improving teacher and principal performance and, ultimately, improving student outcomes. Qualitative and quantitative evidence of system impacts can serve as a justification for the time and money necessary to fully implement the EE System. To hire an independent contractor, the state of Wisconsin requires agencies to use a competitive bid process in most cases, which begins with the issuance of a Request for Proposal (RFP) detailing the work the contractor will do (see Appendix A for a description of the RFP process). In response to a request from DPI, this report provides information that DPI will use to develop an RFP to hire an independent evaluator who will conduct research on EE System implementation and outcomes over the course of several years.

This report will provide guidance for DPI regarding potential research questions, data collection, metrics, and data analysis, drawing from methods other states use to evaluate the efficacy of their own evaluation systems. The report begins with background information that describes the development and theory of change of educator evaluation systems as well as the components of Wisconsin’s EE System. The report then outlines a logic model to inform the evaluation recommendations and present information about the evaluation of educator evaluation systems in other states. This research leads directly to recommendations for research questions, data collection methods, metrics for assessing implementation and educator and student outcomes, and data analysis methods. Finally, the report discusses the limitations of the recommended evaluation methods.

**Background of Educator Evaluations**

Research has consistently shown that teachers’ performance within classroom settings is one of the most important influences on student achievement and growth (Darling-Hammond 2000; Rivkin et al. 2005; Rockoff 2004). In a 2004 study of 79 urban and rural schools in Tennessee, researchers found that the difference in student achievement gains between a non-effective and an effective teacher is one third of a standard deviation for reading and a greater standard deviation for mathematics (Nye et al. 2004). Their findings are similar to those of other studies of teacher impacts and indicate that “teacher effects are real” (Nye et al. 2004, 17). Consequently, in the past several years, efforts to improve the academic achievement of students have focused on the creation of educator evaluation systems. Educator evaluation systems assess educators’ professional practice and its effects on student learning. In addition, these systems provide feedback that informs professional development activities to increase teaching effectiveness. Stakeholders including states, school districts, and educational organizations are increasingly implementing new evaluation systems or improving their existing systems.
Public interest and research on the effects of evaluation systems, along with new evaluation practices and methods, have transformed the structure of educator evaluation systems. Though researchers have gathered little information thus far on the relationship between the implementation of educator evaluation systems and student outcomes, research on the general principles and components of these systems continues to grow. This section of the report reviews the existing literature on the development and different components of educator evaluation systems.

**Traditional Educator Evaluation Systems**

Understanding earlier educator evaluation systems can inform how these systems evolved over time. Traditional educator evaluation systems differ from today’s systems in a number of ways. First, earlier educator evaluation systems relied heavily on teacher qualifications and credentials. For example, Laura Goe, a leading research scientist at Educational Testing Service, notes that earlier evaluation systems assessed teacher quality based on an educator’s certification status, degree level, or licensure test scores. Goe (2009, 17) believes that teacher qualifications are “weak indicators of teacher effectiveness” and do not fully measure effective performance. Empirical studies also suggest that assessing teacher education and qualifications does not fully capture teacher effectiveness and quality. Jennifer King Rice, an associate professor of education policy at the University of Maryland and an associate at the Economic Policy Institute, has published numerous articles and book chapters on education policy issues such as equity, efficiency, and the sufficiency of the U.S. education system. In her 2003 research analysis on teacher quality, Rice (2003, 6) states that assessments should include more sophisticated measures “of what teachers know and can do” rather than conventional measures, such as highest degree earned. Both Goe (2009) and Rice (2003) agree that though educational credentials do matter in teacher assessments, they are not the most appropriate measures by which to assess teacher effectiveness.

A second characteristic of traditional educator evaluation systems is the use of single classroom observations. Evaluation systems place a great deal of weight on classroom observations, as they provide an in-person display of classroom content, management, and interaction between educators and students. Traditional evaluation systems have a tendency to rely exclusively on single annual classroom observations, usually conducted by school principals (Toch and Rothman 2008). However, as Toch (2008) notes, checklists used within classroom observations often lack sufficiently comprehensive assessments. He states, “In most instances, the evaluator just checks off either ‘satisfactory’ or ‘unsatisfactory’ for checklist items that ask if educators are ‘presentably dressed,’ and if the educator ‘started on time’” (Toch 2008, 23).

Lastly, earlier evaluation systems provide summative rather than formative evaluations of teacher performance. Summative evaluations are those that provide
educators with an overview of their performance but do not provide guidance or support designed to improve performance. Mathers and Oliva (2008) state that school districts primarily used summative evaluations to make decisions on dismissals, tenure, salary, transfers, and personnel assignments. They argue that for evaluation systems to function effectively, formative evaluations must be implemented into the system as well. They define the intent of formative evaluations as “provid[ing] teachers with feedback on how to improve performance and what types of professional development opportunities will enhance their practice” (Mathers and Oliva 2008, 3). An effective system, they argue, should incorporate both types of evaluations.

In summary, earlier evaluation systems assessed educators on conventional educator attributes such as degree completion, specific coursework, and experiences. Many argue against the use of such measures as they fail to fully capture teacher effectiveness (Goe et al. 2009; Mathers and Oliva 2008; Toch and Rothman 2008). Secondly, traditional educator evaluation systems depend heavily on a single annual classroom observations, which many scholars now believe is a poor and unfair form of assessing teacher performance (Toch and Rothman 2008). Lastly, traditional evaluation systems are more summative in nature rather than formative. Though summative evaluations provide educators with a summary of their performance and help administrators make decisions on salary, tenure, and job dismissals, they fail to provide formative evaluations that produce meaningful feedback for performance improvement.

**Emerging Educator Evaluation Systems**

Policy initiatives and recent research on transforming traditional one-dimensional systems have led to new methods of evaluation which use more multi-dimensional, formative, and data-driven processes. Most educator evaluation systems today incorporate these changes to improve teacher accountability as well as to provide support and professional development for educators who may be underperforming. Comprehending how newer evaluation systems function and what they look like can help enhance understanding of why these systems shifted from one-dimensional systems to more multi-dimensional data-driven systems that allow for educator improvement and support.

**Multi-Dimensional Use of Evaluation Measures**

One of the strongest arguments in favor of improving educator evaluation systems is the previous lack of comprehensive measures that fully assess teacher performance. Traditional educator evaluation systems, as noted earlier, rely heavily on classroom observations and fail to utilize other measures of effective teaching. In addition, it is difficult to compare principals or teachers from different schools using traditional systems. More recent systems use a variety of measures to help determine teacher quality performance. Aside from the conventional use of classroom observations, newer systems now include value-added models and school surveys. Using a variety of measures in evaluation
ensures fairness of the assessments and provides more accurate and comprehensive evaluations of teacher performance.

Many emerging evaluation systems utilize value-added modeling (VAM). VAM is a statistical evaluation method that gathers student test score data to estimate how much teachers contribute to annual student learning growth (McCaffrey et al. 2003). VAM identifies student growth by comparing a student’s test scores from the beginning and end of the school year, making statistical adjustments to control for variables that may influence student growth but are outside of a teacher’s control such as students’ socioeconomic background, their parents’ level of education, and their English language status. Controlling for these factors allows VAM to identify the portion of a student’s academic growth for which the teacher is directly responsible.

The emergence of VAM in evaluation systems occurred for multiple reasons. First, data-driven methods such as VAM gained popularity as a result of the conventional view that teachers should enhance student learning over time. Next, VAM is seen as a strong indicator of future teacher performance, as opposed to previously used indicators such as the teacher’s number of years of experience, education attainment level, and specific advanced coursework. Lastly, federal education initiatives and laws encourage school districts to use VAM. For example, President George W. Bush’s administration included incentives in the Teacher Incentive Fund for school districts to assess teachers based on VAM. Additionally, President Barack Obama’s Race to the Top program requires states to use VAM as a condition of competing for a portion of the $4.3 billion fund (Glazerman et al. 2011).

