WCER Working Paper No. 2012-11
December 2012

The Role of Cognitive Validity Testing to Understand Leadership Practice in the Development of CALL, the Comprehensive Assessment of Leadership for Learning

Mark H. Blitz
Postdoctoral Fellow
Wisconsin Center for Education Research
University of Wisconsin–Madison
mblitz@wisc.edu

Jason Salisbury
Doctoral Candidate
Wisconsin Center for Education Research
University of Wisconsin–Madison
jsalisbury@wisc.edu)

WCER
Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison • http://www.wcer.wisc.edu/

WCER working papers are available on the Internet at http://www.wcer.wisc.edu/publications/workingPapers/index.php. Recommended citation:

> Blitz, M. H., & Salisbury, S. (2012). *The Role of Cognitive Validity Testing to Understand Leadership Practice in the Development of CALL, the Comprehensive Assessment of Leadership for Learning* (WCER Working Paper No. 2012-11). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Paper prepared for the University Council for Educational Administration Conference, November 16, 2012, Denver, Colorado. The research reported was supported by the U.S. Department of Education Institute of Education Sciences (Award R305A090265) and by the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison. Please do not cite without permission. Any opinions, findings, or conclusions expressed in this paper are those of the author and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions.

The Role of Cognitive Validity Testing to Understand Leadership Practice in the Development of CALL, the Comprehensive Assessment of Leadership for Learning

Mark H. Blitz and **Jason Salisbury**

Educators operate in an age of assessment. Consequently, researchers and policy makers work to develop assessments that accurately measure school success. A critical charge for developers of these assessments is to create tools that prove useful for the practitioners being assessed. Cognitive validity testing is one process to ensure assessment that is relevant to schoolwide practice. This paper discusses the importance of cognitive validity testing in the development of CALL, the Comprehensive Assessment of Leadership for Learning. Funded by the Institute for Educational Sciences (IES), CALL specifically focuses on using technology to support data-driven instructional leadership. Data-driven instructional leadership has become commonplace for school leaders, which has led researchers to examine the type of data leaders use. Student test scores and trait-based surveys illuminate areas of strength and weakness in schools, but school leaders do not have information on the work behind making improvements in those areas (Anderson, Leithwood, & Strauss, 2010). CALL is an online formative assessment of school leadership designed to examine schoolwide leadership tasks, and it uses a task-based distributed leadership framework to examine leadership practice across a school.

While the intent of developing a formative feedback assessment on distributed instructional leadership has been a noble one, the challenges of developing such a tool move beyond correlating items and identifying standards. Because CALL is task-based rather than individual- or trait-based, the challenges in developing this tool concern identifying specific leadership tasks and developing these tasks into universally accessible survey items. This paper discusses these challenges, as well as the solutions the research team devised. The presentation of the findings from this study is organized around three sample items used during cognitive validity interviews with CALL participants. Following the presentation of findings, we discuss the lessons learned that would contribute to the ongoing work of survey development, as well as provide further insight into schoolwide leadership practice. Our work was guided by the following research question:

> In developing a task-based 360-degree assessment of school leadership, how do practitioner user-testing experiences inform the work of capturing practices within a uniform survey?

Literature Review

Research has documented that school leadership plays a key, yet often indirect role in student learning (Leithwood & Riehl, 2005, Leithwood, Seashore Louis, Anderson, & Wahlstrom, 2004; Hallinger & Heck, 1996). More specifically, research has demonstrated that particular actions by school leadership play a vital role in allowing schools and teachers to meet

the learning needs of their students (Bryk & Driscoll, 1985). These actions include promoting a rigorous curriculum, developing a culture of mutual responsibility for student learning, and nurturing instructional practices that are effective in working with students.

The term *leadership for learning* has increased prominence in research related to educational leadership. However, the question remains as to how to understand or describe what leadership for learning looks like in everyday practice. Distributed leadership is a meaningful lens through which we can gain a ground-level understanding of leadership (Spillane, Halverson, & Diamond, 2004; Spillane & Diamond, 2007). Spillane and colleagues (2004) explain how leadership is situational and stretched across an organization, necessitating researchers to examine leadership activities and practices, as opposed to individual leaders; this shifts the unit of analysis from an individual to actions and practices. Furthermore, in using a distributed perspective, artifacts and situational factors become essential to the understanding of leadership (Spillane & Diamond, 2007). These artifacts or situational factors can include any type of tool or resource used in the process of leadership, such as a school improvement plan, a routine used to deal with specific situations, or a meeting agenda.

