The Design and Validation of the Comprehensive Assessment of Leadership for Learning (CALL) Formative School Leader Assessment

Professors Carolyn Kelley, Richard Halverson and Eric Camburn
Department of Educational Leadership and Policy Analysis
University of Wisconsin-Madison
270 Education Building
1000 Bascom Mall
Madison, WI 53706

Mark Blitz, Post-Doctoral Scholar
Wisconsin Center for Education Research
University of Wisconsin-Madison

Jason Salisbury, Doctoral Student
Educational Leadership and Policy Analysis
University of Wisconsin-Madison

Seann Dikkers, Doctoral Student
Department of Curriculum and Instruction
University of Wisconsin-Madison

Matthew Clifford
American Institutes for Research

March 17, 2012

The Design and Validation of the Comprehensive Assessment of Leadership for Learning (CALL) Formative School Leader Assessment

Carolyn Kelley, Richard Halverson, Eric Camburn, Mark Blitz, Jason Salisbury, and Seann Dikkers, University of Wisconsin-Madison, and Matthew Clifford, American Institutes for Research

Background and Purpose

Many middle and high school principals struggle with structural reforms that have limited impact on classroom practice and student learning because they leave the instructional cultures of their schools untouched (Hargreaves et al, 2007; McLaughlin & Talbert, 2001; Siskin, 1995). Recent changes to educational accountability policy have raised the stakes for school reform by holding principals accountable for the success or failure of their schools (Karhuse, 2012). While the principal plays a central role in laying the groundwork for advancing student learning, issues such as large school size, complexity of organizational cultures, and norms of teacher autonomy and isolation characteristic of many middle and high schools highlight the importance of assessing and developing *leadership throughout the school* rather than simply focusing on a single *school leader*. Leadership, however, is not merely a generic feature of organizations. Rather, leaders engage in a series of tasks that establish the conditions for teaching and learning in schools (Spillane, Halverson & Diamond, 2004). Improving organizational leadership would require tools that help researchers and practitioners identify these tasks, determining who does them and then measuring the degree to which tasks actually improve teaching and learning (Halverson, 2007). Tools that provide information on the key tasks of leadership practice would provide middle and high school principals with valuable feedback to support the improvement of leadership across the school

organization, and support to guide the ongoing development of instructional leadership throughout the continuum from novice to expert practice.

In education, formative feedback is typically studied in terms of motivating or directing *student* learning (Brophy, 1981; Schwartz & White, 2000; Black & Wiliam, 1999). However, formative feedback is increasingly important for guiding *adult* learning. The recent proliferation of benchmark assessment systems in public schools (c.f. Means, Padilla & Gallagher, 2010) demonstrates the felt need for educators to receive timely, standardized feedback on the progress of student learning. School leaders also need timely information on the progress of local initiatives in professional development, resource allocation, assessments and school safety to improve teaching and learning across the school. This study focuses on the development of a formative assessment instrument designed to strengthen instructional capacity by providing timely information on leadership task enactment across the school organization at the middle and high school levels.

Specifically, this paper describes the design and validation of the Comprehensive Assessment of Leadership for Learning (CALL), a 360-degree, on-line, formative assessment and feedback system for middle and high school leadership. CALL is designed to assess specific leadership practices or tasks characteristic of high performing middle and high schools (Halverson, 2007; Kelley, 2010; Kelley & Shaw, 2009). The survey captures leadership practices and school cultures across five domains of leadership practice:

1. Focus on Learning
2. Monitoring Teaching and Learning

3. Building Nested Learning Communities

4. Acquiring and Allocating Resources

5. Maintaining a Safe and Effective Learning Environment

Consistent with research on distributed leadership, the CALL survey defines leadership as distributed across the entire school organization, rather than as the actions or behaviors of a single person. Thus, the CALL survey examines the set of leadership practices carried out by formal and informal leaders distributed throughout the school. This is unique to CALL and its design as a formative organizational assessment instrument, rather than as a summative assessment of the leadership of the principal. The focus on leadership is consistent with our theoretical framing of leadership as distributed across the school organization (Spillane, Halverson & Diamond, 2001), and with the principles of effective feedback, which suggest that to motivate and direct improvements in performance, feedback should focus on the task and task performance, not on the individual person or the person's self-concept (DeNisi & Kluger, 2000).

Development of the CALL Survey

Development of the survey instrument began in 2009 with the support of a 4-year grant from the U.S. Department of Education to design and validate the survey instrument. The process of designing survey items involved attention to design issues consistent with the formative nature of the assessment, to ensure that the survey met what we viewed as four essential design criteria. Specifically, each item was designed to be:

- Aligned with *research on effective middle and high schools;*

- Grounded in *leadership practices rather than opinions* about the leader; and

- Framed to *transparently communicate the underlying theory of action*, so that the process of taking the survey would serve as a developmental experience for school leaders and instructional staff; and

- Consistent with *best practices in survey design.*

To ensure that these criteria were met, an initial draft of the survey was developed by the research team based on rubrics created by Professor Richard Halverson in conjunction with the University of Pittsburgh Institute for Learning (IFL) (for rubrics, see www.callsurvey.org/resources/rubrics). These rubrics were consistent with research conducted by Carolyn Kelley and James Shaw on leadership in schools that had consistently closed achievement gaps and improved overall student learning (Kelley & Shaw, 2009). We also conducted extensive reviews of research on effective middle and high school leadership, and on each of the domains of practice to ensure that item development was consistent with the research literature on effective leadership for learning, and more specifically on the practices represented by specific survey domains, sub-domains, and individual items.

