Comparing CALL and VAL-ED:

An Illustrative Application of a Decision Matrix for Selecting Among Leadership Feedback Instruments

WCER Working Paper No. 2015-5
October 2015

Peter Goff
Wisconsin Center for Education Research
University of Wisconsin–Madison
pgoff@wisc.edu

Jason Salisbury
Iowa State University

Mark Blitz
Wisconsin Center for Education Research
University of Wisconsin–Madison

Comparing CALL and VAL-ED:
An Illustrative Application of a Decision Matrix for Selecting Among Leadership Feedback Instruments

Peter Goff, Jason Salisbury, and Mark Blitz

In the current "Age of Information" and era of high-stakes accountability, policymakers as well as state and district leaders promote the use of data to inform leadership decision-making and professional development (Elmore, 2000; Halverson, Grigg, Prichett, & Thomas, 2007; Robinson, Lloyd, & Rowe, 2008). This focus on effective school leadership has been widely recognized as an essential strategy to advance student learning (Leithwood & Seashore-Louis, 2011; Marzano, Waters, & McNulty, 2005). Therefore, education researchers have sought to develop assessment tools that provide information on the effectiveness of school leadership while concurrently providing data to school leaders to promote leadership improvement. While many of these assessment tools are grounded in existing research on leadership and teaching practices that promote student learning, the multitude of available instruments leaves schools and policy makers in the position of having to select an assessment tool with little guidance. This lack of guidance creates the potential for a situation in which schools and policy makers may select instruments based on market-based influences as opposed to research-based factors.

The privileged space that accountability movements have afforded leadership assessments heightens the need for schools and policy makers to have strategies at their fingertips to assure that leadership assessment tools are selected in a manner that meet place-specific needs and preferences. The privileging of assessment tools in today's educational landscape positions leadership feedback instruments as a driving force in school improvement processes, professional development discussions, and high-stakes policy implementations within schools nationwide. Furthermore, multisource feedback instruments are frequently used as sources of data to assess the effectiveness of all levels of programs in educational organizations (Guthrie, 1990; Oakes, 1986). The goal of this paper is to present a decision matrix that educators, policy makers, and researchers can use when selecting a leadership assessment tool. Our decision matrix encourages potential consumers of leadership assessments to consider the psychometric properties of the assessment; the model of leadership; the contextual relevance of the system; and the actionability of the system. The paper concludes by illustrating the application of our decision matrix through a comparison of two prominent measures of school leadership: the Comprehensive Assessment of Leadership for Learning and Vanderbilt Assessment of Leadership in Education.

Current Trends in Performance Indicators

Schools across the United States have faced increased levels of scrutiny in the form of performance indicators or accountability assessments during the last 30 years in response to forces such as the *A Nation at Risk* commission report and No Child Left Behind legislation. Much of this pressure has been exerted through accountability systems aimed at instructional

practices and student learning, but there is an increasing focus on assessing school leadership practices due to the strong connection between school-level leadership and student learning (Camburn, Huff, Goldring, & May, 2010). This movement can be thought of as an attempt to recouple localized educational practices to external expectations (Diamond, 2012, 2007; Hallett, 2010; Spillane & Burch, 2006). Recoupling schools through accountability processes can be explained through the bureaucratic-rational choice model (Spillane, Diamond, Hallett, Halverson, & Burch, 2002), which argues that schools and school personnel shift practices based on accountability mandates and their subsequent incentive structures. In this system, performance indicators become the instantiation of accountability pressures leading to individual and organizational change (Bryk & Hermanson, 1993).

Performance indicator systems can be seen as tools that enable districts, policy makers, and researchers to collect and access data related to an organization's performance and the relationships among key components of the organization (Burnstein, Oakes, & Guiton, 1992; Selden, 1994). Indicator systems are employed to achieve five goals: (1) relating the current performance of an organization; (2) advancing various policy agendas; (3) serving as the basis for accountability systems; (4) assessing programs or policies; and (5) functioning as information management systems (Ogawa & Collom, 2000). In thinking about leadership assessment tools in education as performance indicators, their function can be to meet the above five goals, often more than one at the same time.