Given the importance of school leadership in promoting academic success, it becomes necessary to find ways to meaningfully assess school leadership practices and use that information to provide leaders with actionable knowledge that can impact their ability to improve practices within their organization (Thomas, Holdaway, & Ward, 2000). While the concept of *distributed leadership* helps us understand what we need to tease apart and examine regarding leadership for learning, it does not address the question of *how* we assess leadership for learning. One fruitful method for completing this task is through a research-based survey that specifically focuses on schoolwide practices, as opposed to individual leaders (Kelley et al, 2012). Surveys have been demonstrated to provide a meaningful snapshot of current practices within an organization related to specific phenomena (Basit, 2010). Basit (2010) notes that surveys can be used seamlessly to compare and contrast different groups within or across organizations, while providing timely reliable data. CALL positions itself as one such instrument (Kelley et al, 2012).

However, for CALL to be an effective research tool and formative feedback instrument for schools, the instrument must accurately capture leadership practices in schools. A large part of accurately capturing these practices is ensuring that survey participants fully understand what is being asked of them and that they are able to comment on the validity of survey material (Desimone & Le Floch, 2004). The idea of cognitive validity addresses this concern (Ruiz-Primo, Schultz, Li, & Shavelson, 2006). Cognitive validity is loosely defined as evidence of alignment between survey users' thoughts, beliefs, and feelings with the intended outcomes of a given survey instrument (Karabenick et al., 2007). Addressing issues of cognitive validity early on in survey development helps ensure participants' accuracy in understanding the meaning behind questions and their ability to accurately self-report individual and organizational practices (Mullens, 1998). Furthermore, addressing cognitive validity in survey development helps reduce personal biases and judgments when individuals respond to survey questions (Mullens et al,

1999). Beatty and Willis (2007) claim that the process of cognitive interviewing is as much art as it is science. While the method is intended to strengthen an instrument for quantitative research, cognitive validity testing is qualitative in nature.

At its heart, cognitive validity testing is concerned with how users experience a survey, whether or not they consider the items in alignment with intent of the survey designers, and if they believe the instrument adequately addresses the predetermined item construct (Karabenick et al., 2007). Karabenick and colleagues (2007) further state that the process of cognitive pretesting can be a valuable tool in cognitively validating surveys, and that cognitive pretesting is beneficial during the development phase of a survey instrument. Conducting focus groups and cognitive interviews early in the survey design process provides researchers a tool to increase the fidelity of their instrument to their research question. By interviewing participants during or soon after survey administration, survey developers can record the thought process behind responses (Jobe, Tourangeau, & Smith, 1993). However, there is little research on how to conduct pretesting, and the majority of research studies employing a survey design offer little explanation of their pretesting process (Presser et al., 2004).

Beatty and Willis (2007) emphasize the importance of thinking aloud and probing to uncover the users' thought processes. Beatty and Willis identify different types of probes that are both proactive and reactive in a given cognitive interview. Having the flexibility to engage participants in discussions around their thought processes supports the notion that each user brings a unique set of experiences and perspectives, and the interviewer should be an active listener in order to capture the particular thought process for each participant.

Collins (2003) identified four aspects of a "question-and-answer" model of cognitive methods: comprehension, retrieval of information, judgement, and response. In conducting cognitive pretesting, researchers should use these elements in seeking to understand a user's processes in responding to questions. Collins also asserts that because cognitive interviews are qualitative, they cannot provide quantitative evidence that a revised survey questionnaire is better suited to fulfill the intended goal. Therefore, it would be worthwhile to conduct quantitative reliability and Rasch analyses to determine the quality of the instrument.

In their research on developing an instrument to assess principal preparation across international contexts, Wildy and Clarke (2009) claim that "instruments, no matter how standardized, rest on certain assumptions" (p. 108). The challenge facing these researchers was to develop a survey instrument for users from 13 countries. English was not always the first language of the prospective users. As a result, identifying the appropriate terminology for the survey items was a recurrent theme within their findings. In addition, the researchers recognized the challenge of offering accurate response options for users. If a user did not *fully* identify with a particular response, the user did not select that response.