Distributed leadership analyses propose that practice is composed of macro- and micro-tasks (Spillane, Halverson &Diamond, 2004). Macro-tasks refer to the general organizational tasks, such as providing adequate resources, planning and designing professional development, that organize much of the work of school leadership. Micro-tasks articulate these general responsibilities into the day-to-day activities of school leaders. Our survey design work translated micro-tasks into items that described practices that could be observed by teachers, leaders and staff in a typical school context. Our focus on middle and high school leadership contexts led us to describe micro-tasks to

reflect the appropriate departmental, grade level and instructional staff (e.g. special education, counseling, instructional coaches and mentor) contexts. The CALL survey articulated the work of school from five leadership macro-tasks into 115 items that described micro-tasks relevant for improving learning. The tasks described in the five domains include: 1) focus on learning; 2) monitoring teaching and learning; 3) establishing nested learning communities; 4) acquiring and allocating resources; and 5) ensuring safe and effective learning environments. Each macro-task, or domain, is organized into 4-5 sub-domains, which contain the specific items. The following sections provide a brief description of the research upon which the domains are based.

Focus on Learning. The CALL *Focus on Learning* domain contains four sub-domains: maintaining a school-wide focus on learning; formal leaders are recognized as instructional leaders; collaborative design of an integrated learning plan; and providing appropriate services for students who traditionally struggle. In maintaining a school-wide focus on learning, leaders regularly engage the school community and staff in ongoing conversations that serve as a foundation of a collective understanding of student learning (Goldring & Pasternak, 1994; Waters & Marzano, 2006). As a result, there is a collaboratively developed, shared vision of student learning that reflects the actual practices and aspirations of teachers (Hallinger, 2003; Leithwood, Jantzi, et al., 2004). School leaders regularly discuss student achievement data and concrete examples of instructional practice with teachers (Halverson, et al., 2004; Kelley & Shaw, 2009).

Through their substantive and symbolic interactions with the school community, formal leaders are recognized as instructional leaders. School staff and stakeholders recognize the principal as an instructional leader in the school and consistently seek

his/her input on a variety of instructional issues (Hallinger, 2005; Hallinger & Heck, 2002). School leaders regularly engage in public instructional leadership activities such as learning walks or classroom visits (Abrutyn, 2006; Biddle & Saha, 2006). School leaders work with teachers to organize professional development and curriculum design, and are active participants in the sessions (Goldring & Rallis, 1993; Pearce & Conger, 2003).

These highly effective middle and high schools are also characterized by collaborative design of an integrated learning plan. Strategies to improve student academic performance are the regular focus of faculty meetings (Scribner, Sawyer, Watson & Myers, 2007). The school has a collective instructional planning process that uses student multiple sources of data to coordinate specific instructional initiatives toward overall goals of student achievement (Knapp, Copland & Talbert, 2003; Leithwood et al., 2010; Yang, Goldstein, Rath & Hill, 1999). The school's learning plan integrates intermittent measures of student progress toward learning goals (Halverson, 2010).

Finally, effective collaborative planning is carried through to implementation, as appropriate educational experiences and services are provided for students who traditionally struggle. Special needs staff work together and with teachers to plan appropriate services (Frattura & Capper, 2007), and services are usually provided in the context of the regular classroom (Vaughn & Linan-Thompson, 2003). Leaders work with teachers to develop and monitor differentiated instructional practices for students who traditionally struggle (Fuchs, Mock, Morgan & Young, 2003; Lawrence-Brown, 2004). Teachers consistently use pre-assessment tools as a basis for differentiation in all content

areas (Hoover & Patton, 2008), and differentiation of instruction is regularly observed across subject areas (Hall, 2007; Tomlinson & McTighe, 2006).

Monitoring Teaching and Learning. The Monitoring Teaching and Learning domain contains four sub-domains: formative evaluation of student learning; summative evaluation of student learning; formative evaluation of teaching; and summative evaluation of teaching.

To carry out effective formative assessment of student learning, leaders provide structured opportunities at grade level or subject matter meetings for teachers to share practices for providing meaningful, systematic feedback on student performance (Fisher & Adler, 1999; Wenglinsky, 2002). Leaders recognize the value of formative assessments and provide opportunities for teachers to collaboratively redesign assessments in light of school learning goals (Black & Wiliam, 2004; Hallinger & Heck, 1996). The school successfully uses a systematic method for providing intermittent measures of student learning in order to predict and shape student-learning outcomes across classrooms and grade levels (Erickson, 2007; Hamilton, Halverson, Jackson, Mandinach, Supovitz & Wayman, 2009).

Because CALL has been designed in a particular student learning accountability context, summative evaluation of student learning connects classroom instruction and learning outcomes to performance on high stakes tests, such as state tests and college entrance examinations. Evaluations of student performance are based on multiple sources of data including student self-evaluation and/or self-reflection (Knapp, et al., 2003). Teachers and staff have multiple annual opportunities to collaboratively reflect on

student achievement data and redesign the school instructional program in light of the data (Smylie & Wenzel, 2003).

To provide leadership in formative evaluation of teaching, school leaders invest time in both formative and summative teacher evaluation and regularly provide feedback on teaching (High & Achilles, 1986; Marzano, et al., 2005). School leaders provide guidance for individual teachers to find resources to improve practice that are integrated into teacher and school improvement planning (Nelson & Sassi, 2005; Wahlstrom & Louis, 2008). Faculty meetings include samples of typical and exemplary student performance, to clarify teaching and learning tasks and distinguish levels of performance (Cawelti, 1997; Waters et al., 2003).

School leaders also participate in regular, meaningful summative evaluation of teaching. Evaluation policies are developed with the input of teachers and staff and are reviewed annually (Clark et al., 1980; Marzano et al., 2005). Occasions for evaluation are targeted to measure the staff's ability to effectively carry out the school's major instructional initiatives (Hallinger & Heck, 1998). The evaluation process draws on multiple classroom visits by multiple observers (Blase & Blase, 1999). Evaluation practices are used to document poor teaching as well as to provide valuable feedback for accomplished teachers (MDRC, 2007; Quint et al., 2007). The design of the evaluation process integrates measures of student learning and is linked with the school and teacher's professional development plan (Leithwood & Montgomery, 1982; Wahlstrom & Louis, 2008).

Building Nested Learning Communities. The Building Nested Learning Communities Domain contains four sub-domains: collaborative school-wide focus on

problems of teaching and learning; professional learning; socially distributed leadership; and collegial relationships.