Academic researchers and scholars, however, caution that utilizing VAM as the sole form of assessment fails to fully represent and capture teacher effectiveness. Most scholars recommend the use of VAM coupled with other measures such as classroom observations, artifacts, surveys, and “observational information from school principals or peer assessors” (Harris 2010, 2). Douglas N. Harris, an economist at Tulane University, believes that the “issue is not whether to use value-added, but how to do so” (2010, 2). Harris, who has conducted and published research on the effects of VAM and its impact in evaluation systems, recognizes that value-added measures have the ability to filter out-of-school factors from a teacher’s performance. Referred to as the Cardinal Rule of Accountability, which Harris defines as “holding people accountable for what they can control,” VAM can provide a summative assessment of teacher performance (2010, 3). However, Harris advocates combining VAM with other measures of assessment to achieve both summative and formative evaluations. Harris (2010) is not alone in his support for multidimensional use of VAM and other classroom-related evidence. Steven Glazerman of Mathematica Policy Research and other researchers who specialize in the analysis of educator evaluation systems stress the incorporation of multiple sources of information that can optimally “reflect the full range of teacher performance” (2011, 4). The
incorporation of multi-dimensional measures requires other bodies of evidence of effective teaching coupled with VAM scores and traditional classroom observations.

Other evidence that can help measure teacher performance includes classroom portfolios, student surveys, principal assessments, and student assessments. Classroom portfolios are collections of school artifacts, including lesson plans, student work, written teacher reflections, and other instructional materials, used often in newer evaluation systems (Goe et al. 2008). Goe et al. (2008) note that classroom portfolios are not only comprehensive, but can help determine aspects of teacher performance that may not always be visible through traditional classroom observations. Another increasingly used evaluation measure is principal assessments. Unlike evaluations conducted by trained evaluators, evaluations by principals may provide more information about teacher performance because they include teacher interviews before and after the evaluation to help identify possible areas of improvement (Goe et al. 2008). Lastly, student assessments can be used as evidence of effective teaching. Though there are concerns that student assessments may be biased, several studies have shown that student ratings can be a reliable source of information about a teacher’s performance (Kyriakides 2005; Peterson, Wahlquist, and Bone 2000; Wilkerson et al. 2000). Wilkerson et al. (2000) found that student ratings were highly correlated with student achievement when compared to principal and teacher-self ratings. Unlike other evaluators, such as principals, trained evaluators, and teachers, students have the most contact with, communication with, and exposure to teacher performance. Thus, an incorporation of evidence including classroom portfolios, principal, and student assessments in addition to VAM can lead to more accurate and comprehensive assessments of teachers.

Providing Support and Meaningful Feedback
Implementing effective educator evaluation systems goes beyond identifying effective and ineffective teachers. Many systems are focused on providing support and relevant feedback to educators to improve their effectiveness, fostering professional growth and development. In addition, many systems today shift away from earlier forms of summative feedback to both summative and formative feedback to provide meaningful reviews for teachers.

Professional development serves as an essential part of educator evaluation systems. Little (1987) defines professional development as any activity for educators geared toward improving teacher performance. Professional development can include formal activities such as conferences and workshops as well as more informal activities (Little 1987). For example, Schifter and Fosnot (1993) explain that professional development can take the form of mentoring, co-teaching, and group discussions, and Grossman et al. (2001) mention that professional development can include teacher networking, study groups, or even book clubs dedicated to talking about performance improvements and experiences. These definitions of professional development illustrate that professional growth can take place almost anywhere and can be formal and
informal, as long as the participation is intended to improve teacher performance and classroom instruction.

As a result, fostering professional development and learning opportunities for educators can give educators insights on how their instruction affects student learning as well as how to improve their instructional practices. Peter Youngs, an associate professor of educational policy at Michigan State University who has conducted research on state and district policies related to teacher evaluation and professional development, argues for comprehensive professional development that provides “strong content focus, engage[s] teachers as active learners, [is] of sufficient duration, and involve[s] participation with colleagues, [and] help[s] teachers acquire both content knowledge and pedagogical content knowledge” (2013, 7). Professional development is a growing component of new and improving educator evaluation systems. In a research analysis of five educator evaluation systems, Little (2009) found established ongoing professional development and career advancement opportunities for educators. Little (2009) and Goe et al. (2014) agree that educator evaluation systems tied to professional learning opportunities have greater impact and effectiveness versus systems that fail to include professional development opportunities for teachers. Lastly, most of the newer educator evaluation systems include feedback designed to provide clear and constructive assessments of teacher performance and improvements. Unlike earlier educator evaluation systems, most systems today combine summative and formative feedback. In particular, Goe et al. (2014) reinforce this idea. Specifically, they assert that feedback should not only convey a review of teacher performance but offer suggestions on “how to improve performance and what types of professional opportunities will enhance their practice” (Goe et al. 2014, 5). Feedback is an integral part of the evaluation system because it helps educators identify their performance strengths and weaknesses and offers information that educators can use to improve their performance (Darling-Hammond 2012). Thus, feedback is valuable, allowing for learning and growth.

In sum, educator evaluation systems transformed considerably in recent years, shifting from internal reviews, such as observations and assessments conducted by school district staff, to greater external review and accountability. Further, evaluation systems have shifted from one-dimensional, summative systems to multi-dimensional, formative, and data-driven systems geared toward assessing not only teacher quality but also teacher effectiveness. In addition, newer evaluation systems are expanding professional growth and learning opportunities as well as providing more useful feedback that educators can use as resources for improvement. Educator evaluation systems will no doubt continue to evolve as new research offers improved methods by which to evaluate teaching.

**Wisconsin’s Educator Effectiveness System**

Wisconsin has made great strides in reforming its education system, setting clear outcome goals in areas such as graduation rates, college readiness, and reading
proficiency, and holding high standards for students and educators alike (Wisconsin Department of Public Instruction 2014a). One way in which Wisconsin has sought to improve its education system is by overhauling its teacher and principal evaluation systems. Wisconsin’s approach to evaluating teachers and principals is distinctive in the way in which it balances state and local involvement; uses multiple measures of evaluations, including student achievement; attempts to isolate the impact of teachers on student growth; and aligns evaluation results to inform development opportunities for all educators (Hull 2013).

The development of Wisconsin’s EE System began in 2010 when Wisconsin State Superintendent Tony Evers formed an Educator Effectiveness Design Team that was charged with developing recommendations for a statewide educator evaluation system. The team consisted of various stakeholders, including leaders of state education organizations and educator preparation programs, the Governor’s office, and DPI. In designing the EE System, these stakeholders focused on the guiding principle that the ultimate goal of education is student learning and improvement and effective educators are essential to achieving that goal. Furthermore, effective practice leading to better student achievement requires continuous improvement and monitoring.

To achieve the goal of providing students with highly qualified and effective educators, the team determined that an educator evaluation system would need to identify effective practices aligned with student learning, and that development and evidence of effective educator practices and student learning must be documented. Additionally, the EE System is designed to inform professional development, inform educator preparation programs, support human resource decisions, and provide credible, valid, reliable, comparable, and uniform outcomes across districts (Wisconsin Department of Public Instruction 2011). The resulting evaluation criteria rely on multiple indicators, including student academic growth data, a regulatory framework for implementation, model performance-based observation systems, and methods to support improvement and incentives for performance (Wisconsin Department of Public Instruction 2010). In 2011, the design team recommendations were released in a final report (Wisconsin Department of Public Instruction 2011).

Based in part on the recommendations of the design team, the Wisconsin Legislature passed a comprehensive education reform package (Wis. Stat. § 115.415, 2011) that required DPI to develop, through the administrative rulemaking process, a state system to evaluate educators in public schools, including independent charter schools. The law also required DPI to create an equivalency process to offer school districts a different approach to implementing evaluations, but one that is aligned with the state system. To support the continuous improvement of educator practices, the evaluation system is designed to provide for fair, valid and reliable evaluations by using multiple measures, including educator practices and student outcomes.
Under the state evaluation system, student outcomes make up 50 percent of an educator’s evaluation score (see Table 1). The student outcomes score draws upon multiple measures of student performance, including performance on state assessments as measured using VAM, student/school learning objectives (SLOs), district choice, school-wide reading scores at the elementary and middle-school levels, and graduation rates at the high school level (Wisconsin Legislative Council 2011).

