Developing and Validating the Next Generation of Leadership Evaluation Tools: Formative Assessment for High-Stakes Accountability

IES Education Leadership Research Grants, Goal 5

Project Narrative

1.0 Significance

Our proposal aims to develop and validate the next generation of on-line, formative assessment tools necessary for middle and high schools to establish the conditions for improving student learning. In today's high-stakes accountability world, school and district leaders need guidance to design the kinds of school environments that improve student learning. Elmore (2002) explained that instructional leadership had long been dominated by models of professional autonomy in which leaders provided adequate organizational support and professional guidance for teachers to flourish in loosely coupled school organizations. High-stakes accountability in general, and the No Child Left Behind Act of 2001 (NCLB) in particular, has changed the expectations for school leaders. Leaders must now tighten the connections between classroom practices and school-wide outcome data to build the capacity for intentionally improving student learning across the school. Our tool, called the School Leadership Assessment Tool system, will provide leaders with a set of formative rubrics that can be used by middle and high schools to self-evaluate and to guide the development of critical leadership practices. While the standards-based, theory-driven assessments now in place measure where leadership currently is, formative feedback systems are needed to help school leaders understand the gap between where they are and where they need to be.

Meeting accountability standards has proven particularly challenging for middle and high school leaders. While elementary school leaders have provided most of what we know about sustainable school improvement (Hargreaves et al, 2007), many middle and high school leaders struggle with structural reform programs that leave the instructional cultures of their schools untouched (for review, see Hargreaves et al, 2007; McLaughlin & Talbert, 2001; Siskin, 1994). Middle and secondary school leaders need to build high expectations for all students, and create sustainable connections between classroom teachers across departments, and among regular and special educators, student services providers, and instructional support teachers to make sure that teaching and support services are aligned and targeted for all students to succeed. Elmore (2002) concluded that many school leaders simply lack the relevant knowledge of *how* to shift schools from loosely-coupled to accountability-driven organizations. Although elementary school leaders in an increasing number of schools have now developed practices that can predictably improve student learning outcomes, researchers and practitioners continue to struggle with how to translate these hard-won organizational successes to the middle and high school levels (c.f. Lachat, 2001; Lachat & Smith, 2005; NASSP, 2006).

We base our argument on the recent blossoming of research designed to show why and how school leadership matters for teaching and learning. Our review begins with a discussion of the correlational research that shows *that* leadership matters for student learning. We then review recent research that demonstrates *how* leadership influences student learning by describing a range of leadership functions that improve the conditions of teaching and learning in schools.

These studies have led to the design of leadership standards (e.g. Interstate School Leaders Licensure Consortium [ISLLC]) and tools for assessing leadership. We discuss how assessments such as the Vanderbilt Assessment of Leadership in Education (VAL-Ed) and the Rand Corporation's Leadership Performance Planning Worksheet (LPPW) provide a profile of leadership practice in a given school in terms of the identified leadership functions.

We argue that research on sensemaking and social cognition demonstrates the necessary but insufficient nature of categorizing leadership practices in terms of standards as a mechanism for advancing leadership practice. If we consider standards-based leadership assessment from the perspective of leaders as *learners*, it is clear that assessment systems must provide clearer guidance for leaders to move to the next level of performance. In other words, to provide school leaders with the tools sufficient to advance student learning, formative leadership assessment systems must be developed to complement the standards-based summative assessment systems now being used in many schools. These formative assessments provide feedback about current practice, and clear guidance on what steps leaders need to take to strengthen and advance leadership to improve student learning. Without this guidance, many school leaders may simply interpret the results of the new standards-based assessments in terms of existing practices, and miss the critical moves necessary to improve student learning. Our rationale, then, concludes with our plan to build on the emerging knowledge base of leadership for learning to design, implement and validate a *formative* assessment system for school leadership.

1.1 Leadership matters

We know *that* school leaders play a critical and measurable role in shaping school effectiveness (Hallinger & Heck, 1998; Leithwood & Louis, 2004; Leithwood & Riehl, 2003; Mortimore, 1993; Scheurich, 1998; Waters, Marzano & McNulty, 2003). Although modeling school effects on student learning leaves a significant share of student learning unexplained, about one quarter of the total school effects can be attributed to principal leadership (Hallinger & Heck, 1998; Leithwood & Louis, 2004). Research suggests that leadership behaviors are second only to teacher effects in their impact on student learning.

We know something about the leadership *functions* that contribute to improving student learning. In a comprehensive review, Murphy, Elliot, Goldring and Porter (2006) describe the several functions of learning-centered leadership including establishing a shared vision for learning that establishes high standards for students (e.g. Dwyer, 1986; Newmann, 1997; Bryk & Schneider, 2002); leading the instructional, curricular and assessment programs (e.g. Marzano, Waters, & McNulty, 2005; Murphy & Hallinger, 1985), developing strong communities of learning (e.g. Berman & McLaughlin, 1978; Little, 1982; Bryk & Schneider, 2002); effectively acquiring and allocating resources (Odden & Archibald, 2001; Beck & Murphy, 1996) maintaining a strong organizational culture focused on learning (e.g. Bossert, Dwyer, Rowan, & Lee, 1982; Louis, Kruse and Bryk 1995; Leithwood, Steinbach & Jantzi, 2002) and engaging in social advocacy (e.g. Fullan, 2003; Goldring & Sullivan, 1996; Moll, 1992). These functions indicate the goals towards which leaders should work to improve student learning.

Researchers have also identified the critical *organizational characteristics* that leadership functions must establish to improve the conditions for teaching and learning. Leithwood and Louis (2004), for example, suggest that leadership practice for improving learning involves

transformational practices such as setting new directions, professional development, and organizational redesign. Others have identified such important organizational characteristics that leaders must facilitate such as establishing relational trust (Bryk & Schneider, 2002); shaping school culture (Deal & Peterson, 2003); providing instructional leadership (Fullan, Hill & Crevola, 2006; Hallinger, 2000); supporting systems of distributed leadership (Spillane, 2006; Gronn, 2002), building inclusive service delivery systems for students who struggle (Capper & Frattura, 2007) and actively engaging the community to help address critical context variables that shape student outcomes (Rothstein, 2004; Warren, 2005). Because learning is social and based on prior experience (Brown, Collins & Duguid, 1989), the most effective school leaders support teacher learning by developing *communities of practice* for teachers to problem-solve, share best practices, and learn as a community (Wenger, 1998, Kelley & Shaw, forthcoming). Effective learning environments are built on communities of practice that are knowledge centered, learner centered, and assessment centered (Bransford, Brown & Cotting, 1999). This research provides a knowledge base that identifies a set of leadership functions and organizational characteristics that contribute to higher levels of student learning. Linking leadership functions with these key organizational outcomes serves as the basis for a rich, research-based map that can guide the practical work of school leaders.

1.2 Standards-based school leadership assessment

Recent work in policy development has translated these research efforts into new standards for school leadership practice. The Wallace Foundation, in particular, has made a significant investment in strengthening educational leadership as a tool for school improvement and has reinvigorated efforts to develop meaningful standards and assessments to support the development of school leaders and hold them accountable for leadership behaviors that close achievement gaps and advance learning for all students. Many states (e.g. Arkansas, Iowa, Maryland, Wisconsin) have developed standards and assessments for educational leaders from initial to master levels of practice.

The ISLLC standards provide perhaps the most widely known and used leadership standards. The standards are used to guide preparation program design, state licensure policies, and professional development for educational leaders. In 2008, the Council of Chief State School Officers published the Educational Leadership Policy Standards to elaborate the ways in which the ISLLC standards could advance leadership in the current high-stakes accountability policy environment. The standards include:

- Setting widely shared vision for learning;
- Developing a school culture and instructional program conducive to student learning and staff professional growth;
- Ensuring effective management of the organization, operation, and resources for a safe, efficient, and effective learning environment;
- Collaborating with faculty and community members, responding to diverse community interests and needs, and mobilizing community resources;

- Acting with integrity, fairness, and in an ethical manner; and

- Understanding, responding to, and influencing the political, social, legal, and cultural context (CCSSO, 2008).