First, school leaders build a collaborative school-wide focus on problems of teaching and learning. The school has collaboratively developed a long-term vision for instructional improvement (Hord, 1997; Printy, 2008; Waters & Marzano, 2006). Current programs and teacher planning build on past initiatives and continuous improvement (Collins, 2001; Gates & Watkins, 2010; Louis, 2006; Marzano et al., 2005). Professional development, curriculum design and school improvement planning are linked to the key problems of teaching and learning (Louis, 2006). Discussions of school instructional initiatives are participatory (Goldring & Rallis, 1993; Marks et al., 2002). Faculty committees develop intermediate timelines and benchmarks to determine whether new practices are helping achieve student-learning goals (Wayman et al., 2006).

Highly effective middle and high schools also have rich opportunities for professional learning and growth. The school has developed a long-term plan for focused support of professional growth in key instructional areas that provide differentiated support for individual teacher ability in terms of whole school instructional goals (Marks et al., 2002; Wahlstrom & Louis, 2008). Information is disseminated in alternative media to allow for maximum time for staff to engage in and reflect upon professional development activities (Kelley, 2010). A variety of summative and formative feedback assessments are developed to determine whether the professional development program is helping teachers improve student learning in targeted areas (Maldarez et al., 2007; Spillane et al., 2002; Waters & Marzano, 2006).

Leadership is socially distributed as leaders create structures through which teachers and staff are able to develop initiatives for the school's instructional priorities (Leithwood et al., 2010; Stoll et al., 2006). Leaders solicit staff and teacher feedback about the overall goals as well as the details of the school budget plan (Pearce & Conger, 2003; Spillane et al., 2002). Control over the direction and content of the instructional agenda is shared by leaders, teachers and staff (Hipp et al., 2008; Leithwood & Jantzi, 1999; Mitchell & Sackney, 2006; York-Barr & Duke, 2004).

The school is characterized by collegial relationships that recognize and access teacher expertise. Leaders provide teachers who have expertise in content and pedagogy with structured opportunities to share information, experiences and/or knowledge with other teachers (Hord, 1997; Roehrig et al., 2008). Expert teachers are selected to mentor other teachers on a regular basis, and mentoring training programs help mentors relate their experiences to mentees (Smith & Maclay, 2007). District-level instructional coaches are respected instructional leaders and are known for helping teachers solve problems and introducing new methods and practices (Abell et al., 1995; Blase & Blase, 1999).

Acquiring and Allocating Resources. The Acquiring and Allocating Resources Domain contains five sub-domains: personnel practices; structuring and maintaining time; school resources are focused on student learning; integrating external expertise into the school instructional program; and coordinating and supervising relations with families and the external communities.

Personnel practices are designed to ensure a highly qualified teaching force allocated to maximize student-learning opportunities. Specifically, teachers are certified

and/or meet requirements to teach in their assignments and teachers with specialized qualifications are actively recruited to fill needs (Darling-Hammond et al., 2001). Teacher induction programs are integrated into mentoring and professional development programs. School leaders have developed an incentive system to reward teachers for progress toward school-wide goals (Odden, 2011; Odden & Kelley, 2002).

Regarding the allocation of time, leaders structure professional time to address complex issues of instruction. Time is provided for whole-school, grade and subject matter level planning, curriculum design and reflection (Mertens & Flowers, 2006; Warren & Muth, 1995). Teachers receive feedback on effective uses of instructional planning time.

School resources are focused on student learning (Odden et al., 2007). Leaders perceive they have considerable range of discretion for allocating and acquiring necessary human, material and financial resources (Halverson & Thomas, 2007; Kelley & Shaw, 2009). Leaders base budget decisions on school-wide goals for student learning (Goldring & Pasternak, 1994). Fiscal and performance data are systematically reviewed for making informed decisions (Hallinger et al., 1993). There is a budget process that incorporates staff input and is transparent to stakeholders (Pearce & Conger, 2003).

Highly effective schools integrate external expertise into the school instructional program. Leaders continuously seek out expertise from the district and outside resources. The work of external experts is coordinated and targeted to school instructional goals (Cawelti & Protheroe, 2001). The school has cultivated "critical friends" to provide perspective on school progress. Leaders have strong relations with the district and are able to influence the design of district priorities (Halverson & Thomas, 2007; Kelley &

Shaw, 2009). Most teachers participate in professional networks outside the school

(Desimone et al., 2007).

Another important resource that schools need to utilize effectively is their

relationships with families and external communities. Teachers contact many families

per month to discuss academic progress, strategies for improvement, or to commend

students' successes (Fan, 2001). Families work with leaders to develop programs that

make the school more welcoming and bring community resources into the school

(Epstein & Dauber, 1991). The school regularly sends information to families through a

variety of media, and seeks out effective relationships with the school community to

support student learning (Erickson, 2007; Kelley & Shaw, 2009).

Ensuring a Safe and Effective Learning Environment. Safe learning environments

are a necessary condition for improving student learning (Brophy, 1986). The Ensuring a

Safe and Effective Learning Environment domain contains four sub-domains: clear,

consistent and enforced expectations for student behavior; safe learning environment;

student support services provide a safe haven for students who traditionally struggle; and

buffering the teaching environment.

In order to provide clear, consistent and enforced expectations for student

behavior, teachers and school leaders work together to ensure that discipline policies are

fairly, equitably and consistently enforced (Sheldon & Epstein, 2002; Foster, 2004;

Gottfredson, 2001). Teachers and leaders use data on student conduct and achievement

to review and adjust policies (Halverson, 2007), and students take ownership by

participating in the development and peer-enforcement of behavior policies.

A safe learning environment is supported by safety policies and procedures that

reflect school conditions and that are annually reviewed. A high percentage of students involved in fighting, theft, selling or using drugs, or are perpetrators or victims of harassment is a strong sign that teachers are unable to engage students in the classroom (Osher, Bear, Sprague & Doyle, 2010). Students regularly lead and interact civilly at school-wide assemblies, and school-wide announcements that interrupt classroom teaching typically occur infrequently.