<table>
<thead>
<tr>
<th>Table 1. Wisconsin Educator Effectiveness System Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student Outcomes (50% of Evaluation)</strong></td>
</tr>
<tr>
<td>a) Student and School Learning Objectives</td>
</tr>
<tr>
<td>b) Value-Added State Assessment Data</td>
</tr>
<tr>
<td>c) School-Wide Reading Scores (elementary and middle-school levels)</td>
</tr>
<tr>
<td>d) Graduation Rates (high school level)</td>
</tr>
<tr>
<td>e) District-Specific Outcomes</td>
</tr>
<tr>
<td><strong>Educator Practice (50% of Evaluation)</strong></td>
</tr>
<tr>
<td>a) Interstate Teacher Assessment and Support Consortium (teachers)</td>
</tr>
<tr>
<td>a. Danielson’s <em>Framework for Teaching</em></td>
</tr>
<tr>
<td>b) Interstate School Leaders Licensure Consortium</td>
</tr>
<tr>
<td>Educational Leadership Policy Standards (principals)</td>
</tr>
<tr>
<td>a. Wisconsin Framework for Principal Leadership</td>
</tr>
</tbody>
</table>

Source: Wisconsin Department of Public Instruction (2013)

Although the Design Team intended state assessments to represent a significant proportion of an educator’s student outcomes score, the VAM measure of state assessments may not be attainable until at least 2017, due to the complexity of data collection at the individual student and teacher level (Wisconsin Department of Public Instruction 2013). The SLO portion of the student outcome measure requires educators to identify detailed, measurable goals for students to achieve in an academic year (Wisconsin Department of Public Instruction 2014c). Finally, district choice, graduation rates at the high school level, and schoolwide VAM reading scores at the elementary and middle school level make up a small portion of the outcomes score.

The other 50 percent of an educator’s score evaluates the quality of educator practice as defined by rubrics, including the Danielson Framework for Teaching and the Wisconsin Framework for Principal Leadership, aligned to the 2011
Interstate Teacher Assessment and Support Consortium (InTASC) or standards for principals established under the 2008 Interstate School Leaders Licensure Consortium Educational Leadership Policy Standards (ISLLC) (Wisconsin Legislative Council 2011). The InTASC and ISLLC standards have received national recognition as rigorous and robust research-based standards of professional practice (Wisconsin Department of Public Instruction 2011, 4). Lastly, the EE System requires evaluators to undergo thorough training and certification prior to assessing educators to ensure high, reliable standards are met (Wisconsin Department of Public Instruction 2013).

Under state law, school districts must begin using an evaluation system, whether the state system or an alternative process approved through an application for equivalency, by the 2014-2015 school year. State law requires DPI to create an equivalency application and review process to provide flexibility for school districts looking to utilize alternative educator practice evaluation methods. Alternative models must align to the InTASC or ISLLC standards and create teacher rubrics that align to the following four domains:

- (a) planning and preparation;
- (b) the classroom environment;
- (c) instruction
- (d) professional responsibilities and development (Wisconsin Legislative Council 2012).

DPI has developed a comprehensive system to support districts as they implement the System. For example, DPI contracted with Teachscape©, a program aligned with the nationally recognized Danielson Framework for Teaching, to facilitate online evaluation and observation management, evaluation training, and professional development systems. DPI is also developing additional online training modules, and the agency will provide school districts with funding to alleviate training, licensing, regional support and technical assistance costs. Finally, regional Cooperative Education Services Agencies will support implementation (Wisconsin Department of Public Instruction 2013).

In the 2012-13 school year, DPI initiated a development pilot program, which included more than 600 educators testing individual and isolated measures of the System such as principal practice, teacher practice, and student learning objectives (Wisconsin Department of Public Instruction 2013). According to DPI staff, DPI has expanded upon those efforts in 2013-14 by implementing a full pilot program that includes approximately 1,200 additional educators and has a broader scope. Instead of testing different pieces of the System, participants will be asked to implement it in full. DPI is on track for statewide implementation in the 2014-15 school year, as required by state law.
Methodology

This report draws upon external evaluations and RFPs from other states to determine best practices for studies of educator evaluation systems and inform the development of an evaluation plan for DPI. Research for the report began by searching for educator evaluation system evaluation reports on the website of each state’s department of education and creating a list of states with published evaluation reports. To ensure that all published evaluation reports were located, the authors contacted the educator evaluation teams of every state via email, unless research determined that a state had no consistent statewide educator evaluation program in place. If states did not respond to the initial email, the authors followed up with one additional email and one phone call. Thirty-two of the 45 states with formal evaluation systems were successfully contacted. Written evaluations were retrieved from 15 of these states, including eight states that received Race to the Top funds: Arizona, Connecticut, Delaware, Georgia, Indiana, Massachusetts, Michigan, Minnesota, Nevada, New Hampshire, New Jersey, Ohio, Rhode Island, Tennessee, and Washington. While Massachusetts and Delaware also had published RFPs, the information contained within the RFPs was not specific enough to provide useful information. In sum, the authors located documentation from 15 states, or 33 percent of the 45 states with formal evaluation systems. For a visual representation of these states, see Figure 1; Race to the Top grantees from which a written evaluation was retrieved are highlighted in blue and non-grantees from which an evaluation was retrieved are highlighted in red.

Figure 1. Map of States With Written Evaluations

Source: Authors; Power Point Toolkit (2013)
To ensure that the data retrieval process was thorough, the authors also contacted three evaluation firms recommended by a researcher at the University of Wisconsin-Milwaukee: Westat, Mathematica, and the American Institutes for Research. These large research and evaluation firms provide their clients, including businesses, organizations, and government, with objective research on policies and programs, including program evaluations. The purpose of this step was to use triangulation to locate any evaluations not obtained directly from states. Through this process, the authors located one additional state evaluation as well as useful background information about educator evaluation and general evaluation principles. Because this report draws on multiple sources for evaluation documentation, including state websites, state education officials, and major evaluation firms, it can be stated with reasonable confidence that the number of existing relevant evaluations collected can sufficiently inform this report.

Next, the authors studied the evaluations to determine common practices and novel or unique components related to evaluations. Recommendations were developed after coding research questions, data collection methods, metrics, and data analysis methods used in each evaluation, with attention to practices common across several states, as well as to novel practices that resulted in valuable information.

To develop the recommendations for DPI, authors relied on principles of program evaluation, which require establishing the expected outcomes of a program, and then designing a research methodology that will determine the actual impacts of the program relative to the desired outcomes (Shadish et al. 2001). To define the expected EE System program outcomes, the authors designed a logic model that visually describes how the inputs and activities related to the EE System lead to program outputs and outcomes. Once the expected program outcomes were identified, the authors drew upon the information gathered from other states’ research to recommend research questions, data collection methods, metrics, and data analysis methods that could measure those outcomes.

**Logic Model**

The following logic model visually displays how the EE System works, starting with the inputs and activities required for successful implementation and moving to the outputs and outcomes that are expected as a result (United Way 1996, 3). Logic models often represent a first step in an evaluation process because they demonstrate how a program works in theory, while the evaluation gathers data to determine whether the program is working in practice. The model presents the conceptual linkages between the EE System and expected outputs and outcomes (see the logic model in Table 2).
DPI and the state of Wisconsin, as well as individual schools throughout the state, will invest numerous resources, or inputs, into the EE System. DPI’s EE System team will invest time into developing and implementing the program, and educators and principals across the state will invest time becoming familiar with the program and implementing it in their schools. DPI will also invest funding for tools and training. Training in person and online is also an essential input to ensure that participants are familiar with the system and its use. Finally, inputs will include several tools developed by other organizations, including Teachscape© software, Charlotte Danielson’s Framework for Teaching, and existing high quality common assessments of student learning.

DPI and schools will use these initial inputs to engage in activities, actions that DPI and school districts will take to have the desired impact. All school staff will be trained in the teacher evaluation process, and some district staff will be trained to earn their Teachscape© certification and become certified evaluators. Educators will develop SLOs, Self-Ratings of Performance, and professional practice goals and strategies. After developing SLOs, educators will collect evidence of their students’ progress towards meeting those goals. Educators and principals will be evaluated during their first year of practice and every third year thereafter, and their evaluators will give them substantive written or verbal feedback after each evaluation. Based on identified strengths and growth opportunities, educators will participate in professional development activities, including coaching or mentoring with more experienced educators. Finally, an external evaluator will study the implementation and outcomes of the EE System, and DPI will make adjustments based on the results of that evaluation.