While the ISLLC standards have been widely used to design leadership evaluation efforts across the country, recently there have been several notable assessment efforts to specify the relation of leadership functions with the conditions for improving student learning. VAL-Ed,[1] for example, is a theory-based assessment tool designed to measure principal behaviors. VAL-Ed aims to provide a 360 degree perspective on leadership behaviors in order to produce a quantitative diagnostic profile of instructional leadership at the school level, with scores on scales and subscales across a six-by-six grid of core components of leadership practice on one axis and the key processes related to instructional leadership on the other. The core components include: high standards for student learning, rigorous curriculum, quality instruction, culture of learning and professional behavior, connections to external communities, and performance accountability. The key processes include: planning, implementation, supporting, advocating, communicating, and monitoring. Each cell in the grid then corresponds to the specific practices expected of school leaders that meet the standards. Although the developers of VAL-Ed refer to it as a summative and formative assessment instrument, it is primarily a summative assessment instrument that measures the quality of specific leadership practice in terms of theory-based standards (Porter et al., 2006).

The Rand Corporation's Leadership Performance Planning Worksheet provides a different approach to theory-driven assessment. The LPPW structures the induction experiences of novice school leaders by guiding conversations and reflections with novice principals and their mentors/coaches on key leadership performance areas (Scott, 2008). Building on existing state and national instruments, the tool identifies nine leadership dimensions important for new leaders: personal behavior, resilience, communication and the context of learning, student performance, situational problem solving, learning, supervision of instructional and non-instructional staff, management, and technology. The tool is being used by eight states and a number of preparation programs to support leadership development for new principals.

Efforts such as ISLLC, VAL-Ed, and LPPW illustrate how researchers have operationalized leadership factors that lead to student learning. They have provided important contributions to defining, assessing, and supporting leaders in advancing student learning. Together, these efforts help to identify the characteristics of leadership in a given school that are known to support student learning. They do not, however, focus on providing practicing principals with formative tools to self-assess leadership behaviors and to identify steps that school leaders and teachers across experience levels can take to improve learning in their schools. The next generation of leadership evaluation tools will need formative elements that can act as a roadmap to tell schools where they are as well as where they need to go. This will provide much needed guidance to strengthen the work of principals and their leadership teams in advancing leadership for learning.

---

[1] www.vanderbilt.edu/lsi/valed/index.html

1.3 Why formative assessment matters

Formative assessment provides information crucial for modifying the thinking or behavior of the learner toward intended outcomes (Shute, 2007). Most behaviors generate information. Such information becomes formative feedback when an actor uses it to correct or confirm the initial behavior. Formative feedback is often collected and shared in terms of scaffolded learning environments that provide proximal challenges to guide learners through complex tasks (Collins, Brown & Newman, 1989). In education, formative feedback is typically studied in terms of motivating or directing student learning (Brophy, 1981; Schwartz & White, 2000; Black & Wiliam, 1999). However, accurate formative feedback is equally important for guiding adult learning. Spillane, Reiser and Reimer (2002) describe the conservative pull of existing professional knowledge and skill exercised as school leaders and teachers make sense of new school curricula and policies. Educators, like other professionals, tend to read new information in terms of what they already know, and often miss the salient features of a policy or a program designed to spark new practice (see, for example, Spillane 1998; Coburn, 2005, or Halverson & Clifford, 2006).

However, the high-stakes accountability policy context requires leaders to stretch what they already know in order to build new professional interaction and strengthen student learning environments (Elmore, 2002). While the standards-based, theory-driven assessments we discussed above show where leadership currently is, formative feedback is needed to help school leaders understand the gap between where they are and where they need to be. Since each schools and districts face different problems in terms of things like student mobility, teacher turnover, poverty status and parental involvement, each school leadership team must interpret how to apply relevant standards in unique combinations to their local context. Formative assessment tools will allow school leaders to measure local practices in terms of a developmental arc determined by national standards. Many school districts have already implemented benchmark assessment systems for student learning, such as the Northwest Evaluation Association's Measures of Academic Progress (MAP) program, to provide formative information about how students are progressing through the school instructional program. We argue that school leaders need similar kinds of information, scaffolded in terms of the key tasks of leadership, to create the research-driven conditions for improving student learning so well described in current assessment systems.

A distributed leadership perspective suggests that *tasks* provide the most appropriate level for analyzing school leadership (Spillane, Halverson & Diamond, 2004). Leadership tasks have several advantages for the development of a formative feedback system. First, tasks represent the ways that practitioners organize their work, thus providing a link between theory-driven expectations and everyday practice. Second, tasks exist on two levels. Macro-tasks indicate the general leadership functions (described in the leadership effects research described above) towards which practice aims, and micro-tasks describe the ways in which everyday work is organized within the macro-tasks. Finally, tasks allow for the focus of analysis to shift from actors to actions. Instead of narrowly defining the allocation of task responsibilities to specific positional leadership roles (e.g., principals or department chairs), a task-based approach allows for contextual differences in the allocation of leadership responsibilities within the school. This perspective recognizes that multiple and varying actors are involved in leadership activities

across schools, and provides a more context-sensitive translation of theory to practice in a formative assessment instrument.

2.0 Research plan

The proposed School Leadership Assessment Tools (SLAT) will provide an on-line formative assessment system based on a rubric that will allow teams of school leaders and teachers to assess themselves in terms of core leadership tasks and to receive feedback that will scaffold efforts to improve local practices. Our proposal is grounded in a distributed leadership perspective that focuses on identifying and evaluating the leadership tasks necessary to improve learning in schools. SLAT will focus on leadership tasks rather than leadership roles in order to draw the focus of the assessment away from summative judgment of positional leaders and toward measuring (and understanding) the kinds of work necessary to improve student learning. The resulting SLAT reports can then be used as planning documents to help schools determine which tasks will be necessary to improve leadership for learning and to assign who will be responsible for conducting these tasks.

The SLAT design will build on existing leadership standards and summative assessment tools, and will draw on two recent successful rubric development projects to assess and support school leadership. The initial content for SLAT will be provided by two prior rubric-based evaluation systems developed by the proposal Primary Investigators: Halverson's *School Leadership Rubrics* and Kelley's *Socio-Cognitive Leadership Rubrics (both found in Appendix B)*. Both rubrics have face validity: Halverson's is being used in 15 large urban school districts and Kelley's is currently being used in over 50 schools.

The *School Leadership Rubrics,* developed by Richard Halverson in collaboration with the Institute for Learning at the University of Pittsburgh, identify the specific tasks that school leaders should pursue to advance student learning. These rubrics organize school leadership into five central tasks:

- Maintaining a focus on learning;
- Monitoring teaching and learning;
- Building a nested learning community;
- Acquiring and allocating resources; and
- Maintaining a safe learning environment.

The rubrics were initially developed in 2004, and have been used as a core professional development tool for districts including the Dallas, Austin and Minneapolis Public Schools and the Los Angeles Unified School District.

The *Socio-Cognitive Leadership Rubrics*, developed by Carolyn Kelley and Jim Shaw in conjunction with their work with the Wallace Foundation, define where effective school leaders focus their attention and describe a shared cognitive decision-making approach used in effective

schools. These rubrics define a cognitive decision-making model that is evidence-based and shared, and is applied to four dimensions of leadership for learning:

- Advancing equity and excellence in student learning;
- Developing teacher capacity;
- Managing and aligning resources; and
- Building and engaging community.

They have been used as a reflective tool by 50 educational leaders in districts throughout the state of Wisconsin, and have been triangulated with the effective leadership practices of schools and districts across the country that have successfully closed achievement gaps and improved learning for all students.