Most, if not all, of the literature on cognitive validity testing focuses on the method itself. The information gathered during the cognitive pretesting is used to refine the instrument;

however, researchers may also have an opportunity to learn more about their targeted area of inquiry through the development process. Previous research does not reference such opportunities. This study seeks to reveal these layers of findings, in addition to contributing to survey design methodology.

The complexity of the task within this study concerns conducting cognitive validity testing on a practice-centered survey. To create a continuum of responses for each survey item, a distributed leadership survey developer would need to guard against double-barreled questions and responses that attempt to present a spectrum of practice (Bassili & Scott 1996). Overall, creating a task-based rather than individual-based survey assessment presents various challenges, most of which the CALL research team was able to successfully address, as is recounted in this paper.

Methods

From the beginning stages of developing the CALL instrument, researchers sought practitioner expertise and input in constructing each survey item (Blitz, Milanowski, & Clifford, 2011). In Year 2 of the 4-year grant from IES, the CALL research team conducted a pilot test of the survey instrument. CALL researchers administered the survey in six schools in Wisconsin and conducted pre- and post-interviews. One purpose of the pilot study was to conduct cognitive validity testing to gather feedback from pilot participants on using the CALL system and taking the survey in order to inform survey revision and refinement. Soon after CALL was administered in each pilot school, CALL researchers conducted 24 interviews with principals, associate principals, teachers, department chairs, guidance counselors, and activities directors who took the survey. Researchers selected specific items to present to the participants during the interviews in order to observe the users' thought processes and rationale in choosing a response for each item. We coded the transcribed interview data to identify elements of the survey that would need to be addressed, such as language and syntax, items and relevancy to practice, and participant perspective versus item intent. The following section presents our findings.

Findings

As part of the cognitive validity testing, we presented the participants with specific items from the survey to help us refine the instrument. Within this process, we also learned how developing a task-based leadership assessment survey informs what we know about how leadership practice is perceived across various school community members. In the following section we present three of the survey items we gave to the participants. The findings are organized around these items since they either were the direct source of the findings or provide specific examples of more general findings. The following categories of findings are located within discussion of the three sample survey items:

- accessible language,
- leadership practice beyond school walls,
- socially desirable responding,

- working with 360-degree perspectives,
- applying appropriate terminology, and
- identifying appropriate practices.

Each sample survey item is presented as it was during this study. Current iterations of the items that reflect the necessary revisions can be found in the appendices.

Accessible Language

Item A, *Predictive Power of Formative Assessments*, shown in Figure 1, asks school leaders and staff about the school's use of formative assessment of student learning. One challenge for survey developers has been to create survey items artfully and succinctly while also not alienating the target population of practitioners. While academic researchers use language to describe practices in a school environment, that language may not be used by those operating in such an environment. This survey item illustrates that challenge, especially in the use of the term *formative assessment*. In responding to this item, the following participant conflated two different types of student assessments:

> Respondent: I answered "C" on that because being on the data team we were looking at the two different tests that were given, the 8th grade level, and then the sophomore level, and … it was almost comical because the 8th grade level was so easy, and then the 10th grade level … they're not even taught that unless they're in an advanced class.
>
> Interviewer: Mhm. And these are formative assessments?
>
> R: Yes.
>
> I: And what are they?
>
> R: The WKCE [state standardized test].

Figure 1
Item A: Predictive Power of Formative Assessments

The formative assessment program in our school:

a) Does not exist. (We don't have a schoolwide formative assessment program.)
b) Exists, but I don't know how well it predicts student performance on state tests.
c) Exists, but does not accurately predict student performance on state tests.
d) Exists and accurately predicts student performance on state tests.

The participant, a classroom educational assistant, responded more to the term *assessment* than *formative*. His school did use a schoolwide formative assessment program, but the school leaders do not refer to these assessments as "formative" even if that is the intended purpose. Other participants in other schools also conflated *formative* and *summative*. These

participants, as well as others, would have benefitted from embedded definitions of survey terms within the survey. However, this presented a dilemma for developers: how to create survey items that are succinct yet clear and accessible to practitioners.