As a result of these inputs and activities, the necessary components of the intervention will be implemented. These outputs are the products that immediately result from the inputs and activities of DPI and school districts. Numerical outputs include the number of schools implementing the EE System and the level of implementation; the number of staff who are certified evaluators; the number of staff trained in the evaluation process; the number of staff who receive an educator effectiveness rating and the amount of feedback accompanying those ratings; the educator-evaluator ratio; and the number and quality of mentoring relationships that develop between more and less experienced educators.

The activities associated with the intervention are expected to produce changes in educators’ and students’ short-term, intermediate, and long-term outcomes. These outcomes are the end benefits that will result from this program for educators and students. In the short-term, anticipated outcomes include improved educator awareness of their own strengths and weaknesses compared to their awareness at the beginning of the program, improved principal awareness of both their own and their schools’ strengths and weaknesses compared to their awareness at the beginning of the program, and improved quality and focus of educator mentoring and professional development activities based on the evaluation feedback.
compared to mentoring and professional development prior to program implementation. Two to four years after the statewide implementation of the program, a number of intermediate outcomes will result. These include an improvement in the quality and rigor of SLOs and the process of developing SLOs as educators become more familiar with the system, compared with SLOs prepared in the first year of the program; continually improving mentoring and professional development for educators; and improvement in educator progress towards meeting personal performance goals that will lead to improved educational practices compared with practices in the program’s first year. Finally, the anticipated long-term outcome of the EE System is an improvement in student educational outcomes resulting from improvements in principal and educator practice, compared with student outcomes prior to the implementation of the program. These educational outcomes may include common assessment scores, graduation or attendance rates, or other measures.
## Table 2. Wisconsin EE System Logic Model

<table>
<thead>
<tr>
<th>Inputs</th>
<th>Activities</th>
<th>Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time of DPI EE System team</td>
<td>District staff will be trained to pass Teachscape© certification to become certified evaluators</td>
<td>Number of staff who are certified evaluators</td>
</tr>
<tr>
<td>Time of teachers, principals, and district staff</td>
<td>All educators and principals will be trained in the teacher evaluation process</td>
<td>Number of educators who are trained in the evaluation process</td>
</tr>
<tr>
<td>Teachscape© software</td>
<td>Educators will develop SLOs, Self-Ratings of Performance, and professional practice goals and strategies</td>
<td>Evaluator/educator ratio</td>
</tr>
<tr>
<td>The Danielson Framework for Teaching</td>
<td>Educators will collect evidence of student progress towards meeting SLOs</td>
<td>Number of SLOs developed and amount of evidence of student progress collected</td>
</tr>
<tr>
<td>DPI funding for tools and training</td>
<td>Teachers and principals will be evaluated during their first year of practice and every third year thereafter</td>
<td>Number of educators who receive educator effectiveness score summaries</td>
</tr>
<tr>
<td>In person and online training modules</td>
<td>Educators will receive written or verbal formative feedback as a result of evaluations</td>
<td>Amount of feedback given to educators based on evaluations</td>
</tr>
<tr>
<td>High-quality common assessments of student learning</td>
<td>Educators will coach and mentor each other based on identified strengths and growth opportunities</td>
<td>Number of mentor relationships between more and less experienced educators</td>
</tr>
<tr>
<td></td>
<td>External evaluator will study implementation and outcomes and DPI and school districts will adjust program based on the evaluation</td>
<td>Number of schools implementing different aspects of the educator effectiveness program, including basic state requirements and optional best practices</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Short-term (&lt; 1 year)</th>
<th>Intermediate (2-4 years)</th>
<th>Long-term (&gt;4 years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved educator self-awareness of strengths and weaknesses</td>
<td>Increased quality and rigor of SLOs and strategies designed to reach goals</td>
<td>Improvement in student educational outcomes, including test scores and graduation rates</td>
<td></td>
</tr>
<tr>
<td>Improved principal awareness of school’s strengths and weaknesses</td>
<td>Improved process of developing SLOs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Improved quality and focus of mentoring and professional development activities for educators</td>
<td>Improvement in educator progress toward achieving personal professional practice goals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Increased teacher leadership in the school</td>
<td>Continually improving mentoring and professional development for educators</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Improved educational practices designed to improve student performance</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Authors.
Evaluation Reviews

Following the development of the logic model and the collection of state evaluation reports, the authors analyzed each evaluation to highlight common and novel practices relating to research questions, data collection methods, metrics to measure implementation and educator and student outcomes, and data analysis. This section provides the results of this analysis.

Research Questions

Within each evaluation, research questions center on how educator evaluation programs and statewide systems operate, whether they meet stated goals and objectives, and how they could be improved. Broadly, these evaluations seek to determine implementation fidelity (whether school districts implement the system as the state intended), educators’ perceptions of and experiences with their respective systems, impacts on teacher and principal behavior, and short- and long-term student outcomes.

Implementation-related questions seek to identify district-specific obstacles and challenges, unintended consequences of reform, and the percentage of mandated tasks completed. New Jersey’s evaluation asks about barriers to implementation (Firestone et al. 2013), while Minnesota analyzes challenges by questioning how individual districts address each difficulty (FHI360 2013). Additionally, Arizona, Minnesota, and Ohio seek to find whether implementation resulted in any positive or negative unintended consequences to existing practices. Examples of unintended consequences could include whether teachers encourage students to take responsibility for their learning or whether districts enact systems in which the training is so comprehensive that it becomes overwhelming. Illinois attempts to discern whether evaluators completed required training and certification programs or utilized training resources, as well as how much time educators spent using provided resources (Quarterly Report 2013). Finally, Nevada asks questions regarding implementation fidelity and possible improvements to the system (Rabinowitz et al. 2013).

Most state evaluations also attempt to decipher educators’ perceptions and experiences with implementation to understand burdens on teachers and changes to their workloads, how to ease these burdens, and whether evaluators and educators addressed concerns about these burdens. Regarding educator perspectives, Arizona asks to what extent participants perceived “that initial implementation of the new teacher effectiveness measures resulted in: (a) teachers increasingly reflecting about their practice; (b) more instruction-focused interactions between teachers and administrators; (c) changes in instructional practice; and/or (d) teacher collaboration/knowledge sharing” (Arizona Department of Education 2013, 13). New Jersey and Connecticut examine how educators experience the system and why experiences vary between districts,
individuals, grade level, and subject taught (Firestone et al. 2013; Donaldson et al. 2013).

Some evaluations also seek to determine the impact evaluation systems have on educator behavior. Ohio examines “changes in individual instructional practices and levels of embedded change within [local education authorities]” (MGT of America and Susan Zellner 2012, 9). Similarly, Connecticut seeks to determine how educators changed their practices as a result of the new system (Donaldson et al. 2013). While most evaluations attempt to discern educator perceptions, these evaluations are distinctive insofar as they specifically ask about how educators changed their behaviors.

Finally, two states look at whether the presence of evaluation systems correlate with improved student outcomes. Ohio’s evaluation asks about its system’s impact on student achievement (MGT of America and Susan Zellner 2012). Michigan’s evaluation, in part, focuses on “devising rigorous, fair, and useful procedures for measuring student growth and estimating teachers’ contributions to their students’ achievement” (Rowan et al. 2013, 2). To do so, Michigan designs research questions to fit within a value-added framework.

**Research Question Recommendations**

After reviewing each evaluation, the authors recommend that Wisconsin select five broad research questions relating to the topics mentioned above: implementation, educator perceptions of program experiences, the burdens of reform, changes to teacher and educator behavior, and long-term student outcomes. Including these questions will ensure consistency with best practices and the proposed logic model. For each topic, the evaluation should include specific sub-questions. Further, each question should be framed in a manner that results in the creation of clearly measurable and defined metrics, discussed further below. See Appendix B for sample research questions.