The dimensions and elements of each set of rubrics will provide the initial content for SLAT. The School Leadership Rubrics provide a model for breaking down the elements into specific leadership tasks and for how to articulate tasks across quality dimensions to provide the formative feedback context. The Socio-Cognitive Leadership Rubrics provide the reflective questions that will spark the elicitation of evidence needed to justify a particular rating (e.g., How have you worked to build teacher capacity to meet student needs and raise student achievement? How have you worked to support all teachers to grow professionally and engage in reflective practice (including teachers who struggle)?). The re-design process we propose will combine elements of the two models, and draw on the leadership functions and standards built into the VAL-Ed, ISLLC and LPPW models. Because SLAT will provide specific task descriptions on how these leadership practices in sub-areas (such as student services, school safety, data-driven instruction, and subject-matter based instructional leadership) are articulated across quality dimensions, the rubric design process will be informed by rigorous literature reviews in each of these critical areas of school practice.

SLAT will be designed around a task-based rubric for measuring leadership practice. The rubric will be developed around six key *dimensions* of leadership practice as suggested by research on leadership practice. Each dimension will be divided into four to six *elements* that describe the salient aspects of the leadership practice; and each element will be further subdivided into five to seven leadership *tasks* that describe the activities that correspond to each element. The task descriptions will be articulated across three quality levels: needs attention, proficient, and exemplary, in order to provide a basis for assessing current practice as well as an indicator of how leaders could think about improvement within a particular task. SLAT will have a hierarchical structure of task items within elements, element scales within dimensions. Scores will be determined at the task level – element scale scores will consist of averages of task scores, and dimension scale scores will be calculated as averages of element scale ratings. The School Leadership Rubrics provide an example of the rubric design in the area of "collaborative design of integrated learning plan." (Figure 1) This example shows how the presence or absence of practice-level tasks provide evidence for each element at the practitioner level, and how the task descriptions are articulated across qualitative markers.

SLAT will be delivered in a Web-based assessment system. The Web-based system will be developed and launched by Web- and database programmers at the Wisconsin Department of Information Technology (DoIT) and the Wisconsin Center for Education Research (WCER), with assistance from the Learning Point Associates and the Consortium for Policy Research in Education (CPRE) research teams. The system will be developed to enable user-testing and validation operations (described below). For example, the system will collect demographic information about raters, and guide raters sequentially through the evaluation process. The system interface will include prompts about the kinds of evidence appropriate to reflect upon for assessing each leadership task. The Web interface will allow raters to zoom in on the specific tasks of an individual element, or to scan the developmental sequence across tasks to determine the appropriate rating. The Web tool will be connected to a database to record information about how users navigate the system as well as to collect all rating evidence. Data will then be reported to researchers in terms of individual and collective ratings within tasks, within elements, within dimensions and across the evaluation system as a whole to determine aggregate ratings. Data will be reported to raters in terms of the school's existing leadership levels and task-based indicators for subsequent action. The Web-based system should structure the entire evaluation process into a 30-minute timeframe in order to enable raters to complete the process in a reasonable time. School teams will be able to return to the system as a tool to facilitate professional development and school planning.

1.2 Collaborative design of integrated learning plan

The school-wide focus on learning needs to be operationalized into a viable shared learning plan. Teachers and teaching staff need to participate in the development and review of this plan, and to understand how the plan informs daily instructional practice. School leaders set up the structures to develop, implement and review the learning plan.

Needs Attention	Proficient	Exemplary
Teachers are left to their own devices to come up with instructional strategies. Strategies to improve student academic performance are rarely discussed at faculty meetings. School-wide planning for instruction is either not done or is an exercise that exists apart from the actual instructional practices of the school.	Teachers and leaders work together to refine and develop instructional strategies. The school has developed a structured, collective instructional planning process that coordinates specific instructional initiatives toward overall goals of student achievement. Strategies to improve student academic performance are discussed at faculty meetings. The school plan reflects the priorities the district learning priorities.	Strategies to improve student academic performance are the regular focus of faculty meetings. The school has developed a structured, collective instructional planning process that uses student achievement data to coordinate specific instructional initiatives toward overall goals of student achievement. The plan integrates intermittent measures of student progress toward learning goals. The school plan is well-integrated with the district learning plan.

Figure 1: **Sample item from the School Leadership Rubrics.**

The SLAT developers will work draw on the WCER Technical Services application development team's extensive experience in both open and closed source databased-backed web applications. We propose developing the SLAT infrastructure on a Microsoft Windows stack:

- Server 2008
- IIS 7.0
- .Net 3.0
- SQL Server 2008

The interactive client-side components will be developed to use either Adobe *Flash* or Microsoft *Silverlight* depending on the exact feature needs set generated by the development process. The system will be developed to be platform neutral and browser independent. This set of technologies is supported by all participants on the project and will enable all members to engage on the design and development process. This approach would allow all members to host and test these technologies. Indeed, most educational agencies would be able to support this standard line-of-business combination of technologies.

In addition, the development team can draw on extensive WCER experience in designing, building, and supporting data-rich, real-time applications in educational settings. Currently deployed applications include test-to-standard & standard-to-standard alignment tools, large, branching survey instruments, state-wide data collection and validation tools as well as web-based collaboration and instructional environments. Our user-focused design process are informed by the latest research on development practices and well-aligned with the practices proposed by the research team.

2.1 SLAT collaborative design research plan

The SLAT collaborative design process will (a) develop formative assessment tools that measure leadership quality in terms of the world in which leaders work, and (b) conduct a coordinated set of seven studies that provide evidence for the validity of inferences based on SLAT. The work proposed falls into two main categories: the design, implementation, and iterative redesign of the SLAT Web-based system; and the proposed studies to establish the SLAT validity. First we will describe the assembly of a collaborative design team that will be responsible for the rubric and Web-system development, then detail the nine empirical studies we propose to conduct to guide the development and validation processes. Though SLAT will be developed as a formative assessment tool, it is still important to use the kinds of analysis appropriate to determine the quality of summative evaluations in the validation process in order to demonstrate (1) that self-assessors can and do use the tool to make assessments that are based on the actual features of the situation, and not simply on the assessor's idiosyncratic interpretation of the evidence and the rating scales or rubrics, and (2) that the instrument itself is related to desired leadership outcomes, such as enhanced leadership processes, school climate, and student learning.

Our overall SLAT development model is guided by core concepts of collaborative design (c.f. Edelson, 2002; Danesi, Gardan & Gardan, 2006). Collaborative design processes involve teams of researchers, practitioners, and designers in efforts to build tools that can be better implemented in contexts of practice. In the SLAT design, researchers will have the primary responsibility for bringing together the ideas that guide rubric construction and validation study design (described below); designers will be responsible for developing the Web-based assessment system; and practitioners will contribute to describing tasks across quality dimensions to ensure feedback that provides clear guidance for leadership development and

sensitivity to variations in school context. The SLAT collaborative design team will include investigators described in section 5.0, and a group of master practitioners nominated by the Association of Wisconsin School Administrators (AWSA) (Letter of Agreement in Appendix A). Nominees will include middle and high school principals who have completed the Wisconsin Master Educator Assessment Process (WMEAP), a rigorous assessment process that is recognized by the Wisconsin Department of Public Instruction. The SLAT design process will involve two sub-groups of practitioners, four for the Middle School team, and four for the High School team.

The work of the collaborative design process will fall into two phases: Phase 1: *SLAT Development* (Years 1-2), and Phase 2: *SLAT Validation* (Years 3-4).