Student support services provide a safe haven for students who traditionally struggle. The school effectively identifies students with special needs and successfully provides services to improve learning for most identified students (VanDerHeyden, 2007). Leaders work with teachers across the school to continually revise plans for improving attendance, dropout and graduation rates for students who traditionally struggle (Christle, et al., 2007). An extensive pool of adult mentors and advocates contact students in need to provide academic and social assistance (Humphrey, et al., 2009).

Leaders also buffer the teaching environment to ensure that effective instruction can occur. Leaders are able to help teachers deal with parent concerns when needed (Leithwood, Menzies, et al., 1994). In addition, leaders are able to relate the message of successful achievement at the school to district and community leaders. Leaders are successful advocates for district resources and filter them effectively to teachers (Berg, et al., 2066). Leaders have established and regularly review reliable procedures to control access to the classroom (Waters, et al., 2003), and teachers welcome classroom visitors.

Practices rather than Perceptions

CALL is designed to capture levels of leadership for learning by measuring existing leadership and learning practices from the perspective of school leaders and staff, and providing feedback to strengthen leadership. CALL is designed to provide feedback in three ways:

- Through transparency in the design of assessment items so that learning occurs as educators take the assessment;

- By providing assessment results that identify established levels of expertise, patterns in response items, and more traditional statistical summaries of results; and

- By providing leveled guidance on next steps for strengthening and building principal and distributed leadership for learning.

The overall feedback system is described elsewhere (Kelley, Kimball & Clifford, 2010); here we focus on the transparency of item design. A major goal of the survey design process was to ground survey items in choice options that reflecting actual practices, rather than framing responses in terms of perceptions of leadership practice (e.g., "strongly agree" to "strongly disagree;" or "not at all" to "to a great extent").

To the extent possible, the survey relies on prevalence of practices (e.g., what is the number of times per week teachers meet to talk about instruction?) rather than perceptions (e.g., to what extent do you think your principal is an effective instructional leader?) to gather data on leadership practices. By being explicit about a choice set ranging from low to high levels of practice, the survey provides clearer information about best practices underlying the assessment items and attempts to contextualize item response choices. The resulting survey has a relatively high cognitive demand, but items reflect actual practices in schools, consistent with a clearly specified model of leadership.

In addition, the leadership domains and rubrics that underlie survey design are available to participating schools and districts, and provide a clear identification of critical elements of effective leadership for learning, specified in the five CALL domains of leadership practice.

A consistent comment we received from practitioners who have taken the survey is that it is comprehensive and that taking the survey provided them with an opportunity to think about the things that they should do, that they do well, and that they need to work on in their leadership practice.   Because CALL reflects a model of distributed leadership, broad participation in the survey helps build awareness of leadership practices and challenges across the school community.

Consistent with Best Practices in Survey Design

The CALL survey design process began with initial construct identification and survey development based on the Halverson rubrics in Fall 2009. Beginning in Spring 2010, research to support item selection and construct validation was undertaken at the University of Wisconsin-Madison and through the North Central Regional Education Laboratory (which was acquired by the American Institutes for Research in 2010).

UW-Madison Item Selection and Construct Validation

UW-Madison researchers assembled two practitioner focus groups to review and improve the draft survey design. The middle school practitioner group consisted of two principals, an assistant principal, a school psychologist, a former Title 1 reading specialist, a special education teacher, and a language arts coordinator for the district. The high school practitioner group consisted of a principal, a department chair, a special education teacher, an assistant principal, and a former high school principal. All

practitioners were drawn from different schools, although some of the schools were located in the same district.

The groups met seven times over the course of four months. For the first meeting, the group members were asked to take the survey, recording any preliminary observations. For each subsequent meeting, the middle school and high school groups met separately in order to capture school-level feedback. In each meeting, the practitioners examined a specific domain closely, with the goal of providing feedback on the appropriate use and clarity of language, appropriateness to school level, importance of the question, including items that were not focusing on critical features of the construct or missing items, advice on who in the school should answer the question, and whether there should be any format changes to the questions. Individuals were also asked to determine whether the response options reflected the appropriate range of practice in middle and high schools.

The researchers used the track changes function in Word to capture the comments in the focus groups. After each meeting, survey items from that domain were reviewed and modified to incorporate practitioner feedback. Changes included adjusting language and defining terms, addition of new items, revision of items to better address core issues, creating multiple items out of a single item to eliminate double-barreled questions, and changing response options to more accurately reflect gradations of practice (Blitz & Clifford, 2010).

Overall, practitioner response informed item selection, reframing, and affirmed the content validity of the survey, suggesting that the items in the survey capture the

range of critical leadership practices to improve student learning at the middle and high school levels.

North Central Regional Education Laboratory (NCREL) Focus Groups.

Upon completion of CALL survey version one, item selection and construct validity inquiry was launched by NCREL with seventy-eight school-level leaders from eleven middle and high schools in Illinois and Wisconsin. Administrative and teacher leaders involved in providing round one feedback were all members of school leadership teams, who are likely to use CALL data for decision-making. Item selection and construct validity components involved: (a) completing version one of the online CALL survey, (b) rating the clarity of CALL survey items, and (c) providing feedback on the utility of CALL data and application to school-level decision-making.

Regarding survey design, the NCREL focus group data also showed that this initial draft of the survey was comprehensive and reflects major school leadership systems and actions, and taking the survey prompted self-reflection on leadership quality. Of concern to the leadership teams was the length of the survey, which required in excess of 35 minutes for respondents to complete. Suggestions included splitting the survey into pieces, or allowing individuals to leave the survey and return later to complete remaining items.

In addition, the NCREL focus groups provided similar feedback as the UW-Madison practitioner groups regarding terminology, double-barreled questions, reorganizing items to speed response time, making questions and responses more concise, and eliminating question redundancy.