There was a clear difference in semantics between survey developer language and practitioner language. Consistently, participants commented on the disconnect between the survey language and the language they use in their schools:

- "Vocabulary could have been easier, maybe. I'm just thinking of the average person, not necessarily myself, because I'm not … but I'm just thinking the vocabulary might have been a little bit stiff."
- "One of the things I'm talking about [is] dumbing down the test … maybe not as much 'researcher-ese.'"
- "And [the teachers] don't think that it relates to them, and this is somebody in an ivory tower kind of aspect. And so I think it has to be user friendly, and language is the first step."
- "I'm old school. I get caught up with the ... I get frustrated with the educator speak."

Other participants did not mind the presence of jargon but would have preferred that it reflect the wording they use in their district. These participants are responding to the task-based nature of the assessment, recognizing the effort to understand and capture specific leadership practices. While a purpose of a task-based survey is to focus on more discrete practices, the participants still wanted more specificity and relevance to their school culture.

Leadership Practice Beyond School Walls

One participant claimed that the school does not have a schoolwide formative assessment program, but when probed about a program that the school actually did use, the respondent reconsidered her response and confirmed that they did in fact have one. The difference was, she explained, that this was a district-wide initiative, something that they do not control in their school. Therefore, this item was more challenging for participants since the construct was further removed from their domains of practice. Moreover, participants would often cite that survey items inquired about leadership practices that were beyond the school walls. Especially in the case of the principals, participants expressed the need for the CALL survey to focus on district-level leadership practices as well:

> I found it somewhat frustrating because it asks questions with relation to … with regard to various initiatives or ways that we do things that here in this district are really more district-wide things. So we might say, you know, we disagree that that's not being done here, but it's not being done because it's not part of what we can do.

At the time of this pilot, the CALL survey was not offered to district leaders. To be sure, they would have insight into school-level leadership practice; however, they would more

effectively be able to speak to how district-level leadership practices impact school-level practice. This finding has already informed future work of CALL, as will be discussed later.

For the next iteration of CALL, the research team reconstructed this item and added a "filter" question to cut down on the cognitive demands of the survey and eliminate double-barreled responses. The current iteration of the item, located in Appendix A, contains an embedded definition of *formative assessment* to ensure universal understanding of the item construct.

Socially Desirable Responding

Another item discussed with participants, Item B shown in Figure 2, measured the construct, *Taking Responsibility for Student Learning*. This item represents a critical component of instructional leadership, and the findings yielded from the validity testing were just as informative as the survey data gathered after it was administered. As is the case with most surveys, whether on a Likert scale or not, responses for questions on the CALL survey were ordered in a continuum of least effective to optimal practices. Survey developers were aware of respondents' inclinations to select socially desirable responses: those that would reflect positively on them and their school. However, this item revealed that certain "trigger" words elicit socially desirable responses more so than placement of responses within the continuum. The following excerpt presents a participant's rationale for his selection of choice "B" rather than "C," the optimal response within this construct:

Figure 2
Item B: Taking Responsibility for Student Learning

Responsibility for learning for students who have been identified for special services, (e.g. special education or English language learning students):

a) Is regarded by classroom teachers as primarily the responsibility of instructional support staff.
b) Is regarded by classroom teachers as a shared responsibility between classroom teachers and instructional support staff.
c) Is regarded by classroom teachers as primarily their own responsibility and support staff aid in classroom learning designs.

> Interviewer: And why did you say B?
>
> Respondent: Because I think that anytime we have students, we are all a support team for that student. So it doesn't matter what our job is, our job is to help the students be successful. So if it's the support staff, it's the [Learning Disability] teacher, or the counselor, or whoever, you know? We're looking at the whole … I think we should be looking at the whole child, instead of content only.

The participant alludes to sharing responsibility among all school staff members who work with a given student. Although the CALL theory of action views the responsibility of

teaching all students, regardless of learning style, as that of the primary classroom teacher, the participants in this study were drawn to the trigger word "shared."

Responding according to social desirability was further evident in the following exchange in which a teacher reflected on his response as well as those of others:

> Respondent: Well, I would hope … it would be my hope that they would answer that it's "B," a shared responsibility.
>
> Interviewer: Right.
>
> R: But I don't know if people answered it [that way].
>
> I: So how did you answer it?
>
> R: I answered it by "B."
>
> I: Because…?
>
> R: That's just … well see, that's my background anyway. And so I do believe that it is a shared responsibility. It's just not one person responsible.