In Phase 1, the collaborative design teams will critically review the existing rubric sets, to examine the task descriptions and articulations appropriate to middle and high school contexts, and to suggest revisions. The results of the five studies (described in Section 4.2) conducted in Phase 1 will inform the design process. The design work will result in a new rubric with revised elements and task descriptions that address the design team's best sense of critical practices. In Year 2, the collaborative design teams will conduct a pilot implementation of the Web-based SLAT system with middle and high school leaders in the Racine (WI) school district (Letter of Agreement in Appendix A). Raters will use paper-based and Web-based versions of SLAT to determine the degree to which the medium influences rating decisions. The team will meet to review the rubrics and the Web-based evaluation system before and after pilot tests (Year 2). An executive committee of Primary Investigators Halverson and Kelley will oversee and coordinate the development teams; Investigator Clifford and researchers from Learning Point will lead the Phase 1 validation studies.

In Phase 2, Primary Investigators Halverson and Kelley will coordinate a multi-district effort to obtain reliability and validity data on the SLAT system. We plan to work with middle and high schools in several medium-to-large sized districts including Madison, WI, Fairfax County, VA and the El Paso (TX) Public School systems (Letters of Agreement in Appendix A). Investigators Milanowski and Kimball will lead the design and implementation of the three validation studies (described in Section 4.3). During the validation phases of Years 3-4, Primary Investigators Halverson and Kelley will work to integrate new information arising from implementation back into the system design.

The SLAT collaborative design team will rely on information gathered during nine studies that will take place across the development and validation phases. The team will continue to meet throughout the project to review data generated and to integrate this information to inform system design. Table 1 summarizes these studies and the following sections provide detail about each study.

Table 1
Timeline of Studies

	Year
Phase 1: SLAT development	1 & 2
Study 1: Procedure for reviewing constructs	1
Study 2: Procedure for selecting items	1

	Year
Study 3: Procedure for examining content validity	1
Study 4: User-testing	2
Study 5: Item distribution analyses	2
Phase 2: SLAT validation	3 & 4
Study 6: Inter-assessor agreement assessment ratings	3 & 4
Study 7: Relationship of ratings with other indicators of leadership quality	3 & 4
Study 8: Construct validity of formative assessment ratings	3 & 4
Study 9: Consequential validity	3 & 4

2.2 Phase 1: SLAT development (Years 1-2)

Study 1: Procedure for reviewing constructs to be "tapped" by the instrument (Year 1)

SLAT construct definition will be a synthetic process that draws upon pre-existing evaluation instruments, interviews with school leaders, and well-founded theoretical work completed by members of the development team. The collaborative design process will be supplemented and its results tested by a procedure for reviewing the constructs tapped by the instrument. First, we will have a small group of school leadership experts and school leaders not involved in the collaborative development review the definitions of the SLAT dimensions and elements (subdimensions) within the dimensions. We will ask this group to first assign elements to dimensions, then identify aspects of the dimensions and elements that overlap, identify unclear aspects of the constructs, and generate aspects of school leadership that they perceive to be missing from the tools. This review will serve as a check that we have defined clear and coherent constructs that do not miss major aspects of instructional leadership.

Study 2: Item selection procedure (Year 1)

The initial review of the items to be included in the operational SLAT tool will start with items from the School Leadership and Socio-Cognitive rubrics, examples from other rubrics as identified through the literature review, and additional items written by the development team. The scale format that we will use to articulate the tasks across element quality areas will be a 3-point range for each school leadership task as used in the Halverson and Kelley rubrics. When a sufficient number of items describing tasks have been developed, investigators Clifford and Condon, with assistance from Milanowski, will conduct an initial content validity analyses using a simplified version of Lawshe's (1975) method, to ensure that the items of each tool are representative of the constructs of school leadership the tools are attempting to measure. Using this method, the degree of content validity is assessed based on the extent to which members of a small expert panel (consisting of the design team and 2-3 other experts in educational leadership) perceive overlap between the item and the definition of the element it is intended to represent. A content validity index will be calculated based on panel members' judgments of items' relevance to the element construct. Items with low content validity indices will be removed. If needed, other items will be developed to replace them. We will aim to have 5-7 task items for each element at the end of this stage. We will follow Loevinger's (1957) recommendation that the

number of items per subscale (dimension or element) should equate to the importance of that subscale in measuring the overall construct.

Study 3: Content validity (Year 2)

Content validity studies address the relevance and representativeness of the content upon which the items are based and the technical quality of the items (Messick, 1995; AERA/APA/NCME Standards, 1999; Wilson, 2005), and the content validity study occurs after items have been drafted. Content validity is basically a measure of agreement among informed respondents that the content is salient for the assessment purpose (Lawshe, 1975). The SLAT tools will be developed in accordance with processes for establishing content validity (Wolfe & Smith, 2007; Grant & Davis, 1997; Rubio, Berg-Weber, Tebb, Lee & Rausch, 2003). Conducting a content validity study involves four steps, including: (a) determining who will review the instrument, (b) preparing reviewers for the content validity study, (c) creating a content validity survey, and (d) analyzing measures to determine instrument validity (Grant & Davis, 1997; Rubio, Berg-Weber, Tebb, Lee & Rausch, 2003).

The status of school improvement and leadership practice research suggests expert review should include content experts on school improvement and school leadership, master principals, and measurement specialists (Gable & Wolf, 1993). The content validity study will convene a panel of twenty content experts nominated by members of AWSA, and ten additional content experts in the specific task areas (e.g. instructional leadership, special education, student services, subject matter, and school finance). After an orientation to the content validity process, each expert will be asked to complete a content validity survey, which contains each school improvement diagnostic item. The domains of the content validity survey are included in Table 2 below. The survey will also require respondents to explain ratings and comment on the instrument.

Table 2
Content Validity Domain Rating Scales

	Clarity	**Representativeness**	**Factors**	**Comprehensiveness**
Item sample	• Item is not clear • Item needs major revisions to be clear • Item needs minor revisions to be clear • Item is clear	• Item is not representative • Item needs major revisions to be representative • Item needs minor revisions to be representative • Item is representative	• Item is not a factor • Item needs major revisions • Item needs minor revisions • Item is representative.	• Item should be deleted • Item should be retained

The analysis plan involves computation of three measures (see Rubio, et al, 2003). First, inter-rater agreement determines the extent to which experts were reliable in ratings. The four-point scale for Clarity and Representativeness will be converted to a dichotomous scale, where ratings 1 and 2 are equal to 1 and ratings 3 and 4 are equal to 2. The number of items associated with 1 and 2, and the number of items that are 100% Representative will be divided by the total

number of items. According to Lynn (1986), inter-rater agreement between .7 and .8 for the survey in its entirety are optimal, given that the number of raters exceeds five. Second, content validity will be calculated per item counting the number of panelists rating the item at a 3 or 4 for Representativeness, then dividing that number by the total number of panelists. Should item analysis result in a value less than .8, the item will be revised. Third, a factorial validity index will be calculated to determine the degree to which panelists correctly associate items with factors. The calculation counts the number of panelists correctly associating items with the factor, and then dividing that number by the total number of panelists. The average across all items will be calculated to determine the overall score. We expect the item analysis quotient to exceed .8, but if not, the item will be revised. If the total survey quotient falls below .8, the tool will be revised. If significant revisions should occur, the process will repeat until necessary agreement is achieved.

Study 4: User-testing (Year 2)

Determining SLAT administrative procedures will involve pilot testing the SLAT system with 3 middle and 2 high schools in a school districts nominated by our AWSA colleagues. We will pilot the SLAT system with 3-4 formal school leaders (e.g. school principals, assistant principals, department chairs, or guidance directors) and 6-8 teachers in each school. Data will be generated from the pilot to test and refine the SLAT rubrics and to test the usability of the Web-based system. A user testing process ascertains the utility of products, particularly computer programs, prior to scaled product development, and user testing is vital in product development because users are less likely to employ tools when more beneficial and easier tools are readily available (Kuniavsky, 2003; Nielsen, 2002). We are proposing two types of user testing, cognitive walk-throughs and reflective interviews to ensure SLAT is engaging and practical in the contexts of its intended use in the pilot sites.