From a utility perspective, the focus groups indicated that they did not have access to these data from other sources. They believed that these data would assist them in improvement planning, particularly if they could be combined with some demographic data that could show response differences among different groups (departments, leadership team members, etc.) within the school. They suggested that we consider providing access to research associated with constructs to reinforce the importance of leadership system quality for school improvement, and provide access to other, similar schools and high scoring schools, so leaders can connect with others about how to improve practice.

Year 2: Pilot Testing

The survey was pilot tested in 2010-11 with 1784 educators in eleven school districts in the Midwestern and Southern United States. In addition, three rounds of interviews were conducted around school context, survey administration, and utility of feedback with the principals and other survey users (leadership team members, teachers, and other staff) in six schools.

In addition to using the pilot testing as an opportunity to explore the utility and practicality of the survey, the pilots provided an important opportunity to test the web-based survey platform and identify any particular challenges associated with large-scale survey administration. Six of the pilot schools were involved in three rounds of interviews regarding school context, survey administration, and design and utility of feedback. Round one was an interview with the school principal, which focused on understanding the organizational and leadership context of the school. These data were designed to triangulate survey data to enable us to check on the ability of the survey to

capture critical context and leadership factors that shape the leadership challenge and impact student learning.

Round two involved interviews with the principal and teachers at each school to discuss survey administration, and to capture principal and teacher experiences in taking the survey. From this round, interviewees indicated that taking the survey was an important professional development opportunity for themselves and other staff members. The survey was designed to communicate a clear theory of action to survey respondents through the focus of questions and the response options, which were ordered to be transparent in terms of levels of practice. Teachers reported that the survey promoted informal conversations about professional practice among teachers; created expectations for future action in the school; and responding to the survey was perceived as a means of communication with the principal about the school's strengths and weaknesses.

Round three of data collection involved an interview with the principal to discuss survey results. In addition, the CALL instrument provides formative feedback to school leaders through summary results by domain, sub-domain, and item. Based on these results, school leaders also receive suggestions on leadership practices they could engage in to strengthen leadership practices shown to improve teaching and learning. The feedback designs reflect research on the principles of effective feedback (DeNisi and Kluger, 2000; Wimer, 2002), professional development (Darling-Hammond et al, 2007; Goldring et al, 2009; Gusky, 2005; Smylie et al, 2005) and adult learning (Knowles, 1980; Merriam & Caffarella, 1999; Wood & Thompson, 1980), and an analysis of features of formative feedback systems with purposes similar to the CALL system to

inform the development of a conceptual framework of research-based principals of effective feedback design.

In the interview with the principal, we specifically asked for information on how the principal would use the results of the CALL survey and what form of feedback would be most useful to principals. The survey is designed as a standards-based rather than norm-referenced survey, and principals were more interested in data communicating levels of practice (a frequency distribution across item responses) rather than summary data that would likely be used for comparative purposes (my school received an average of "3" on this response and the other schools in the district received a "2.5").

Principals also identified specific design features that promote effective communication of results and support mobilization for improvement, including:

- The transparency of the survey, which communicates effective research-based practices to survey participants and gives the school a sense of areas it may need to work on;

- Presentations of survey results that promote clear identification of current and desired practices;

- Item-by-item guidance on steps the school could take to improve practice for that particular item, connected to a larger vision or theory of action for effective leadership practice.

- The survey's focus on distributed leadership rather than the leadership of a particular individual. Principals felt that this design communicated the important role that all teachers play in taking ownership of leadership for learning in the school.

Statistical Analyses of Year 2 Survey Pilot Data

In addition to the qualitative information collected through interviews and focus groups, UW-Madison researchers used year two survey pilot data to begin analysis of the statistical properties of the survey to inform further survey refinement. Some features of the CALL survey include intentionally designing the survey instrument itself as a meaningful professional growth activity. At times, designing the survey instrument to provide formative information for school leadership development was in tension with other goals of survey design, such as designing to optimize statistical properties of the survey. Our design team recognized that weighting one set of goals for survey design at times interfered with addressing other goals of survey design, but throughout the process, we prioritized item design that reinforced formative features of the survey as well as maximizing statistical properties.

Examples of the trade-offs in designing items for statistical properties versus formative feedback occur in the design of item response sets that are ordered so the respondents can clearly identify desirable and undesirable practices, that focus on practices rather than perceptions, and that include items that have response choices that are theoretically desirable but occur rarely in schools. These design features increase the risk of socially desirable responses, increase the cognitive demand of the survey, and include items that may reduce scale reliability. For example, a survey item focusing on the use of technology to support student learning includes high-end options that are atypical in most schools. However, because the survey is intending to communicate best practices to encourage schools to strive for more effective practice, we chose to leave in high-end options even though few schools choose those options. These decisions make

the survey results robust for communicating a theory of action to survey respondents, and provide valuable data for discussion in staff meetings, but slightly reduce the robustness of statistical properties of the survey. Further research is needed to explore these trade-offs, but we believe prioritizing the formative goals of the instrument enhance its utility as a vehicle for formative conversations to improve leadership practice.

Reliability Analysis. Reliability is a basic measure of the validity of a scale. Conceptually, reliability is defined as the degree to which a scale is free from errors of measurement. Measurement errors will be higher to the extent that different measurements of the same person vary. For example, if we give Jill a math test on Tuesday and the same test on Friday, we would expect fairly similar results. If we do this for all of the children in a school and get very different results between Tuesday and Friday our test is unreliable. Reliability is operationalized as a measure of the degree of consistency between multiple, equivalent measurements of the same construct. Reliability is higher when multiple measurements are more consistent with each other and lower when measurements are less consistent. An important property of reliability statistics is that they tend to increase with a greater number of measurements.