Often in this exercise, respondents (especially principals) were conflicted over whether to answer according to what they felt was *actually* happening in the school versus what they *hoped* was happening. Respondents were well aware of how the results would inform decision-making, and they considered who would be reading the results. This speaks to the formative quality of taking a practice-focused survey. Feeling inclined to select socially desirable responses calls attention to survey developers' use of language; it also allows for participants to communicate to school leaders and reflect on their own practice, thereby enacting the formative use of the instrument.

Working with 360-Degree Perspectives

The reflection on Item B also brought to light the ramifications of a 360-degree survey. All staff members in each participating school were invited to take the survey: all teachers, administrators, student support staff, and instructional support staff were eligible. A wider range of input would contribute to a clear picture of schoolwide leadership practices. However, with the advent of more discrete practices, certain participants expressed frustration that they could not speak to the practices about which the survey inquired. In discussing Item B, a school counselor stated the following: "I thought this had nothing to do with me. It has something to do with the staff, but answering it, like I said, personally and honestly I just couldn't relate to it. I really couldn't." To be sure, there are some items in the survey more accommodating to roles such as school counselors. However, other participants from other schools expressed similar complications with answering Item B. An Activities Director in another school responded thusly to Item B: "A teacher who might be teaching upper level courses, they (sic) have no idea what

our Special Ed population is like, what our ELL population is like, what our lower level learning population.… They don't have access to that, they don't know." It is unclear if the participant is alluding to shortcomings with the survey, the school staff's practices, neither, or both. What matters more, however, is that this item, through the cognitive validity testing alone, has validated its own existence. Another school counselor or administrator in this school or other schools may find this item much less problematic. Therefore, the participants' perspectives of this item are an important finding and data point for participating schools.

The current iteration of Item B, as shown in Appendix B, contains the same construct but with revised language. The trigger word "shared" has been removed, and the survey developers added emphasis to the key words that distinguish each response. The word "primarily" was added to the stem with emphasis in order for respondents to consider one role (e.g., teacher) versus another role (e.g., student support services).

Applying Appropriate Terminology

Item C, shown in Figure 3, focuses on the construct *Norms around Informal Leadership*. Measuring this construct is important in gauging school climate and trust. The wording of the item within the pilot version elicits tones of conflict and mistrust. However, the identification of the subject of the item, *informal leaders*, varied among participating schools. One participant raised this concern before being asked to consider Item C specifically:

Figure 3
Item C: Norms around Informal Leadership

In my school, **informal leaders**:

a) Often seem to thwart or undermine the instructional agenda of formal leaders.
b) Are typically not engaged with the instructional agenda of formal leaders.
c) Support formal leaders in efforts to advance the school instructional agenda.
d) Take the lead along with formal leaders to shape and advance the school instructional agenda.

> [W]hat was a struggle for me was that there weren't good definitions for some of the terms. For example, the one that I remember clearly was an "informal leader." What is that? Because you know one person can interpret that one way and somebody else … and I thought, 'Oh my gosh, am I an informal leader because I'm on the data team? I never signed up to be a leader but does that…?' You know, I don't know how other people perceive that term.

The definition of an informal leader varied from school to school and from person to person. One participant viewed an informal leader as the union representative, while another participant had a less formal definition: "People whose opinions, decisions, statements people tend to look up to. People [who] tend to have a certain charisma." Either definition would be correct since the meaning of the term depends upon the school community and culture. The "*who*" in this case, and throughout the CALL survey, is much less important than the "*what*." As

a result, the current iteration of Item C (Appendix C) does not contain the term *informal leaders*; rather, the focus in on faculty as a whole and their responses to changes in school.

Finding appropriate terminology was a consistent challenge across the survey. The need was for survey developers to identify universal terminology that would be accessible to any school in any context. For example, while some participants viewed special education teachers as instructional support staff, other participants viewed them as teachers. Therefore, it was critical for survey developers to either embed definitions in the survey item or use more general terminology.