- *Cognitive walk-throughs*. User testing in *Step Two* of the design research process asks school administrators at pilot schools to interact with draft SLAT documents and technology during simulations. As users interact with SLAT, we will request that they "think aloud" as they attend to case data and interact with instrument content. Cognitive walk-throughs (also called think-aloud) interviews are commonly employed in instrument and computer interface development (Ericsson & Simon, 1993; Dumas, 2002). We will conduct cognitive walk-throughs with raters (3 leaders and 3 teachers) in a middle school and a high school in the pilot study. Once the SLAT system is developed (Year 3), we will use cognitive interviews again with another, larger sample of school principals to establish construct validity (see below *4.3 - Study 8*).

- *Reflective interviews.* Reflective interviews involve videotaping users as they interact with the system, then asking them to reflect on their decision-making processes in a post-observation interview (Kuniavsky, 2003). These interviews will provide reflective opportunities to discuss initial interactions and hypothesize about SLAT use in a real-world context. Guiding questions would address topics such as: How relevant and appropriate is the SLAT language for prospective users?; To what extent do the principals use SLAT as intended? What, if any, aspects are used, adapted, or ignored during the simulation?; To what extent do tools assist principals to set problems that can be addressed in their, or the school's,

practices?; and How, if at all, would administrators and prospective principals improve the content or form of SLAT?

Analysis of the user testing process will focus on the types, amount of time, and use of SLAT and case material used by the principals during the user test session. The descriptive analyses will provide the collaborative design team with usability data on the SLAT system.

Study 5: Item distribution analyses (Year 2)

The five-school pilot-study results will enable us to inspect item distribution (Clark & Watson, 1995). We will eliminate items that are highly skewed or unbalanced, which indicate items for which almost everyone chooses Needs Attention or Exemplary, for instance. We will retain items that have good spread, or variability, among the sample taking the measure. We anticipate, however, that some challenging aspects of leadership practice need to be strengthened in many middle and high schools, and we will review items identified in the distributional analysis to determine whether items rank consistently low because of poor item design, or if items rank low because they reflect areas of consistently low performance across schools.

- *Dimensionality.* We will also investigate the degree to which the items developed for each SLAT subscale are related, as expected if items tap a common construct. We will first examine the range and pattern of interitem correlations within and across elements and dimensions. We would like to see correlations within dimensions range from .25 to .50 and cluster around the mean correlation, and that correlations across dimensions would be less than .25. It is important to note that we will not retain highly intercorrelated items in the final scales because they are redundant with other items and will not offer new information. Exploratory factor analyses will also help us determine the dimensionality of the items and get a preliminary indication of whether the items assigned to subdimensions within the instrument seem to be tapping them. Although we realize that with only approximately 50 raters, clustered in 5 schools, our analyses will provide limited insight, still we feel as though items that load weakly on the expected dimension (factor) are will make good candidates for deletion from the scale that we will test with a wider collection of schools in the Phase 2 research.

- *Subscales.* We will analyze correlation matrices and do exploratory factor analyses within dimensions to help substantiate that separate element-level scales exist within the instrument. Element level scales only make sense if the intrasubscale item (task) correlations are higher than the intersubscale item correlations. We expect intersubscale correlations to be significantly greater than zero but clearly less than the intrasubscale correlations (e.g., target for correlations of .20).

- *Internal Consistency Reliability.* In addition to validity, we must also examine internal consistency reliability in developing our tools. Internal consistency is a necessary but not sufficient condition for homogeneity, or unidimensionality. The tools need to show adequate internal consistency at the element and dimension levels, which we will examine using Cronbach's alpha. We will follow Nunnally's (1978) recommendation that scale reliability should be at least .80 for the element-level scales. Broader dimension-level scales would be

expected to have about the same level of alpha, with the larger number of items counteracting the greater diversity of item content.

- *Inter-rater Agreement.* We will also examine the agreement among the pilot test raters. We will do this by calculating the percent absolute agreement and the standard deviation of the average ratings made on each element (sub-dimension) for each school. Our goal would be to have 75% absolute agreement and an average standard deviation of .5 or less, which is one-half the distance between the three level integer–denominated scales of the rubrics. Because each school team rates one target (leadership practices in that school), and because the specific level of the rating is important, correlational measures of agreement are not useful here.

After the pilot test, we will take stock of the measurement and usability properties of the assessment and make modifications as needed. We will undoubtedly find that at least a few items will have to be deleted and perhaps some new items written. We may also have to modify presentation formats and instructions to improve clarity and usability. These activities will be performed by the collaborative design team in Year 2.

2.3 Phase 2: SLAT validation (Years 3-4)

We propose to obtain evidence of the reliability and validity of the SLAT formative assessment system in middle and high schools in four urban school districts from across the country (Madison, WI; Racine, WI; El Paso, TX; and Fairfax Co. VA) and in several smaller rural and suburban districts that will be nominated by our colleagues at AWSA. To ensure consistency across schools, we will recruit teams of raters in each school to include the principal, an assistant principal for instruction, an assistant principal for discipline/dean, the department chair/lead teacher for English/Language Arts, the department chair/lead teacher for math, and the leaders of the guidance/student services department. We will also randomly pick 6 teachers in each school to complete the assessment. Each rater will complete the assessment process individually. The SLAT validation process will involve four studies based on the collected data: evidence that different assessors agree on ratings of performance (Study 6), evidence that the ratings are measuring the performance dimensions or constructs they are intended to measure (Study 7), evidence that the assessment ratings are related to other indicators of school or leader performance (Study 8), and evidence that implementation of the assessment is related to changes in leadership practice (Study 9).

Though we will be attempting to "build in" validity of the SLAT tools during the development phase of the project, and will be assessing aspects of validity during that phase, in the third year we will conduct a series of more extensive studies to evaluate the tools by assessing the validity of the judgments made using the tools. This evaluation effort will emphasize the collection of four types of evidence: evidence of rater agreement on SLAT ratings, evidence that the judgments are measuring the performance dimensions or constructs they are intended to measure, evidence that the judgments of leadership performance are related to other indicators of school or leader performance, and evidence that the use of the tools has some of the consequences we intend for their formative use. Our approach to quantitative analyses will generally be to collect validity evidence within each district, treating such evidence as the result of a single validity study. We will then combine results across districts to get a more

reliable estimate of the relationship using meta-analytic procedures. The approach will accommodate the use of the different alternative indicators of school or leader performance (e.g., student achievement tests) that are likely to be used by the participating districts. Because we will be working with four districts in this effort, we expect to have data from a fairly large sample of schools and school leadership teams. Table 3 below shows the number of schools at each level in each of the urban districts in our study (these numbers will be supplemented by a sample of rural and suburban districts as nominated by AWSA).

Table 3
Number of Urban Schools Within Study Districts by School Level

District	Middle	High	Total
El Paso, Texas	17	13	30
Fairfax County, Virginia	25	25	50
Madison, Wisconsin	14	4	18
Racine, Wisconsin	7	4	11
Other smaller districts	TBD	TBD	TBD
Combined:	63	46	109

While collecting validity evidence in multiple school districts, instead of just one, is analytically more complex, it has the advantage of allowing us to get an idea of the importance of enactment to the strength of the relationships that constitute evidence of validity, and how well validity is likely to generalize across districts. It is important to show that validity evidence is consistent across districts. This is because the validity of assessments made using any high inference tool such as the one we are developing depends not only on the instrumentation (the defined performance dimensions and the rating scales or rubrics) and the formal procedures (e.g., how evidence is to be collected, how evidence from multiple sources is to be combined) but also on the way the process is enacted and the specific judges that make inferences from evidence to ratings. Of necessity, one collects validity evidence not only on the instrument but also on the implementation.