To evaluate fidelity of implementation specifically, the state should ask questions identical to those in WestEd’s validity study of Nevada’s educator evaluation framework. These questions include: “Is the infrastructure around data collection sufficient and effective? Can we suggest any changes, improvements, or refinements to current processes? Is there content and construct validity? Are the processes around their use feasible and relatively low-burden, to ensure fidelity, sustainability, and ability to scale up?” (Rabinowitz et al. 2013, 8). Asking these questions ensures that evaluators continually improve processes, metrics, and data collection. In addition, these questions allow the state to review many of the processes identified in the logic model, such as evaluator training, educator development of SLOs, and the creation of mentoring relationships.
**Data Collection**

Most states utilize a combination of qualitative and quantitative data collection. The evaluations collect quantitative data to find correlations between variables, statistically significant differences between variables, effect sizes of the educator evaluation intervention, and singular variable effects on each study’s dependent variable. Additionally, the evaluations collect qualitative data to provide context for metrics and performance data; elucidate subjective assessments, opinions, and problems with the educator evaluation program; and provide examples and explanations that clarify best practices or failures that occurred in isolated circumstances.

Nearly all evaluations rely on the collection of survey data. These involve open- and closed-ended questions that allow teachers and principals to assess satisfaction, adequacy of training prior to implementation, and changes in perceptions of workloads. Survey targets include teachers, principals, or both. Survey administration varies by state, but online administration is most common. To clarify survey results, many states engage in site visits in which teachers, evaluators, and educators participate in focus groups, interviews, or both. Extensive classroom observation also occurs in Michigan and New Hampshire. All evaluations make use of at least one of these methods.

Although surveys include large numbers of respondents, focus group and interview participants had to be chosen through sampling due to the time constraints. Nevada and Connecticut use stratified sampling to achieve this goal, recognizing differences between school district may affect each district’s ability to implement its educator evaluation system. Broadly, this sampling separates schools into categories based on specific characteristics, such as student demographics or location in an urban or rural area, and randomly samples within each stratum to ensure evaluators compare alike groups and draw subjects from each school type (Shadish et al. 2002). In Nevada, stratifying and coding schools was done by type (preschool, middle school, or high school), size, district, socioeconomic index (defined as the percentage of students receiving reduced-price or free lunch), the percentage of special education students, and the percentage of English language learners (Rabinowitz et al. 2013).

While each state recognizes the importance of collecting and analyzing quantitative data, no consensus has emerged on its most effective use. Quantitative data vary by state, often centering on closed-ended survey questions, summaries of survey data, and measurement of task completion relating to implementation. Evaluations also collect teacher and principal ratings to create a picture of a district’s distribution of scores.

Lastly, while data remain sparse, a growing movement exists to use quantitative data to determine the effect of implementing an educator evaluation system on student outcomes. Michigan and New Hampshire include data collection.
strategies to measure student outcomes in their pilot evaluations. Each state has explicitly expressed a commitment to take a similar approach in subsequent statewide evaluations. Michigan calculates value-added scores based on data collected on effectiveness ratings (Rowan et al. 2013) while New Hampshire explicitly uses student growth percentiles for educators in “tested” grades (New Hampshire Department of Education 2013).

**Data Collection Recommendations**

DPI should utilize a combination of quantitative and qualitative data collection strategies in its evaluation. The data collection, along with the metrics discussed below, will inform the “outputs” and “outcomes” of the logic model, allowing DPI to assess directly whether it is implementing the EE System correctly as well as student and educator outcomes. Most evaluations conduct online surveys as a cost-effective method of administration which allows for easy tabulation of large numbers of qualitative responses while mitigating interviewer bias. Surveys also allow an opportunity for diverse stakeholders to voice their opinions, which could lead to increased system buy-in. To eliminate the possibility of bias among participants, these surveys should be anonymous. Finally, qualitative data such as that gleaned from surveys can provide nuanced, detailed information regarding the implementation and perceptions of the EE System that quantitative data would not capture.

Therefore, DPI should send an online survey to teachers and principals approximately every six months to assess EE System implementation. Sending out surveys any more often would be burdensome; doing so any less often would fail to give an adequate picture of implementation. Further, DPI already has the ability to send surveys to schools across the state, so this recommendation would result in minimal additional burden. DPI should also conduct interviews or focus groups to gather more detailed qualitative information about program implementation and outcomes.

Additionally, stratified sampling is a useful technique. It would be infeasible and cost-prohibitive for DPI to visit every school in a district and conduct focus groups, interviews, or both. Therefore, DPI should use samples of districts with particular characteristics to efficiently collect implementation data. Moreover, within the sampled schools, DPI should select teachers and principals to interview and partake in focus groups using a stratified sample that draws educators from different grades and subject areas. To minimize response bias, school officials should not choose survey, interview, and focus group participants. As DPI has extensive expertise in conducting surveys and focus groups, this report largely defers to DPI’s judgment to determine the appropriate mixture of qualitative data collection methods, relevant stratification criteria, individual inclusion criteria for participating in interviews or focus groups, questions used, and how to structure each. Sample questions can be found in Appendix C.
**Metrics**

Each evaluation uses metrics to measure outcomes and answer specific research questions. This section outlines metrics to assess implementation, educator outcomes, and student outcomes. Broadly, implementation-related metrics examine and attempt to measure educator perceptions of educator evaluation systems, educator understanding of such systems, the level of educator progress in completing implementation tasks, and whether school-level evaluators had appropriate resources to provide training (including those needed to conduct classroom observations). This final metric, while uncommon, is crucial to implementation. It is impossible for evaluators to implement an educator evaluation system, and for DPI to realize the outcomes outlined in the proposed logic model, if districts do not have the resources with which to do so. Table 3 presents implementation metrics across the states, and describes how the metrics answer research questions and relate to the logic model.

<table>
<thead>
<tr>
<th>Metric Category</th>
<th>State-Specific Metrics</th>
<th>Relationship to Research Questions</th>
<th>Relationship to Logic Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher and Educator Perceptions of Efficacy</td>
<td>Arizona: Measured agreement with the following: the number of formal classroom observations was adequate to assess my performance, the pre-observation conference(s) fully prepared me for what to expect during my observation(s), the training I received on the new evaluation process was adequate for me to effectively participate this year, the new teacher evaluation process represents an improvement over prior teacher evaluations</td>
<td>Did the evaluation system create opportunities for professional growth? How did educators experience the evaluation system? How was [the evaluation system] enacted in each site? Was there fidelity to the state model?</td>
<td>Outputs: Number of staff who are certified evaluators, number of educators who are trained in the evaluation process, evaluation/educator ratio, number of educators who receive educator effectiveness score summaries, number of schools implementing different aspects of the educator effectiveness program, including basic state requirements and optional best practices</td>
</tr>
<tr>
<td><strong>System Understanding/Level of Educator Readiness</strong></td>
<td>Arizona: Rates of agreement with the following: the pre-observation conference(s) fully prepared me for what to expect during my observation(s), the criteria on which I was evaluated were made clear to me. Rhode Island: Measures comfort with the evaluation process</td>
<td>How did educators experience the evaluation system? How was the evaluation system enacted in each site? Was there fidelity to the state model?</td>
<td>Activity: All educators and principals will be trained in the teacher evaluation process.</td>
</tr>
<tr>
<td><strong>Resource Adequacy</strong></td>
<td>Indiana: “What level of concern do you have regarding each of the following items?” on a 5-point scale: Resources to conduct classroom observations, resources to collect student performance data, resources to provide training for evaluators, resources to provide training for staff, resources for increased compensation component. Minnesota: Rates of agreement for teachers and administrators with the statement: “I am confident that my district will provide me with the opportunities and resources to achieve [system] goals.”</td>
<td>How was the evaluation system enacted in each site? How did educators experience the evaluation system?</td>
<td>Input: DPI funding for tools and training</td>
</tr>
<tr>
<td><strong>Quantitative Measures</strong></td>
<td>Michigan: Number of teachers evaluated, median number of observations per teacher. Percent of principals reporting conferences, teachers reporting conferences, principals using growth tools in evaluations, teachers receiving mid-year report, teachers receiving end-of-year report, and teachers assigned effectiveness rating. Median hours spent on training, observing probationary teachers, observing tenured teachers, rating/reporting probationary teachers, rating/reporting tenured teachers. Median days/year spent on teacher evaluation.</td>
<td>How was the evaluation system enacted in each site? To what extent has implementation diverged across individuals or subgroups? What explains these variations?</td>
<td>Outputs: Number of staff who are certified evaluators, number of educators who are trained in the evaluation process, evaluation/educator ratio, number of educators who receive educator effectiveness score summaries, number of schools implementing different aspects of the educator effectiveness program, including basic state requirements and optional best practices</td>
</tr>
</tbody>
</table>

Sources: MGT of America and Susan Zellner 2012; Cole et al. 2012; FHI 360 2013; Donaldson et al. 2014; Rhode Island Department of Education 2013; Authors.
As with implementation metrics, educator outcome metrics tend to include perceptions. However, the perceptions measured are different and relate to how teachers and educators may have changed their practices as a result of the educator evaluation system. Moreover, only a few evaluations use quantitative measures to assess educator outcomes. For example, Arizona employs a variety of measures, including distributions of teacher scores and statistical relationships among components (such as student surveys, observations, and student growth) (Arizona Department of Education 2013). Table 4 outlines these findings.