Identifying Appropriate Practices

Respondents taking a survey with a standard Likert scale usually are able to select an option with which they feel relatively comfortable, especially when given a neutral option. In developing a survey that contains more discrete tasks, the CALL survey developers faced the significant challenge of not only identifying the appropriate practices but also determining the appropriate range. For example, in reflecting upon Item C, a participating principal examined each response and felt that she could justify each response as the appropriate one because all practices occurred within the building. With the addition of the phrase "in general," the participant would be directed to identify the typical response rather than the "correct" one. Therefore, the CALL survey developers made that necessary change.

While some items like Item C presented options that were all plausible, participants made claims to the contrary regarding other items. When asked to rate the difficulty of taking the survey on a scale of 1–4, with "1" being easy, one participant offered the following explanation:

> I would say a two or almost a three just because sometimes the responses, like … none of the four seemed to really depict what I thought was going on here. And so I had to spend some time thinking about which ones sort of more, you know, was more closely aligned to what I thought of what we were doing. There were a lot of times where I felt like it would be better if there were five or six options so it would be a little more nuanced.

The respondent's desire for the item responses to reflect more of "what we were doing" presents perhaps the greatest challenge for developers of task-based surveys of discrete practices. It would be impossible to capture the nuances of the various forms of the same basic practice across all schools. Even when asked to select the option that "best" describes a certain practice, respondents still expressed reticence in selecting an option that did not fully capture the characteristics of the practice.

To be sure, there are advantages and disadvantages in constructing a distributed leadership assessment rather than an individual trait-based survey. Developing a survey that is universally accessible presents one challenge. Developing a practice-based survey presents yet another challenge. Engaging in cognitive validity testing contributes to mitigating the complexity

of such an endeavor. Survey developers need to make compromises in certain areas of the process, as is discussed in the following section.

Lessons Learned

In conducting this pilot study for CALL, researchers acquired input for refining the instrument, but also gained insight into the process of measuring leadership practice, especially when holding the perspective that leadership is not limited to a single individual. Constructing a non-traditional survey tool produces a number of challenges in addition to the inherent values. The following section presents an informal "cost-benefit" analysis of constructing and using a 360-degree task-based assessment of instructional leadership. Based on the findings, we identified three tradeoffs that accompany the development of such a tool. This insight should contribute to the work of developing leadership assessment tools.

Tradeoff One: Focusing on Leadership *Practice*

Developers of CALL have been guided by this theory of action: In order to authentically measure leadership effectiveness, one would need to examine the actual practices of leadership, practices that extend beyond an individual. As a result, the CALL survey inquires about areas that delve more deeply into larger and more general areas of practice. The benefit of such a tool is many-fold. For one, participants engage in a formative reflection process while taking the survey. They are led to consider specific instances of interaction, teaching, feedback, and collaboration. In adopting a distributed leadership model, the CALL instrument presents items situated beyond the principal's office. As a result, teachers reflect on their own practice as teacher leaders, informal leaders, or as classroom teachers only. A consequence of such a feature, however, is that teachers would gravitate toward more socially desirable responses, especially if the inquired-about practice would reflect upon them more than those in formal leadership positions. Incorporating the 360-degree feature of the instrument would likely balance the socially desirable responses from the more realistic ones, but that feature also presents a significant tradeoff as will be discussed in the next subsection.

Furthermore, items to be rated according to a continuum of practice are less reliable than a standard Likert scale, according to psychometric analyses (Camburn & Salisbury, 2012). To be sure, creating a universal scale for all survey items would likely increase reliability; it would also reduce the richness of the data for school leaders. The more general the items, the less specific the feedback would be for schools. In addition, having items scored on a 5-point scale, but not necessarily on the same 5-point scale throughout the survey, provides both challenges and opportunities for participants. While a given score in one domain of CALL would not necessarily align with another domain score, the participants have the opportunity to delve more deeply into these results at the item level to unpack the reasons behind the scores. This would be a benefit to those who relish data use, and a potential burden on those who want more summative results without contributing their own inquiry to the analysis.

Tradeoff Two: Maximizing and Limiting the 360-degree Quality

The most effective leadership assessments are those that incorporate multiple perspectives (Goldring, Porter, Murphy, Elliott, & Cravens, 2009). The CALL survey incorporates ratings from *all* instruction-related staff *within* the school building. This means that central office staff are not eligible to take the current iteration of CALL, and that a wide array of staff within the building do take the survey, even if they are not privy to some of the practices inquired about on the survey. This study revealed that a number of staff who do not work within a classroom, or even teachers who work in other areas of the school (both physically and subject-wise), are not always aware of what is happening throughout the building. While this could be a consequence of operating within large comprehensive institutions, it also demonstrates the need for leaders to engage all staff members into the everyday functions of the school and to make public the practices that are generally kept private. While support staff perspectives are important in measuring leadership effectiveness, it is also important for school leaders to be able to disaggregate results by school position in order to understand how perspectives vary according to role and which staff members are more in tune with the functions of the school. In a forthcoming paper, CALL researchers will report on variation by role in CALL.