Study 6: Inter-judge agreement of formative assessment ratings (Year 3)

Though the SLAT tools will be designed to be used by school leadership teams for formative rather than summative purposes, it is still important to demonstrate that school leaders who assess the leadership performance in their own school can and do use the tool to make assessments that are based on the actual features of the situation, and not simply on the leadership team member's idiosyncratic interpretation of the evidence and the rating scales or rubrics. We will be asking school leadership team members to make an independent, initial judgment of school leadership performance at the beginning of the school year, and another at the end of the year. We will collect these independent assessments and examine their agreement. Because the absolute scale level at which school leaders rate their performance is important, and because each set of raters (teams) rates only one object (their team performance) we will measure agreement using the percent absolute agreement (at the task level) and the standard deviation of the judges' ratings of leadership at each school (at the element level,) where scores are derived from the task ratings.

We will collect self-report information from team members about the leadership team demographics (e.g., position, tenure with the school, past leadership experience) and will use this information, along with information about other school characteristics (e.g., school level, relative level of student achievement within the district), to look for factors that might account for differences in agreement across schools. We will formally test for the effects of these factors using regression models in which the agreement indices are treated as functions of these factors, and the coefficients compared across districts.

We will investigate the agreement in element level averages derived from judgments by the school leadership teams and from judgments made by individuals who are knowledgeable about the school but not members of the leadership team. These would include district-level staff such as principals' coaches or mentors, and teachers in the school who are not members of the leadership team. It is important to look for agreement with judges from outside the leadership team in order to assess the possibility that team members self-assessment of their team performance is not due to wishful thinking or leniency toward one's self or school. Such self-assessments are unlikely to be a useful guide to performance improvement.

We will calculate these agreement measures within schools, average them within districts, then combine across districts. Our desired standard of agreement is to have an average absolute agreement at the task level of 75% across the districts participating in the project, and an average standard deviation of .5 or less (one half of the distance between rubric levels, which are given integer values) at the element level. Where substantial differences between districts exist, we will investigate potential causes of lower (or higher) agreement by looking at how the districts implement the tool, including factors such as the training of raters, the evidence gathering process, who makes the ratings, and the purpose of the ratings.

Study 7: Relationship of ratings with other indicators of leader or school performance

We will be looking at the relationship between formative assessment ratings and three other indicators of leader and school performance: measures of student academic achievement, results of school climate/culture surveys, and summative performance evaluation ratings of school leaders.

7a. Student achievement. To assess the relationship between the assessment ratings and student achievement, we will compile or develop value-added measures of school average student achievement, then correlate these with ratings (including ratings of the dimensions and an average across the dimensions). The district of Madison can provide school-level value-added indicators that they develop for their own internal research and accountability purposes. Madison's indicators were developed in conjunction with WCER's Value-Added Research Center, using a model similar to that described below. For the other districts, we will have to develop such indicators.

We plan to use 6^{th} to 7^{th} and 7^{th} to 8^{th} grade value-added estimates in reading and mathematics as indicators of middle school effectiveness, and 8^{th} to 9^{th} and 9^{th} to 10^{th} and grade value-added estimates in reading and mathematics as indicators of high school effectiveness, where 9^{th} grade tests are available. In other cases, we will have to use state 8^{th} grade tests as the "pretest." For these districts, we will ensure that students included in the value-added analysis

are those with continuous enrollment in a single school from the beginning of 9th grade to the date of the 10th grade assessment. In all of the districts, existing state standards-based testing programs provide the needed 6th, 7th, 8th, and 10th grade reading and mathematics test scores.

The estimation of school-level value added in reading and math for districts without their own value-added systems will be done using a fixed-effects regression approach that has been used by the Wisconsin Center for Education Research Value-Added Research Center to estimate value added in the Milwaukee and Chicago public schools. In this model, the left-hand-side variable in the regression is test score for a subject in a grade, and the right-hand-side of the regression includes test scores in the same subject (and, possibly, other subjects) in previous years; a vector of student characteristics such as sex, ethnicity, special education, and free or reduced price lunch eligibility; and a full set of school fixed effects. Formally, the simplest model is described in Figure 2, where Y_{it} is student i's test score in the current school year t; Y_{it-1} is student i's score(s) in the previous year t-1, X_{it} is a vector of student characteristics, as is a fixed effect for the school s attended by student i in year t, and e_{it} is the error term.

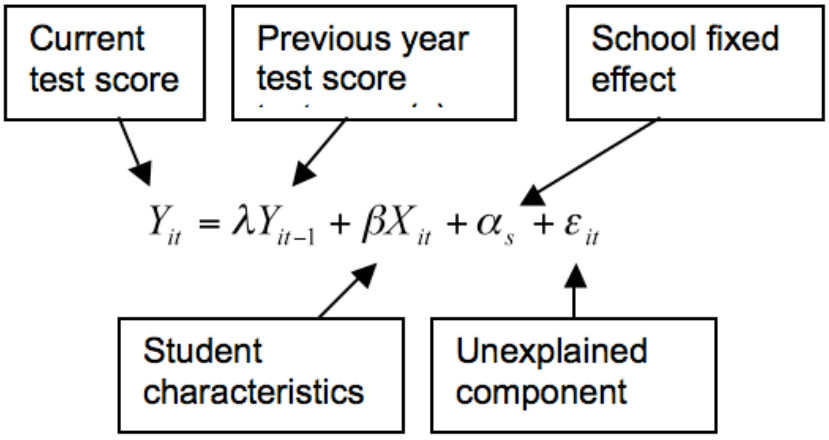

Figure 2. **Value-added model.**

To measure value added, the fixed effects α_s are centered around zero using enrollment as weights. The centered fixed effect for a given school thus indicates the degree to which on average student achievement deviates from the average level in the district. It is equal to the number of extra points on the test that students who attended school *s* scored relative to observationally similar students throughout the district. To combine estimates across grades, we will standardize these estimates for each grade (and subject) and calculate the weighted average across grades (with the number tested in each grade as the weights).

The value-added model described above is adaptable to the testing and data specifications of different school districts. It can be adapted to accommodate retention, mid-year testing, changes in the testing regime, summer school, and other special features. This model can also be extended to include measures of school characteristics (e.g., mobility, school-level percent of students eligible for free or reduced price lunch). The school effects (α_s above) then represent value added net of these characteristics. We will use a number of model variations to assess the sensitivity of the value-added estimates to model variations. When estimating value added as an

indicator of school performance for use in a validity study, it may be appropriate to control for as many characteristics that may be outside leadership teams' control as possible. Yet it may also be that, analogous to estimating classroom value added, adding controls may underestimate "true" effects (Ballou, Sanders, & Wright, 2004) and may be perceived as operationalizing lower expectations for some schools. It may also not matter which approach is taken. By using models both with and without such controls, we will get an idea of the impact of different sets of controls on the relationship between evaluation ratings and value-added.

Next, formative assessment ratings on each dimension and the average across dimensions will be correlated with value-added indicators for both the contemporaneous year and the next year (because leader effects may take more than one year to show up). We will calculate correlations with both composite evaluation scores (representing a summary of rated performance) and ratings on separate performance dimensions to determine which performance dimension ratings best predict student achievement.

Our approach will be to treat the analysis of the assessment rating–student achievement relationship for each district as a separate study, and is necessitated by the different tests and testing patterns used by each district. (This is a common approach in criterion-related validity studies of performance evaluation and selection tests in industry.) We will then combine results across districts to get a more reliable estimate of the relationship using meta-analytic procedures (e.g., Shaddish & Haddock, 1994). Table 4 presents a hypothetical example of how this analysis will be summarized for each performance dimension and for the average across dimensions.

Table 4
Hypothetical Validity Results for One Dimension

		Correlations of average assessed performance with:	
District	***N* (schools)**	**Reading value-added**	**Math value-added**
A	30	.21	.24
B	50	.24	.20
C	18	.26	.19
D	11	.18	.26
Combined:	109	.23	.22
Lower & upper limits:		.04-.41	.02-.40
P value:		.02	.03

It is important to recognize that summaries of existing research on principal influences on student achievement (e.g., Hallinger & Heck, 1996) suggest that the effect is largely indirect, mediated by conditions principals can establish or influence that in turn effect teachers, whose efforts are more directly related to student achievement. Thus correlations of above .30 are not likely to be found. Yet, correlations in the .20-.30 range would still be substantively meaningful. Correlations at that level were found between measures of leader behavior and student achievement in a meta-analysis by Waters, Marzano, and McNulty (2003).