A byproduct of system recoupling is the hegemonic control that performance indicators maintain over school practices (Bryk & Hermanson, 1993). Because of their external legitimacy and political importance, performance indicators often drive internal conversations related to school improvement, professional practice and learning, curriculum, and student learning. School leaders know that failure to move their organization in accordance with the indicator targets risks sanctions and increased bureaucratic pressure. As a result, conversations about professional practice and professional growth with schools are often heavily influenced by the set of performance indicators the state, district, or school select. Put another way, the professional dialogue within a school is inextricably coupled to the measured performance indicators. Historically, these indicators have focused on shifting the practices of educational leaders by focusing on the core technology of teaching and learning, but current movements are focused on more directly assessing leadership practices that research has connected to increased student learning (Grissom, Kalogrides, & Loeb, 2015). Heightened performance indicator focus on leadership practices within schools will undoubtedly strengthen the control that these instruments have over professional conversations within schools. This paradigm makes selection of the best leadership assessment instrument a critical step in shaping school policy and educator development.

Surveys have a long history as a method for collecting data related to performance in education leadership (Hallinger & Heck, 1996), and criticism of their use includes concerns about users' abilities to accurately recall information or behaviors being investigated and potentially high levels of skipped or unanswered questions (see Burnstein et al., 1995; Hilton,

1989; Levine, Chambers, Ixtlac, & Hikido, 1998; Rubin & Baddeley, 1989 for a more detailed description). However, Burnstein et al. (1995) also note that since the mid-1980s, "the quality of education indicators has steadily improved, particularly in (sic) indicators of school and classroom processes" (p. xiv). Furthermore, surveys that generate composite indicators[1]—a common practice among leadership assessment surveys—demonstrate the ability to differentiate between professionals who engage in specific types of practice and those who do not (Mayer, 1999). In a review of the reliability and validity of survey use in educational research, Desimone and Le Floch (2004) concluded there was ample evidence that surveys provide "meaningful, substantive, and informative data" related to practices within schools. Regardless of the established appropriateness of surveys as a performance indicator for school leadership within an accountability context, there is still a pressing need to align the performance indicator being used to the needs of the given practitioners, policy makers, or researchers.

Selecting an Appropriate Feedback Instrument

As noted, multiple feedback instruments exist for assessing the practices of school leaders. The taxing issue is selecting a tool that provides schools with information that is appropriate for their needs and context, a decision often made based on psychometric properties of the instrument, state evaluation policies, or the marketing strategy of various instruments. The sections below outline a decision matrix for schools, policy makers, and researchers to consider when selecting an educational leadership assessment tool. This matrix includes four decision points: (1) psychometric properties of the instrument; (2) model of leadership assumed by the instrument; (3) feasibility of implementation; and (4) actionability of the feedback.

Psychometric Properties of the Feedback Instrument

A basic requirement for any feedback instrument designed to enact changes in professional practice within a school is that the data collection instrument be designed in a psychometrically sound manner. In fact, Ogawa and Collom (2000) reference validity and reliability as the most commonly listed standard that educational indicators are held to. A failure to achieve this threshold results in schools receiving feedback that lacks trustworthiness and clarity of focus. Broadly speaking, indicatory instruments need to be designed and tested to ensure that gathered data are reliable—they consistently generate similar results in similar circumstances—and valid—they measure the desired variable. Indicator instruments that measure leadership practices in a valid and reliable fashion are more likely to provide practitioners, policy makers, and researchers with data that represent what a given school is accomplishing in relation to a desired set of indicators or standards.

Validity. Validity refers to the ability of the instrument to measure what it is intended to measure. This requirement is of particular importance for measures of leadership, which are

---

[1] When using the term composite indicators, we reference the practice of combining multiple related survey items into a single score. For example, a composite score of instructional leadership could include individual items related to a leader's ability to provide feedback on instructional practices, manage the professional development agenda of a school, facilitate curricular decisions, and engage in sense-making activities related to data.

inherently latent—that is, in contrast to manifest measures such as years of experience or certification, they cannot be directly and objectively ascertained. Rather, measures of leadership are determined through triangulation: surveys ask multiple questions about manifest behaviors, policies, and practices with the understanding that no one item fully captures the latent construct of leadership, yet the common covariance among these items can.