Table 4. Educator Outcome Metrics

<table>
<thead>
<tr>
<th>Metric Category</th>
<th>State-Specific Metrics</th>
<th>Relationship to Research Questions</th>
<th>Relationship to Logic Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Educator perceptions</td>
<td>Arizona: Whether teachers changed their educational practices or increased collaboration</td>
<td>To what extent did participants perceive that initial implementation of the new teacher effectiveness measures resulted in teachers increasingly reflecting about their practice; more instruction-focused interactions between teachers and administrators; changes in instructional practice; and/or teacher collaboration/knowledge sharing?</td>
<td>Outcomes: Improved educator self-awareness of strengths and weaknesses, improved principal awareness of school’s strengths and weaknesses</td>
</tr>
<tr>
<td></td>
<td>Ohio: Teacher and administrator perceptions of the system’s effect on behavior and performance</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Connecticut: Self-reported changes to educational practices</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rhode Island: Self-reported improvement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative measures</td>
<td>Arizona: Distributions of teacher scores and statistical relationships between different components (such as student surveys, observations, and student growth)</td>
<td>Did [the evaluation system] create opportunities for professional growth for teachers and administrators?</td>
<td>Outcomes: Improved quality and focus of mentoring and professional development activities for educators, improvement in educator progress toward achieving personal professional practice goals</td>
</tr>
<tr>
<td></td>
<td>Rhode Island: Breakdowns of teacher and administrator scores, including overall effectiveness ratings and scores in the different components of the final rating</td>
<td>To what extent were these differentiated across individuals or subgroups?</td>
<td></td>
</tr>
</tbody>
</table>

Sources: Arizona Department of Education 2013; MGT of America, and Susan Zellner 2012; Donaldson et al. 2013; Rhode Island Department of Education 2013; Authors

Lastly, this report sought to investigate student outcome metrics. Most of the evaluations reviewed did not attempt to measure student outcomes because they only evaluated pilot programs or were only conducted over one year. Therefore, many did not have sufficient data to determine the impact of educator evaluation
systems on student outcomes. However, as Table 5 shows, Michigan, Nevada, and Tennessee all attempted to measure student outcomes in some way.

**Table 5. Student Outcome Metrics**

<table>
<thead>
<tr>
<th>Metric Category</th>
<th>State-Specific Metrics</th>
<th>Relationship to Research Questions</th>
<th>Relationship to Logic Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student Outcomes</strong></td>
<td>Michigan: VAM</td>
<td>What is the impact on student outcomes?</td>
<td>Outcomes: Improved educational practices designed to improve student performance, improvement in student educational outcomes, including test scores and graduation rates</td>
</tr>
<tr>
<td></td>
<td>Nevada: VAM, controlling for school type, size, district, socioeconomic index (percent free or reduced-price lunch), percent special education classification, percent english language learner classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tennessee: Growth on state assessments</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sources: Rowan et al. 2013; Rabinowitz et al. 2013; Tennessee Department of Education 2013; MGT of America, and Susan Zellner 2012; Authors.

**Metric Recommendations**

Based on the evaluations reviewed, DPI should frame its metrics in a manner that easily answers its research questions. Using the logic model, this report provides sample metrics for DPI in Appendix D. Broadly, to assess implementation, DPI should ask about educator perceptions of how the EE System is being implemented, educator comfort and understanding of the EE System, the percentage of evaluations completed, and the number of evaluations that each evaluator conducts. Further, of particular importance to assessing implementation is determining whether evaluators have the resources they need, as well as how much additional time the EE System requires of evaluators. To assess educator outcomes, DPI should ask educators about perceptions of their own improvement and whether teachers and districts have changed any policies or practices. Finally, the authors support the state’s use of VAM. This is not only a useful measure of student outcomes but ensures consistency with state law. Such metrics allow DPI to assess the countable “outputs” in the logic model and begin to address outcome measures, such as improved educator and principal awareness of the system, the implementation of teacher leadership and professional development initiatives, and longer-term student outcomes.

**Data Analysis**

The extent to which evaluations analyzed the data collected and the methods undertaken to do so is largely a product of the type of data collected (qualitative
vs. quantitative). Qualitative data analysis methods include coding survey, focus group, and interview responses as well as information gleaned from observations. Case studies illustrate trends that illuminate best practices or problematic situations.

Quantitative analysis ranged from rudimentary calculations of agreement rates, teacher scores, and test results, to the creation of descriptive statistics such as mean, median, and mode, to more advanced statistical analysis methods. Evaluators used Microsoft Excel, STATA, or SPSS to perform analysis when adequate data were present to calculate cross-tabulations\(^2\), t-tests\(^3\) for determining statistical significance and to determine correlations.

Most importantly, because causality is impossible to determine without a randomized controlled trial, which is often infeasible in the field of education, these evaluations are correlational studies that “simply observ[e] the size and direction of a relationship among variables” (Shadish et al. 2002, 12). Further, “correlations also do little to rule out alternative explanations for a relationship between two variables such as education and income. That relationship may not be causal at all but rather due to a third variable…that causes both education and high income” (Shadish et al. 2002, 7). A further discussion of this limitation follows below. However, while causality cannot be determined, correlations consistent with the logic model provide some support for the effect of the system.

Certain evaluations employ other quantitative methods. Connecticut assesses whether variations in data are based on district demographics, grade levels, and whether an educator has tenure (Donaldson et al. 2013). As referenced above, Michigan investigates the possibility of using VAM to evaluate teachers (Rowan et al. 2013). Delaware employs an alternative method to analyze its survey results, using psychometric testing to assess construct validity, or the “validity of inferences about the higher order constructs that represent sampling particulars” (Shadish et al. 2002, 38).

**Data Analysis Recommendations**

Findings from focus groups and interviews that clarify problems or best practices should be published with all evaluation reports to provide insight into how well the EE System is working, generate hypotheses about any unintended consequences, raise questions, and illustrate problems or successes. Quantitatively, DPI should utilize the following analytical techniques: calculating descriptive statistics for variables, t-tests to determine significance of variables, and VAM for determining student outcomes. Variables worth testing include

---

\(^2\) A cross-tabulation is a table with two or more dimensions “that records the number (frequency) of respondents that have the specific characteristics described in the cells of the table” (Qualtrics 2011).

\(^3\) A t-test is a method to determine if a variable is statistically significant and if there are differences between two variables.
measures of student socioeconomic status such as free or reduced-price lunch, district size, and the number of English language learners. The amalgamation of these methods should provide a holistic picture of implementation and the EE System.

DPI could also consider utilizing an interrupted time series regression model to analyze student outcomes. This model would allow DPI to compare variables before and after implementation. In this case, implementation would function as the “interruption,” or treatment. To do this, “a basic time-series design requires one treatment group with many observations before and after a treatment” (Shadish et al. 2002, 175). Because DPI has baseline data on every student in the state, it could use this model to analyze student outcomes before and after implementation. This will allow DPI to determine whether the EE System is correlated with improvement in long-term student outcomes. As defined by the logic model, this is the ultimate goal of the implementation of the EE System. However, there are some limitations of this model, which are discussed below.