7b. School climate/culture measures. Since an important aspect of school leadership is the development of a productive school culture, which in turn has been shown to be associated with student achievement (Heck, Larsen & Marcoulides, 1990), school climate/culture survey results will be used to provide an indicator of the results of leadership. The dimensions of school culture we intend to assess and relate to leader ratings include three that research has linked to student achievement: professional learning community (Bryk, Camburn & Seashore Louis, 1999; Louis & Marks, 1998; Newmann & Wehlege, 1995), collective efficacy (e.g., Goddard, 2002), and academic press (Hoy, Sweetland, & Smith, 2002; Lee & Smith, 1999). In some of our districts, existing school climate/culture surveys collect data from school staff on these constructs. We will work with the districts to modify their surveys to collect the needed data. In addition, we have allocated funds in the project to administer our own survey to ensure appropriate and consistent measurement of key constructs. Existing scales shown to be reliable and related to student achievement by prior research will be used. Since the reliability school surveys depend, *inter alia*, on a high teacher response rate to obtain reliable school averages, we will work with districts to develop ways of ensuring high response rates and will assess the reliability of school averages before using them as indicators.

7c. Leader summative ratings. In many of the participating districts, principals and assistant principals receive summative performance evaluations each year or once every few years. These evaluations are intended to reflect principal performance with respect to district priorities and conceptions of principal performance. Though some of the performance constructs assessed by these evaluations differ from those in our formative system, we would expect that there should be some degree of positive correlation between principal summative ratings and formative ratings of school leadership performance using SLAT. This is because the formative system and the districts' summative systems both reference many of the same leadership activities and results. Thus the summative ratings provide alternative measurements of some of the formative system's constructs. To begin this phase of the assessment evaluation, we will conduct a content analysis of each of the participating district's summative assessment systems in order to identify those performance dimensions that have substantial conceptual overlap with our formative system, and those that have little conceptual overlap. We would expect ratings on these dimensions to have substantial correlations (.4-.6) with the similar formative ratings. Where content does not overlap substantially, we would expect lower correlations. This pattern of correlations would be evidence for the validity of formative assessment ratings. Next, we will administer the VAL-Ed evaluation instrument in each school site to provide a common measure of school leadership across the schools and districts. Because Val-Ed has substantial conceptual overlap with SLAT, we would expect substantial (.4-.6) correlations between formative and summative ratings.

Study 8: Construct validity of formative assessment ratings

We will be looking at how well the ratings are measuring the leadership performance dimensions they are intended to measure (their construct validity) in four ways.

8a. How users justify ratings. We will collect evidence about how the raters make judgments using the assessment system, including how they gather evidence and their judgment processes in making the ratings. Analysis of raters' response processes is one type of validity evidence recognized by the *Standards for Educational and Psychological Testing* (American Educational

Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) and advocated by scholars of validity such as Messick (1994). Though we will be using cognitive interviews extensively in the development process to understand how the SLAT tools are used to make judgments, we will also do more focused cognitive interviews in the evaluation study. We will interview a sample of school leadership teams who have used the tools in each district using a semi-structured protocol. This protocol will elicit information from the individual team members (judges) about the evidence they collected and considered, how this evidence was interpreted, and the basis on which specific judgments about the school's level on the SLAT rubrics were made. The interviews will be recorded, transcribed, and then analyzed to assess (a) the relevance of the evidence collected and considered to constructs measured by the assessment; (b) the degree to which the evidence justifies the ratings given; and (c) any difficulties judges had collecting or interpreting evidence or using the rubrics to make judgments. The validity of the judgments will be supported to the extent that judges collected relevant evidence, recorded it accurately, interpreted it consistently, made rating decisions based on that evidence and not other unrelated factors, and reported little difficulty making rating decisions. We will do these interviews of one randomly-selected member from a random sample of 30 schools, with the sample allocated to districts proportional to the number of schools, for a total of 30 leaders.

8b. Correlations across ratings. A second type of construct validity evidence will be obtained by looking at the interrelationships between the judgments made on the SLAT dimensions. First, we will examine the pattern of correlations among the dimension ratings. If the assessment dimensions represent a set of distinct but related constructs, as we expect, we will find that the dimension ratings will be correlated with each other, but not so highly that one dimension rating is a nearly perfect predictor of another. (This would indicate that the ratings are not measuring distinct constructs.). Dimension intercorrelations in the .4-.7 range would indicate the expected level of relationship, but still be consistent with interpreting the dimension ratings as measurements of distinct performance dimensions. We will also examine the means and distributions of the dimension ratings. Though it is possible that most of the (leaders, schools) in the participating districts are high performers, this is unlikely given that the upper levels of the assessment rating scales were designed to represent a high level of practice. We would therefore expect the average ratings to be at or below the midpoint of the rating scales, and the distribution of scores to be such that a substantial number (around 25%) of schools were rated at the lower two levels of the scales.

We will also conduct a hierarchical confirmatory factor analyses of both the beginning and end of year SLAT ratings. We plan to test a three level model in which each of the task ratings loads substantially (.6-.8) on first order factors corresponding to the appropriate SLAT element, and these factors load substantially (.4-.6) on the appropriate second order factors representing the SLAT dimensions. The dimensions themselves will be allowed to correlate, and we would expect a substantial correlation between them .We would expect to have a good model fit (e.g., RMSEA <. 6, Non-Normed Fit Index >.9) without having to include separate paths between individual task ratings or element factors (which would suggest that there are other constructs or error sources common to one or more of the SLAT tasks or elements).

8c. Correlations with outcome indicators. A third type of construct validity evidence that will be examined is the pattern of the correlations between formative assessment dimension

ratings and the outcome indicators discussed above. Though we believe that all of the dimensions of the SLAT formative assessment are important and should be related to the indicators of leader or school performance discussed in the sections above, we also expect some dimensions to be more strongly related to certain indicators. For example, we would expect that the formative assessment dimensions involving building school community would be more strongly related to culture/climate survey measurements of constructs such as professional community than to student achievement, due to the closer conceptual and causal relationship between behaviors and results related to community building as defined in the tool and the constructs measures by the surveys.

Study 9: Consequential validity

Tools for the formative assessment of school leadership are intended to help leadership teams focus on specific behaviors, practices, or results that are related to school performance. Thus, one consequence of the use of the tools should be that school leadership teams make efforts to improve their leadership practice as the latter is defined by the tools. As part of the third year evaluation study, we will collect evidence that leadership teams using the SLAT are focusing efforts on behaviors and performances emphasized in the tools. We will request that members of leadership teams in the all schools in the study districts complete brief Web-based surveys twice each semester asking them about the school leadership issues they have been working on and the use they have made of the SLAT tools. This survey will include items corresponding to each of the leadership dimensions of the tool, and include items related to learning activities team members might undertake related to each dimension. Responses will be compared with the profile of ratings members made on the beginning of the year using the tools. If the tools are acting to guide the learning of leadership teams, we would expect that over the year team members would report undertaking more learning activities related to dimensions on which their team's initial assessment was low, and spending more time on leadership issues related to the dimensions or elements with lower scores. We will also collect school improvement plans and related documents (e.g., management team meeting agendas) describing leadership team activities during the school year from a random sample of 30 schools, with the sample allocated to districts proportional to the number of schools. We will content-analyze them in order to see if the improvement efforts they mention are more strongly related to SLAT dimensions that were assessed as needing improvement. This will provide additional evidence that the use of the SLAT tools is influencing the efforts of the school leadership teams.