Regarding central office participation to even more fully round out the 360-degree quality of the survey, CALL researchers concluded that the tasks within the survey items would be overly discrete for those not in the building to be able to rate. Their participation would likely reflect the dilemma of having instructional support staff take the survey as well: the further removed from the locus of teaching and learning, the more difficult it would be to rate practices impacting that area. District leaders could, however, rate leadership tasks at the district level, thereby contributing to a larger picture of leadership practice as it extends beyond the school walls. As a result of this study, CALL researchers have begun developing a version of CALL for district leaders that will correspond to the current iteration of CALL. The instrument will measure task-based district leadership practice in support of school-level leadership practice.

While having a 360-degree component is essential for thorough assessment of schoolwide leadership, it also presents the challenge of incorporating perspectives of people that may not have the knowledge of or experience within given leadership domains. Whether one believes that they *should* have the capacity to speak to all functions within a school, this does not always align with the reality that the instrument is measuring.

Tradeoff 3: Balancing Semantics

A challenge for any survey developer who wishes to release an assessment tool into a large diverse realm is to use culturally relevant and accessible language. As shown in this study, users of the tool were often frustrated and grew disengaged with the survey due to its academic language and non-universal terms. As is the case with the prior subsection regarding whether one *should* be “in the know” or not, it does not matter if researchers and developers believe participants *should* be familiar with certain key terms. A familiar term in one state or district could be unfamiliar in another state or district, and two different terms may carry the same

definitions. To ensure universal understanding, survey developers need to incorporate definitions of key terms. However, these definitions need to be visible and prominent within a given survey item. A “mouse-over” function in which certain terms have been flagged with a hyperlink in order for users to view a pop-up definition would not ensure exposure to the critical information. Even parenthetical information is subject to inattention. Therefore, within the current iteration of the CALL survey, all terms are defined before or within the stem of the item. While this ensures the likelihood of universal understanding, it also increases the cognitive demand for taking the survey. This tradeoff is the consequence of a task-based survey in that users have more to read in order to be able to speak to the constructs within the survey. This tradeoff, as well as the preceding tradeoffs, are byproducts of survey design. They not only inform survey development but also provide further insight into the culture of schooling and schoolwide distributed leadership.

Conclusion

In developing assessment tools, each design decision is accompanied by tradeoffs and compromises. Survey developers need to consider these within a cost-benefit analysis. Engaging participants in the iterative design of a tool would greatly inform developers in this analysis. While this study explored the results from the cognitive validity testing of a formative assessment tool measuring distributive leadership, the lessons learned should also promote the value of conducting cognitive validity testing in the development of any tool for practitioner use. Schools are inundated with assessments and evaluations, most of which are for policy makers’ purposes. In order for these assessments to serve the ultimate purpose of school improvement, they must be directly beneficial to those charged with generating the improvement. Furthermore, if a given assessment is intended *for* practitioners, then the assessment should be *by* practitioners, as much as is reasonably possible. Practitioners should be given the tools and opportunities to contribute to the instruments by which they will be measured and assessed. In a practice-based survey, the input of those engaged in the practices is invaluable. Researchers and developers should continue to challenge assumptions about survey design in order to create tools that benefit multiple stakeholders. Engaging in processes to develop such tools will produce tradeoffs, to be sure, but it also will produce insight into school-level practice and practitioner engagement in design.

References

Anderson, S., Leithwood, K., & Strauss, T. (2010). Leading data use in schools: Organizational conditions and practices at the school and district levels. *Leadership and Policy in Schools, 9*(3), 292–327.

Basit, T. N. (2010). *Conducting research in educational contexts.* New York, NY: Continuum International Publishing Group.

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*(3), 390–399.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*(2), 287–311.