**Limitations in Evaluating Educator Evaluation Systems**

The methods, metrics, and analyses proposed here face several limitations, including those specific to statistical modeling methods, validity threats, and survey and focus group data collection. First, it is impossible to establish causality using these methods; Wisconsin will not be able to determine with absolute certainty whether the implementation of the EE System has a direct causal impact on long-term student outcomes. While a randomized experiment would allow DPI to make such a determination, DPI cannot conduct an experiment in which some school districts’ educators partake in the EE System as a treatment while others do not because the law requires all educators and teachers to participate. In short, the inability to establish a control group and randomly assign participants to treatment also makes it impossible to determine a causal relationship between the EE System and gains in student achievement. At best, evaluators will be able to determine associations between variables, correlations, and effect sizes. In addition, qualitative analysis will offer insight into best practices and district-specific contexts inhibiting implementation.

Even if evaluators were able to undertake a natural experiment, regression, value-added, and interrupted time-series statistical modeling designs face numerous internal and external validity threats. Internal validity refers to “the validity of inferences about whether the relationship between two variables is causal” (Shadish et al. 2002, 508). Shadish et al. (2002) note internal threats include history, maturation, attrition, and testing-related biases. In the context of the type of evaluation proposed here, the history threat may arise when events (such as increased school funding or systemic changes in socioeconomic conditions) that may cause chosen outcomes (specifically student achievement levels) to improve occur concurrently with the treatment (implementation of the EE System). Maturation bias results when “naturally occurring changes over time could be
confused with a treatment effect” (Shadish et al. 2002, 55). Here, students over time may become more naturally interested in learning, make larger commitments to their studies, and outcomes may improve irrespective of educator proficiency or involvement levels. Significant program attrition could also occur among students, teachers, or principals, which could cause fluctuations in student outcomes. In this case, losing “respondents to treatment or to measurement can produce artifactual effects if that loss is systematically correlated with conditions” (Shadish et al. 2002, 55).

Finally, evaluators should concern themselves with testing fatigue, the possibility that teachers and educators tire of meeting procedural requirements and constantly filling out forms, and do so inaccurately to finish quickly, as well as bias. Here, repeated exposure to a test may cause participants to change behavior in ways that “game the system.” This gaming can be confused with a treatment effect (Shadish et al. 2002). The presence of any of these biases undermines evaluators’ ability to establish a causal relationship between the treatment (implementation of the EE System) and their chosen dependent variables (short-term goals, student achievement measures, and implementation fidelity). None of the statistical models advocated herein are immune to these potential threats.

Because this report suggests value-added and interrupted time-series statistical modeling as a way to measure student outcomes, the authors next discuss limitations of their use. Even a proper utilization of VAM is not immune to selection, maturation, and testing bias. The ability to draw statistical findings from these data is further complicated by

the non-random assignment of students to teachers and schools, the limitations of particular tests both for measuring the full range of desired knowledge and skills and for measuring learning gains, and the difficulties of disentangling the contributions of many influences on learning—multiple teachers, parents, tutors, specific curricula, and the availability of useful learning materials, as well as other school resources like time, class sizes, and the like (Newton et al. 2010, 4).

Moreover, interrupted time-series models face issues relating to delayed causation and gradual rather than abrupt treatment interventions. Here, the effect of EE System implementation on teacher and educator behavior may diffuse slowly through a population, rather than the abrupt singular instance (implementation) catalyzing instantaneous change. Moreover, “many effects occur with unpredictable time delays that may differ among populations over time” (Shadish et al. 2002, 196). If DPI deems either (or both) is likely to occur, it may want to forgo using this model. Shadish et al. conclude that a concurrent amalgamation of each event during evaluation can make causal inference difficult because

no knowledge exists as to where to place the onset of the expected or desired effect, and effects might be expected at any point after treatment is
implemented. The longer the time after implementation, the more plausible it is to interpret a possible delayed treatment effect in terms of historical factors (2002, 198).

Finally, surveys and focus groups, and the proposed data collection methods, also present validity concerns, the most prominent of which is selection bias. Selection bias can take many forms. For instance, not everyone to whom DPI sends a survey responds, and respondents may differ systematically from the educator population as a whole. In this case, those who actually respond to surveys have not been selected randomly, and thus may have characteristics that differ from the entire population of possible survey respondents, leading to response bias (Bethlehem 2010). Similar concerns are present with focus groups. If those responsible for forming the focus groups fail to select a representative population, or if educators are able to “self-select” into the focus groups, the data focus groups yield may be misleading.

Conclusion

Here, the authors reiterate and synthesize recommendations regarding the research questions, data collection, metrics, and data analysis DPI should consider when crafting its RFP.

- DPI should create broad evaluation research questions that address implementation, educator perceptions of program experiences, the burdens of reform, changes to teacher and educator behavior, long-term student outcomes, and metric and process fidelity. Within each broad question, DPI should have clearly defined sub-questions that are measurable through collectable data.

- DPI should pursue collecting a mixture of quantitative and qualitative data through frequent surveys, and consider using stratified sampling during its evaluation of statewide implementation to choose school districts in which to conduct focus groups and interviews.

- To assess implementation, DPI should ask about:
  - educator perceptions of how the EE System is being implemented;
  - educator comfort and understanding of the system;
  - the percentage of evaluations completed;
  - the number of evaluations that each evaluator conducts;
  - whether evaluators have the resources they need to conduct evaluations;
  - how much additional time the EE System requires of evaluators.

- To assess educator outcomes, DPI should ask about educator perceptions of their own improvement and whether teachers and districts have changed any policies or practices.
• To assess student outcomes, DPI should continue to use VAM, and consider utilization of an interrupted time-series design to determine whether the EE System is associated with improvements in value-added scores.

• DPI should aggregate and publish the survey, focus group, and interview data it collects. To analyze the data, DPI should calculate descriptive statistics for variables, use t-tests to determine variable significance, and continue to employ VAM for determining student outcomes. DPI should also consider utilizing an interrupted time series regression model to analyze student outcomes before and after the implementation to determine whether implementation correlates with changes in student outcomes.

• Finally, DPI should make the data it collects available to higher education institutions such as the La Follette School of Public Affairs. Faculty and students at such institutions may be willing to conduct in-depth, independent, and cost-effective data analysis to confirm findings of DPI’s chosen evaluation firm.
Appendix A: The RFP Process

Request for Proposals (RFPs) inform vendors of the services an agency intends to procure, what requirements a vendor must meet, and what process and evaluation will take place to select the vendor. RFPs enable agencies to solicit proposals in instances where an award cannot be made strictly on specifications or price and several firms are qualified to furnish the product or service (McCain 2005). RFPs provide greater flexibility and allow for more creativity and innovation than other state procurement methods because vendors are given greater discretion in crafting their proposals. This approach to contracting in which bidders are allowed to put forth competing proposals and the procurer can take alternative approaches is more useful for complex projects. Effective RFPs typically reflect the strategy, and short- and long-term business objectives, providing detailed insight upon which suppliers will be able to offer a perspective (Mhay 2014).

Wisconsin statutes and administrative code provide state agencies with the authority to procure services through the issuance of an RFP and establish requirements with which agencies must comply. Within the Department of Administration (DOA) is the State Bureau of Procurement (SBOP) that manages contracts for products and services state agencies purchase. The bureau usually requires competitive bidding when selecting a supplier of a service that is estimated to cost more than $25,000. Furthermore, the agency issuing an RFP must submit a procurement plan to the bureau for approval.

An RFP typically contains an introduction and background, project scope, general and technical requirements, and a list of required supporting documentation. Wisconsin laws require an RFP to be written in clear, concise, and measurable terms, and include standard language, terms, and conditions established by DOA (Mhay 2014). A procurement plan must contain a score sheet, essentially a detailed rubric used in evaluating the proposals. An RFP also contains a calendar of events to inform the vendors of important deadlines such as the last day for written inquiries, notifications of revisions, any interviews or conferences that will be held, the proposal due date, notification of intent to award a contract, and the contract start date (WDOA 2005).

The agency issuing an RFP must establish an evaluation committee and provide DOA with a list of evaluation team members and their titles in order to score the proposals consistently and without bias. Typically, an RFP committee must have at least three members, including at least one person outside the procuring agency. The agency must keep a written record of all meetings, conferences, oral presentations, discussions, negotiations, and evaluations of an RFP (Mhay 2014).