3.0 Personnel

The SLAT proposal brings together a team of researchers and practitioners ideally situated to building the next generation of formative tools for assessing school leadership. Our personnel contributes expertise on school leadership, rubric development, the design of large-scale reliability and validity studies, conducting survey and interview research in schools, data analysis and academic writing. Details on how the work will be distributed among the investigators can be found in the Budget Narrative.

3.1 Principal Investigators

Principal Investigator **Richard Halverson** (Ph.D. Northwestern University) is an associate professor of Educational Leadership and Policy Analysis at the University of Wisconsin–Madison. Halverson is nationally recognized for his research on distributed leadership, data-driven decision-making, teacher evaluation, and technology leadership. He is the author of the widely used *School Leadership Rubrics* developed for the University of Pittsburgh's Institute for Learning. Halverson is a former high school teacher, school technology specialist, curriculum director and school administrator. He is PI of a National Science Foundation CAREER Award funded project to study how schools develop the capacity to engage in data-driven instructional practices, and is a co-founder of the Games, Learning and Society at UW-Madison, an internationally known research group that investigates how cutting edge learning technologies can reshape learning in and out of schools. His research work has involved collecting data from teachers and principals, developing software and Web-based learning and assessment systems, conducting survey research in dozens of schools and districts, and managing large, multi-site research projects.

Co-Principal Investigator **Carolyn Kelley** (Ph.D. Stanford University) is a professor of Educational Leadership and Policy Analysis at the University of Wisconsin–Madison. Professor Kelley conducted research with the Consortium for Policy Research in Education (CPRE) from 1989 to 2002. She is an internationally recognized scholar in teacher compensation policy whose research focuses on the preparation and professional development of school leaders, and teacher evaluation and compensation as elements of strategic human resources management in schools. Her current research focuses on advancing a shared conception of mastery in educational leadership, including developing and documenting the practices of principals who have led their schools to close achievement gaps and significantly advance learning for all students. She is the author of two books: *Doubling Student Performance: A School Leaders' Field Guide to Closing Achievement Gaps and Advancing Learning for All Students* (with James J. Shaw) and *Paying Teachers for What they Know and Do: New and Smarter Compensation Strategies to Improve Schools* (with Allan Odden).

3.2 Investigators

Anthony Milanowski is an assistant research scientist with CPRE at the Wisconsin Center for Education Research. Milanowski's research investigates the effects of management interventions on instruction and student achievement. His work on the CPRE Teacher Compensation Project studied the relationship of teacher evaluation ratings to value-added measures of teacher performance. This project involved collecting data from teachers and principals on their use of and reactions to these evaluation systems, using interviews, surveys, and value-added analyses. Milanowski was also principal investigator on a recently completed IES-funded study of principal performance evaluation in two districts, which involved combining interviews, surveys, and value-added analyses.

Steven Kimball is a researcher with CPRE at the Wisconsin Center for Education Research. Kimball has extensive experience designing and overseeing research using qualitative methods. He was co-PI on an IES-funded study of principal performance evaluation in two districts, and was site manager for one of the research sites in the CPRE Teacher Compensation Project study

of the relationship between standards-based teacher evaluation and value-added measures of teacher performance. He also coordinated a recent meta-evaluation of the 5-year, $55 million Education Initiative of the Chicago Community Trust. This work has included developing interview protocols, and coding and analyzing interview data using software such as NVivo.

Matthew Clifford (M.S. Education Leadership & Policy Analysis; Adult & Continuing Education, University of Wisconsin–Madison). Clifford is an experienced researcher and evaluator specializing in teacher professional development, high school reform, instructional leadership, and K–20 partnerships. Clifford has conducted several research studies with LPA, including an examination of Midwestern university-based principal preparation programs, and a study of state department of education human resource capacity to enact ambitious accountability and instructional reform policy. He is also currently interested in the evaluation of novice and master school principals, distributed leadership, social network analysis, science education and its improvement (his dissertation topic) and teacher evaluation practices (a recent publication).

Christopher A. Condon (Ph.D., Experimental Psychology, University of Arkansas). Dr. Condon was hired as a statistician at LPA in 2008, and primarily works on the REL-Midwest contract. Prior to working at LPA, Condon was a researcher at the Johnson O'Connor Research Foundation, which does large-scale aptitude testing. As a statistician and methodologist at LPA, his responsibilities include working on all phases of the research process including study design, implementation, execution, and reporting. He also serves as a consultant on data analysis matters for others within the company. He is well-versed in software programs such as SPSS, Amos, WINSTEPS, PowerPoint, Word, and Excel. Condon also has working knowledge of statistical procedures such as analysis of variance, regression, factor analysis, structural equations modeling, and item response theory. He has presented research findings in numerous research reports and at many conferences.

4.0 Resources

The SLAT project will be led by the Primary Investigator (Halverson) and the Co-Primary Investigator (Kelley). The management team (Halverson, Kelley, Clifford, Milanowski and Kimball) will meet monthly and will be responsible for oversight and coordination of all projects. Since UW-Madison and Learning Point Associates are located within driving distance, we will meet in person to conduct the majority of our project meetings. The PIs have extensive experience in leading coordinated research projects, as indicated under Personnel. The personnel team also has a record of significant accomplishment in the design of technology-based systems for professional interaction and on the design and analysis of large-scale validation studies. Halverson and Kelley will be directly responsible for the activities of the collaborative design team. Halverson will oversee the coordination of all studies and team efforts across the project, while Clifford, Milanowski and Kimball will design and guide the nine empirical studies described above.

4.1 Wisconsin Center for Education Research

The project will be headquartered in the Wisconsin Center for Education Research (WCER) at the University of Wisconsin–Madison (UW-Madison). WCER is one of the nation's oldest and most highly esteemed university-based education research and development centers. With annual

extramural funding of exceeding $30 million, WCER is home to centers for research on the improvement of mathematics and science education from kindergarten through postsecondary levels, the strategic management of human capital in public education, and value-added achievement, as well as the Minority Student Achievement Network and a multistate collaborative project to develop assessments for English language learners. The WCER Technical Services Department provides statistical consultation, multimedia services, custom software development, and computer support for more than 350 networked computer systems. The department includes a state-of-the-art multimedia studio staffed by multimedia artists, animators, and programmers. WCER's business office provides projects with budgetary, forecasting, accounting and financial management, and human resource management. A professional editor provides assistance with manuscripts, guidance on preparation of human subjects protocols, and editorial and technical oversight for proposals. A public information specialist/photographer helps disseminate research findings through the WCER Web site (www.wcer.wisc.edu/), a quarterly newsletter (*WCER Research Highlights*), a monthly electronic newsletter (*WCER Today*), the university news service, and the national media.

4.2 University of Wisconsin Educational Leadership and Policy Analysis Department (ELPA)

The UW-Madison School of Education is consistently ranked one of the top schools of education in the country. ELPA is ranked second in the country in the 2008 U.S. News & World Report guide to the best graduate schools of education. The University of Wisconsin–Madison is recognized throughout the world as one of this nation's great universities. Its academic reputation has been rated among the top 10 in the country in many areas of study since the beginning of the last century. U.S. News & World Report currently ranks UW-Madison seventh among U.S. public universities.

4.3 Learning Point Associates

Learning Point Associates is a nonprofit educational organization with 25 years of experience working with and for educators and policymakers to transform education systems and student learning. Key to their success is the ability to collaborate productively with other organizations, forging strategic alliances for added value and efficiency. Learning Point Associates is nationally recognized for its work in teacher quality, school leadership development, district and school improvement, afterschool services, high school improvement, and literacy. Since 1984, Learning Point Associates has operated the regional educational laboratory serving the Midwest, which is now known as REL Midwest. Learning Point Associates also operates the National Comprehensive Center for Teacher Quality with our partners Education Commission of the States, ETS, and Vanderbilt University; Great Lakes East Comprehensive Center; Great Lakes West Comprehensive Center; The Center for Comprehensive School Reform and Improvement, and the NCLB Implementation Center.