To extend this example to the CALL data, multiple survey items that measure a particular sub-domain can be viewed as multiple measurements of a construct. For example, the CALL survey includes 6 items that are intended to measure sub-domain 1.2 "Formal Leaders are Recognized as Instructional Leaders." Reliability statistics for the CALL survey indicate the degree to which respondents' answers to items within a sub-domain are consistent. Reliability ranges from 0 (not reliable) to 1 (completely reliable). Extreme examples illustrate how this works in practice. Sub-domain 1.1 has 5 items. If

every respondent answered "1" to all 5 questions in sub-domain 1.1, the reliability for the scale would be 1.0. If everyone answer "4" to all 5 questions in sub-domain 1.1 the reliability of the scale would also be 1.0. If the answers of all respondents to the 5 questions were completely random, the reliability for the scale would approach 0. As stated above, as the number of measurements increases, reliability tends to increase. In the case of the CALL survey, in general, reliability will be higher for sub-domain scales with a larger number of items.

Our goal in instrument design was to achieve a reliability of at least .7 for each of the sub-domains. Reliability analysis based on the CALL Survey Version 1.0 provided mixed results in achievement of that goal, with initial Chronbach's Alpha reliability scores of .7 or above for 11 of the 16 sub-domain scales. For each scale with a reliability score below .7, we have reviewed items in that scale and have added items, or revised items to improve reliability. The reliability analysis is being repeated for CALL Version 2.0 following administration of the revised survey this Spring.

Rasch Analysis. We applied the Rasch model to CALL survey items to better understand scale reliability and the degree to which items within a sub-domain tapped a unitary dimension. The Rasch model is perhaps the simplest item response theory model that considers individual persons' responses relative to the response frequencies of all people. Item "fit" statistics from these models provided a useful diagnostic of how well particular items measured intended sub-domains. Item "difficulties" estimated from Rasch models provide evidence about whether there are sufficient items at all levels of the distribution of people on the scale to validly measure the full range of sample members. Scale reliabilities from the Rasch were similar to the standard Chronbach's

Alpha statistics presented earlier.   Items with poor fit statistics were identified as candidates for deletion or significant revision. Analysis of item difficulties identified sub-domains that would benefit from adding "harder" or "easier" questions. Reliabilities on some sub-domains were identified as low due to the limited number of questions in that domain (the survey had one to three items in some of the domains).

Not surprisingly, several of the items identified for revision, and sub-domains identified for addition of items had already been flagged through the qualitative reviews of the survey (reviews based on findings from the initial focus groups, pilot interviews, and data analysis). The CALL survey was further revised as needed to address issues identified through the Rasch analysis. These changes included adding items to the sub-domain, moving items from one sub-domain to another that had a better conceptual fit, recalibrating response options in survey items with skewed distributions, and refocusing items that were reducing scale reliability.

Variance Decomposition. A variance decomposition was conducted to assess the within-school versus across-school variance of survey items. Similar to other research in education, the decomposition of variance for most sub-domain scales showed more variance within schools than across schools. Typically, approximately 10 to 20 percent of the variance in sub-domain scale scores lied between schools. For formative feedback purposes, we believe that within-school variance can be as important as between-school variance for promoting discussions of differences in practice across classrooms or department within a school.   The between-school variance analysis provides an opportunity to recognize important contextual and performance differences between schools as well. A challenge throughout the process has been to appropriately interpret

and respond to recognized survey diagnostic procedures within a formative assessment context.

Item Frequency Distributions. Item frequency distributions were produced to provide an opportunity to use analysis of frequency distribution to inform survey refinement, and to explore initial results of the CALL survey. Frequency distributions were produced for the teacher and principal versions of the CALL survey.

The results were examined in by CALL researchers in a collaborative meaning-making session, and compared to results of the reliability, rasch, and variance decomposition analyses. Three primary patterns emerged as important for informing survey refinement. First, items that clustered around a single response were identified for possible refinement in terms of adjusting response options to capture variations in practice that were currently being grouped into a single response category.

Second, items that resulted in unexpected results were identified as needing additional refinement to clarify the question or response options. For example, a very high percentage of respondents in the pilot indicated that they participated in or experienced learning walks in the school. For this item, we determined a need to more clearly define the term "learning walks" to ensure that we were capturing actual practice and not a misinterpretation of the meaning of the question and response options.

Third, items that did not successfully distinguishing between schools were identified for refinement or elimination to ensure that the item was successfully distinguishing between schools in important leadership dimensions.

Survey Center Review for Consistency with Research on Effective Survey Design

The statistical analyses and qualitative studies described above helped to inform significant modifications to the survey instrument. In Spring, 2012, CALL researchers subcontracted with the University of Wisconsin-Madison Survey Research Center to provide an expert review of the survey design, support any further refinements in the survey instrument, site recruitment, and survey administration, and to develop a robust web-based platform for administration of the survey at a broader scale.

The survey instrument was reviewed for its conformity with best practice in survey design, including rules about question wording, question structure, response format, reference period, definitions and instructions. The Survey Center made some recommendations regarding:

- Simplifying wording and sentence constructions to promote cognitive processing by respondents that is more accurate and reliable (Dillman 2007; Tourangeau et al. 2000);

- Using "native" terms and phrases instead of "analytic" terms and phrases. Although the Year 1 practitioner focus group and leadership team reviews focused on this issue, the Survey Center review revealed additional analytic terms to consider for revision. Research demonstrates the respondents' ability to comprehend questions and retrieve information is better when the words and phrases used in the question match those used by respondents (e.g., Belson 1982);

- Rewording "noun adjectives." A "noun adjective" is a noun that is used to modify another noun. Use of nouns to modify other nouns may reduce the

respondent's ability to comprehend the phrase and increase likelihood that the meaning of phrase will vary among respondents.

- Eliminate contractions, abbreviations, parentheses, and brackets. The interpretation of these conventions may vary across respondents, be ambiguous, difficult to interpret, and lead to inaccurate responding in your survey (Guzman et al. 2006).

- Use parallel question wording and question structures. This was a particular challenge in the CALL survey, since the items describe actual practices, reducing the ability to common response choice patterns across questions.

- Avoid double-barreled questions. Research demonstrates that double-barreled questions slow respondents' cognitive processing (Bassili and Scott 1996).

- Match the response format implied by the wording of the question to the response format that appears in the response categories.