After the State Bureau of Procurement approves an RFP, the issuing agency makes it available to the public and vendor community through public legal notices, including a posting on VendorNet, the state's procurement website. Agencies must also send an RFP to all minority business enterprises that are
included on a list furnished by DOA, and are required to contact appropriate labor organizations (McCain 2005). At this point, vendors review the RFP and develop proposals. The RFP development phase involves a great deal of communication between the issuing agency and the vendor. Often, vendors will submit written inquiries and the issuing agency will hold conferences and demonstrations for vendors.

Once vendors have submitted their RFP responses, the agency’s evaluation committee reviews the proposals and scores them according to the established RFP criteria. Examples of evaluation criteria include the proposer’s familiarity with doing business in Wisconsin, managerial capacity, technical capability to meet requirements, and cost estimates (WDOA 2005). After the agency determines which proposals merit further consideration, it often negotiates with those vendors on the terms of the contract prior to making an award. Following receipt of each vendor’s final offer, the agency issuing the RFP determines which proposal best meets the specified criteria and awards the contract. Once the contract is awarded, the agency must state in writing the reason for the award and place the statement, as well as supporting documentation, in a file that is open to the public.

Vendors who are not chosen may appeal the agency’s decision. To do so, they must submit a written appeal to the head of the agency or to the secretary of DOA. In order for an appeal to be considered, the vendor must show that a violation of state statute or administrative code occurred.
Appendix B: Sample Research Questions

Implementation⁴:
How was [the evaluation system] enacted in each site?
Was there fidelity to the state model?
Did [the evaluation system] create opportunities for professional growth for teachers and administrators?
To what extent were these differentiated across individuals or subgroups?
What variations occurred and what explains these variations?

Teacher and educator perceptions⁵:
What were teachers’ and administrators’ perceptions of teacher evaluation and the classroom observation rubrics being implemented?

Teacher and educator experiences⁶:
How did educators experience [the evaluation system]?
What variations occurred and what explains these variations?

Impact on teacher and administrator behavior⁷:
To what extent did participants perceive that initial implementation of the new teacher effectiveness measures resulted in:
   (a) teachers increasingly reflecting about their practice;
   (b) more instruction-focused interactions between teachers and administrators;
   (c) changes in instructional practice; and/or
   (d) teacher collaboration/knowledge sharing?

Unintended consequences⁸:
What are the unintended consequences of implementing the model?

Short- and long-term student outcomes⁹:
[What is the] impact on student outcomes?

---

⁴ Donaldson et al. (2014).
⁵ Firestone et al. (2013).
⁶ Donaldson et al. (2014).
⁸ FHI 360 (2013).
⁹ MGT of America, and Susan Zellner (2012).
Appendix C: Sample Survey and Focus Group Questions

Implementation
From Ohio’s evaluation:¹⁰

What successes or benefits have you or your school/district team experienced as a result of implementing the new teacher evaluation methods thus far?

What obstacles or unintended consequences have you or your school/district team encountered as a result of implementing the new teacher evaluation methods…thus far?

From Minnesota’s evaluation:¹¹

Briefly, what orientation training have you received in the new evaluation model?

Would you say the orientation training provided you with a clear understanding of how your district’s model will be implemented? Why or why not?

Did you feel you were prepared to do the goal-setting process? Why or why not?

Did your evaluator appear prepared to do the goal-setting process? Why or why not?

What evaluation activities took place between the goal-setting conference in the fall and the mid-year conference to help you achieve your individual and school improvement goals?

How was the…model implemented by individual schools and districts?

Were the purpose and procedures clear to…districts and schools?

What obstacles and challenges were encountered by…districts and schools and how were these obstacles and challenges addressed?

¹⁰ MGT of America, and Susan Zellner (2012).
¹¹ FHI 360 (2013).
**Teacher and Educator Perceptions**  
From Arizona’s evaluation:  

Survey questions for which teacher indicates agree/disagree/neither with the following statements:

- I have confidence in my evaluator’s ability to accurately rate my instructional practice.
- The post-observation conference(s) provided me with meaningful feedback on how to improve my instruction.
- The number of formal classroom observations was adequate to assess my performance.
- My participation in the new teacher evaluation process has benefited my students.

From Ohio’s evaluation:  

Please indicate the extent to which you feel that the following activities/measures can provide an accurate assessment of your performance as a teacher: (No accuracy/Low accuracy/Moderate accuracy/High accuracy)

- Formal classroom observation(s) by my principal.
- Student learning objectives (SLOs) for the year, established through consultation with my principal.

**Teacher and Educator Experiences**  
From New Jersey’s evaluation:

- How would you describe your role in the implementation of the new teacher evaluation system?
- How would you describe the amount of support your district received in the following dimensions from the New Jersey Department of Education during implementation of the new teacher evaluation system?
- How would you describe the level of support you received from the school district regarding the implementation of the new teacher evaluation system?

---

13 MGT of America, and Susan Zellner (2012).  
14 Firestone et al. (2013).
What is being done in your school and/or school district to ensure the optimal implementation of the new teacher evaluation system? (Please check all that apply, list numerous relevant activities)

From Connecticut’s evaluation:\(^{15}\)
Rate your agreement with the following statements:

I feel that I have received appropriate levels of recognition based on my summative rating.

I feel that I have received appropriate levels of support based on my summative rating.

**Impact on teacher and administrator behavior**
From Massachusetts’ evaluation:\(^{16}\)

Agree/ Strongly Agree/ Neutral/ Disagree/Strongly Disagree (teachers):

The new system provides me with opportunities to reflect on my practice.

The new system provides me with opportunities to improve my practice.

Five-step cycle contributes to my own growth and development.

My evaluator provides me with meaningful feedback on my teaching practice.

I engaged in the following new professional development tasks following my initial and subsequent evaluations (list options to select).

I have changed my teacher practices in the following ways after each evaluation (list options to select).

Agree/ Strongly Agree/ Neutral/ Disagree/Strongly Disagree (principals):

The new system provides me with opportunities to reflect on my practice.

The new system provides me with opportunities to grow and develop.

Through the five-step cycle, I get meaningful feedback on my practice.

This feedback chances my daily operations.

This feedback has changed how I view my job.

\(^{15}\) Donaldson et al. (2014).

\(^{16}\) SRI International (2014).
I have changed my behavior based on this feedback.

**Metric and Process Fidelity**  
From Nevada’s evaluation\(^\text{17}\):

Do the instruments consistently measure the constructs they purport to measure?

Are they being implemented with fidelity?

Do they produce ratings that are appropriate, valid, and reliable? Are the processes around their use feasible and relatively low-burden to ensure fidelity, sustainability, and ability to scale up?

---

\(^{17}\) Rabinowitz et al. (2014).
Appendix D: Sample Metrics

From Michigan’s Evaluation: 18
Median number of teachers evaluated
Median number of observation per teacher
Percent of principals reporting conferences
Percent of teachers reporting conferences
Percent of principals using growth tools in evaluations
Percent of teachers receiving mid-year report
Percent of teachers receiving end-of-year report
Percent of teachers assigned effectiveness rating median hours spent on training
Median hours observing probationary teacher
Median hours observing tenured teacher
Median hours rating/report probationary teacher
Median hours rating/report tenured teacher
Median days/year spent on teacher evaluation
Percent of principals reporting: no training, initial training only, initial training plus team meetings in district
Number of participants who have completed: initial training, team meetings,
follow-up training, individual support, and amount of missing information

From Delaware’s evaluation: 19
I am able to provide the evidence and documentation needed by my evaluator for
him/her to accurately determine my effectiveness
The evaluation process (observations, documentation, and conferences) provides
adequate evidence of my teaching
The evaluation process (observations, documentation, and conferences) provides
an accurate picture of my teaching
Overall, the evaluation process is implemented consistently at my school
Overall, the evaluation process is implemented appropriately at my school
Are the content, materials, timelines, and delivery methods appropriate and
effective?
In general, my evaluator understands the nuts and bolts of the student
improvement component
My district ensures that the evaluation system is implemented consistently
My district ensures that the evaluation system is implemented fairly
My district ensures that the evaluation system is implemented as intended
The state ensures that the evaluation system is implemented consistently
The state ensures that the evaluation system is implemented fairly
The state ensures that the evaluation system is implemented as intended

18 Rowan et al. (2013).
19 Beers (2013)
Works Cited