Blitz, M., Milanowski, T., & Clifford, M. (2011). Content validity as a window to a richer understanding of leadership practice. Annual Conference of the American Education Research Association. New Orleans, LA. April 2011.

Bryk, A. S., & Driscoll, M. E. (1985). *An empirical investigation of the school as community*. Chicago: University of Chicago, Department of Education.

Camburn, E., & Salisbury, J. (2012). Examining the validity and reliability of the comprehensive assessment of leadership for learning (CALL) formative leadership assessment tool: Pilot study results. Annual Conference of the American Education Research Association. Vancouver, BC. April 2012.

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research* 12, 229–238.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26*(1), 1–22.

Firestone, W. A., & Riehl, C. (2005). *A new agenda for research in educational leadership.* New York, NY: Teacher College Press.

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.

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

Jobe, J. B., Tourangeau, R., & Smith, A. F. (1993). Contributions of survey research to the understanding of memory. *Applied Cognitive Psychology, 7*(7), 567–584.

Karabenick, S., Woolley, M., Friedel, J., Ammon, B., Blazevski, J. B. Bonney, C. R., … Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean, *Educational Psychologist, 42*(3),139–151.

Kelley, C., Halverson, R., Camburn, E., Blitz, M., Salisbury, J., Dikkers, S., & Clifford, C. (2012). *The design and validation of the comprehensive assessment of leadership for learning (CALL) formative school leader assessment.* Paper prepared for the Association of Educational Finance and Policy Conference, Boston, Massachusetts, March 17, 2012.

Leithwood, K. and Riehl, C. (2005). What we know about successful school leadership. In W. Firestone and C. Riehl (Eds.), A new agenda: Directions for research on educational leadership. (pp. 22-47). New York: Teachers College Press.

Mullens, J. E., Gayler, K., Goldstein, D., Hildreth, J., Rubenstein, M., Spiggle, T., … Welsh, M. (1999). Measuring classroom instructional processes: Using survey and case study fieldtest results to improve item construction. Working paper series.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*(1), 109–130.

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching, 38*(2), 260–278.

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, 36*(1), 3–34.

Thomas, D. W., Holdaway, E. A., & Ward, K. L. (2000). Policies and practices involved in the evaluation of school principals. *Journal of Personnel Evaluation in Education, 14*(3), 215–240.

Wildy, H., & Clarke, S. (2009). Using cognitive interviews to pilot an international survey of principal preparation: A Western Australian perspective. *Educational Assessment, Evaluation, & Accountability*. *21*(1), 105–117.

APPENDIX A:
CURRENT ITERATION OF ITEM A: Predictive Power of Formative Assessments

The next question asks you to think about a formative assessment program that may be in place at your school. A *formative assessment program* includes schoolwide policies, plans or practices in which teachers regularly gather information on student learning to help design curriculum and form strategies for instruction.

Does your school have a schoolwide formative assessment program?

a. Yes
b. No [SKIP NEXT QUESTION]

Which of the following *best* describes how well the results from formative assessments in your school predict and improve student performance on the state standardized test?

a. I do not know how well the results from formative assessments predict student performance on the state test.
b. Results from formative assessments do not accurately predict student performance on the state test.
c. Results from formative assessments accurately predict student performance on the state test.
d. Results from formative assessments accurately predict student performance on the state test *and* help to improve student performance on the state test.

APPENDIX B:
CURRENT ITERATION OF ITEM B: Taking Responsibility for Student Learning

In most classes in your school, who is *primarily* responsible for teaching students who have been identified as having a specific learning disability?

a. No one takes primary responsibility for teaching these students.
b. The special education teacher.
c. The special education *and* the classroom teacher, but the *special education teacher* develops the classroom learning plans.
d. The special education *and* the classroom teacher, but the *classroom teacher* develops the classroom learning plans.
e. The classroom teacher, with the special education teacher supporting the design and delivery of instruction.

APPENDIX C:
CURRENT ITERATION OF ITEM C: Norms around Informal Leadership

In general, how do teachers and staff respond when school leaders introduce significant changes that affect classroom instruction in your school?

a. School leaders do not introduce significant changes.
b. Teachers and staff work against significant changes.
c. Teachers and staff are generally indifferent to significant changes.
d. Teachers and staff generally support significant changes.
e. Teachers and staff generally work with school leaders to make significant changes.