- Include a reference period when asking about events or behaviors that occurred in the past. A reference period is the period of time you want the respondent to consider when answering a question about an event or behavior that occurred in the past. Reference periods should appear as the first element in a question and be explicit, specific, and an appropriate length for the thing you are asking about (see Schaeffer and Presser 2003).

- Incorporate definitions and instructions into the body of questions to ensure that all respondents have the same information when they answer a question.

Recommended revisions to question design were reviewed and accepted unless they conflicted with the formative feedback survey design goals. In most cases, the recommendations were incorporated into the survey design.

ISLLC 2008 Standard Comparison

CALL is designed as a formative assessment instrument that can support leadership development to achieve educational accountability goals. While CALL is not designed to be used for summative purposes, its formative use in building leadership skills can support schools in achieving state policy goals for educational improvement. In that spirit, we examined the relationship between the CALL survey and the ISLLC standards.

We would not expect a one-to-one correspondence between CALL survey items and ISLLC standards because (1) CALL assesses leadership tasks that take place in the school regardless of who performs them, and ISLLC is focused specifically on an individual leader; (2) CALL assesses leadership tasks, and ISLLC includes a combination of performance tasks and dispositions; and (3) CALL is a 360-degree assessment that draws evidence from school administrators and instructional staff, limiting the ability of the instrument to assess dimensions of principal leadership in the broader policy context, such as legal compliance of leadership decisions.

Having said that, CALL provides data related to many of the ISLLC standards. On average, 4.2 CALL items were identified for each ISLLC 'performance', with the number of items ranging from 0 to 7.9 items per performance. Generally, performances clearly associated with tasks matched with more CALL items and performances

associated with attributes matched with fewer CALL items. Some CALL items apply to more than one ISLLC performance.

Our analysis of the relationship between CALL and the ISLLC standards suggests that CALL is an effective leadership development tool to support principals in building leadership behaviors consistent with the ISLLC standards. Our analysis suggests that CALL could serve as an important formative tool for school leadership in the context of a state leadership accountability system aligned with the ISLLC standards.

Discussion and Future Research Directions

This overview provides a summary of CALL research to date, documenting the development of a formative assessment instrument designed to strengthen distributed leadership for learning in the complex and challenging organizational environment of middle and high schools.

In Spring, 2012, CALL is being administered in 120 middle and high schools across the country. CALL survey results will be compared with value-added test scores, climate survey data, and other measures of leadership to validate the instrument, to our knowledge making CALL the first validated formative leadership assessment instrument. The CALL feedback system is under development as well, to provide specific feedback to school leaders on how to interpret CALL results, and what steps they can take to strengthen distributed leadership for learning in their schools.

CALL researchers are also currently developing an elementary school version of the survey. While the school-level specific instrument provides important context-specific data about leadership for learning in middle and high schools, many districts have expressed interest in administering CALL to all of their schools in order to provide

data to support district-wide leadership development initiatives. We hope to pilot an initial version of the elementary survey in Spring 2013, and continue development and validation of the instrument moving forward.

References

Abrutyn, L. (2006). The most important data. *Education Leadership, 63*(6), 54-57.

Biddle, B. J., & Saha, L. J. (2006). How principals use research. *Education Leadership, 63*(6), 72-77.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability (103rd Yearbook of the National Society fo the Study of Education*. Chicago, IL: University of Chicago Press.

Blase, J., & Blase, J. (1999). Principals' instructional leadership and teacher development: Teacher' perspectives. *Educational Administration Quarterly, 35*(3), 349-378.

Brophy, J. E. (1981). Teacher praise: A functional analysis. *Review of Educational Research, 51*(1), 5–32.

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist, 41*(10), 1069-1077.

Cawelti, G. (1997). *Effects of High School Restructuring: Ten Schools at Work.* Arlington, VA: Education Research Service.

Cawelti, G., & Protheroe, N. (2001). *High student acheivement: How six school districts changed into high-performance systems*. Arlington, VA: Educational Research Service.

Darling-Hammond, L., LaPointe, M., Meyereson, D., & Orr, M. (2007). *Preparing school leaders for a changing world: Executive summary.* Stanford, CA: Stanford University, Stanford Educational Leadership Institute.

DeNisi, A.S., & Kluger, A.N. (2000). Feedback effectiveness: Can 360-degree appraisals be improved? *Academy of Management Executive, 14(1).*

Erickson, F. (2007). Some thoughts on "proximal" formative assessment of student learning. In P. A. Moss (Ed.), *Evidence and decision making* (pp. 186-216). Malden, MA: Blackwell.

Frattura, E. M., & Capper, C. A. (2007). *Leading for social justice: Transforming schools for all learners*. Thousand Oaks, CA: Corwin Press.

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research and Practice, 18*(3), 157-171.

Goldring, E., & Pasternak, R. (1994). Principals' coordinating strategies and school effectiveness. *School Effectiveness and School Improvement, 5*(3), 239-253.

Goldring, E., Porter, A., Murphy, J., Elliott, S. N., & Cravens, X. (2009). Assessing learning-centered leadership: Connections to research, professional standards, and current practices. *Leadership and Policy in Schools 8*, 1-36.

Goldring, E. B., & Rallis, S. F. (1993). *Principals of dynamic schools: Taking charge of change*. Newbury Park, CA: Corwin Press.

Hallinger, P. (2003). Leading educational change: Reflections on the practice of instructional and transformational leadership. *Cambridge Journal of Education, 33*(3), 329-351.

Hallinger, P., & Heck, R. (1996a). The principal's role in school effectiveness: An assessment of methodological progress, 1980-1995. In I. K. L. a. P. H. (Eds.) (Ed.), *International handbook of educational leadership and administration* (pp. 723-783). Dordrecht, NL: Kluwer.

Hallinger, P., & Heck, R. (1996b). Reassessing the principal's role in school effectiveness: A review of empirical research,1980-1995. *Educational Administration Quarterly, 32* (1), 5-44.

Hallinger, P., & Heck, R. (2002). What do you call people with visions? The role of vision, mission, and goals in school leadership and improvement. In I. K. L. a. P. H. (Eds.) (Ed.), *Second international handbook of educational leadership and administration* (pp. 9-40). Dordrecht, NL: Kluwer.

Halverson, R. (2007). A distributed leadership perspective on how leaders use artifacts to create professional community in schools. In L. Stoll and K. S. Louis (Eds.) *Professional learning communities: Divergence, detail and difficulties*. Maidenhead: Open University Press. 93-105.

Halverson, R., & Thomas, C. (2007). *The roles and practices of student services staff as data-driven instructional leaders*: Wisconsin Center for Educational Research.

Hamilton, L., Halverson, R., Jackson, S., Mandinach, E., Supovitz, J., & Wayman, J. (2009). *Using student achievement data to support instructional decision making*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hoover, J. J., & Patton, J. R. (2008). The role of special educators in a multitiered instructional system. *Intervention in School and Clinic, 43*(4), 195-202.

Karhuse, A. (2012, February 10). Ten states receive NCLB waivers. *Principal's Policy Blog,* National Association of Secondary School Principals. Obtained February 21, 2012 from http://nasspblogs.org/principalspolicy/tag/principal_evaluation/

Kelley, C. J., & Shaw, J. J. (2009). *Learning first! A school leader's guide to closing achievement gaps*. Thousand Oaks, CA: Corwin.

Knapp, M. S., Copland, M. A., & Swinnerton, J. A. (2007). Understanding the promise and dynamics of data-informed leadership. In P. A. Moss (Ed.), *Evidence and decision making* (pp. 74-104). Malden, MA: Blackwell.

Knapp, M. S., Copland, M. A., & Talbert, J. F. (2003). *Leading for learning: Reflective tools for school and district leaders*. Seattle: Center for the Study of Teaching and Policy, University of Washington.

Knowles, M. (1980). *The Modern Practice of Adult Education: From Pedagogy to Andragogy.* New York: Cambridge Books.

Lawrence-Brown, D. (2004). Differentiated instruction: Inclusive strategies for standards-based learning that benefit the whole class. *American Secondary Education, 32*(3), 34-62.

Leithwood, K., & Jantzi, D. (1999). Transformational school leadership effects: A replication. *School Effectiveness and School Improvement, 10*(4), 451-479.

Leithwood, K., Menzies, T., & Jantzi, D. (1994). Earning Teachers' Commitment to Curriculum Reform. *Peabody Journal of Education, 69*(4), 38-61.

Leithwood, K., Seashore-Louis, K., Anderson, S. E., & Wahlstrom, K. (2010). *Learning from Leadership: Investigating the Links to Improved Student Learning*. Minneapolis, MN: Wallace Foundation.

Means, B., Padilla, C. & Gallagher, L. (2010). *Use of education data at the local level: From accountability to instructional improvement*. U.S. Department of Education Office of Planning, Evaluation and Policy Development. Retrieved from http://www.ed.gov/about/offices/list/opepd /ppss/reports.html#edtech.

Merriam, S.B. & Cafferella, R.S. (1999). *Learning in Adulthood: A Comprehensive Guide, Second Edition.* San Francisco: Jossey-Bass.

Osher, D., Bear, G.G., Sprague, J.R., & Doyle, D., (2010, January). How can we improve school discipline? *Educational Researcher, 39* (1), 48-58, 49.

Quint, J. C., Akey, T. M., Rappaport, S., & Willner, C. J. (2007). *Instructional Leadership, Teaching Quality, and Student Acheivement: Suggestive Evidence from Three Urban School Districts*: MDRC.

Schwartz, F., & White, K. (2000). Making sense of it all: Giving and getting online course feedback. In K. W. White & B. H. Weight (Eds.), *The online teaching guide: A handbook of attitudes, strategies, and techniques for the virtual classroom* (pp. 57–72). Boston: Allyn and Bacon

Siskin, L. (1995). *The Subjects in Question: Departmental Organization and the High School*. New York: Teachers College Press.

Smylie, M.A., Bennet, A. with Konkol, P., & Fendt, C.R. (2005). What do we know about developing school leaders? A look at existing research and next steps for new study. In W.A. Firestone & C. Riehl (eds.), *A New Agenda for Research in Educational Leadership.* New York: Teachers College Press.

Spillane, J. P. (2005). Distributed Leadership. *The Educational Forum, 69*(2), 143-150.

Spillane, J. P., & Diamond, J. B. (2007). *Distributed Leadership in Practice*. New York, NY: Teachers College Press.

Spillane, J. P., Halverson, R., & Diamond, J. B. (2004). Towards a theory of leadership practice: A distributed perspective. *Journal of Curriculum Studies, 31*(1), 3-34.

Tomlinson, C. A., & McTighe, J. (2006). *Integrating differentiated instruction: Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Vaughn, S., & Linan-Thompson, S. (2003). What is special about special education for students with learning disabilities. *The Journal of Special Education, 37*(3), 140-147.

Waters, T., & Marzano, R. J. (2006). *School district leadership that works: the effect of superintendent leadership on student achievement*. Denver, CO: Mid-continent Research for Education and Learning (McREL).

Waters, T., Marzano, R. J., & McNulty, B. (2003). *Balanced leadership: What 30 years of research tells us about the effect of leadership on pupil achievement. A working paper.* Denver, CO: Mid-continent Research for Education and Learning (McREL).

Wayman, J. C., Midgley, S., & Springfield, S. (2006). *Leadership for data-based decision making: Collaborative educator teams*. Paper presented at the Annual meeting of the American Educational Research Association.

Wimer, S. (2002). The dark side of 360-degree feedback: The popular HR intervention has an ugly side. *Training and Development, 36(9), 37-44.*

Wood, F. H. & Thompson, S.R. (1980). Guidelines for better staff development. *Educational Leadership*, 37 (5) 374-378

Yang, M., Goldstein, H., Rath, T., & Hill, N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of Education, 25*(4), 469-483.