There is no singular, definitive test of instrument validity. Instead, the validity of an instrument is determined by accumulating evidence across various aspects of validity. Here we focus on the most prominent sources of evidence for validity, as advocated by the American Educational Research Association's *Standards for Educational and Psychological Testing* (2014): face validity, content validity, concurrent validity, and predictive validity.

Face validity asks "Does this instrument look as if it measures school leadership?" and is determined through inspection, by the judgment of experts. As an instrument with tangible implications to change the educational experience of children, face validity helps to ensure that teachers and principals view the measure as legitimate and use the instrument to provide careful, thoughtful feedback. Content validity asks "Does this instrument capture all the important aspects of school leadership?" Often content validity is determined through a comprehensive review of a given construct and then comparing instrument topics to review topics. Content validity provides an assurance that the measure adequately captures the breadth and complexity inherent in school leadership. Concurrent validity asks "Can this instrument be used to discriminate among school leaders?" Concurrent validity uses a sample of individuals whose school leadership varies in known ways and ascertains the extent to which the proposed measure can distinguish among them. This aspect of validity in a leadership instrument is particularly important if the instrument is to be used for human capital management decisions, such as performance pay or strategic staffing. Predictive validity asks "Does this instrument predict meaningful change in outcomes of relevance?" Evidence of predictive validity could be demonstrated by improvement on this measure corresponding to improvement in, for example, retention of exemplary teachers, teachers' instructional growth, or students' sense of well-being or belonging. Predictive validity is a critical validity component as it provides key insights into how behaviors today may affect children tomorrow.

Reliability. Reliability reflects the precision of an instrument: an instrument's ability to replicate results with low measurement error. As with validity, reliability is determined through multiple strategies. Here again we refer to the American Educational Research Association's *Standards for Educational and Psychological Testing* (2014) to underscore three approaches to estimating instrument reliability: inter-rater reliability, test-retest reliability, and internal consistency. Each of these common measures presents some challenges to interpretation. We find it may be helpful to note that, in the case of multi-source feedback instruments, measures of inter-rater reliability may not as useful as typically portrayed. The rationale is that different teachers may experience leadership differently. These differences would be reflected in a low measure of inter-rater reliability, and yet instruments that can identify this variation across teachers can provide meaningful data to the leadership team (Goff, 2013).

4

Test-retest reliability is a useful measure to quantify the stability of an instrument, yet there is no agreed upon minimum acceptable value for the correlation between measurements. As with inter-rater reliability, strong test-retest reliability may be detrimental when examining multi-source feedback measures in practice. One aim when using multi-source feedback may be to document change over time; instruments that demonstrate high test-retest reliability measures may be attuned to aspects of leadership that are fixed, and they may have difficulty measuring changes in leadership and organizational practices.

The last commonly used measure of reliability, internal consistency, is typically reported as Cronbach's alpha, with acceptable values cited as greater than 0.70 (Cortina, 1993). Internal consistency is a measure that reflects the ability of the items within a scale to collectively measure the same underlying construct. One well-documented challenge of Cronbach's alpha as a measure of internal consistency is the artificial inflation of alpha with increasing items (Cortina, 1993). Thus, all else being equal, scales with more items will have greater reliability than scales with fewer items.

Model of Leadership Assumed by the System

Assessments of leadership often assume a theoretical model of how the world they are measuring works (descriptive) or should work (prescriptive). Adopting a particular model privileges the theoretical assumptions and values of the system designers over alternative theoretical assumptions and values. This status is especially true in an area such as educational leadership; assessments of school leadership are grounded in research related to school leadership and research-based practices. However, multiple research bodies advance "best-practice" school leadership. Traditionally, these models have included domains such as instructional leadership, transformational leadership, and shared/democratic leadership. More recently, theories of distributed leadership and socially just leadership have entered into the lexicon of school leadership. At some level the theoretical underpinnings of an assessment tool may seem academic or esoteric, but the theory driving the instrument focuses the instrument toward some aspects of leadership and away from others. For example, an assessment grounded in the research on instructional leadership will place greater value on monitoring teaching and learning and less value on establishing a common vision. In contrast, an assessment grounded in research on transformational leadership will place greater emphasis on creating a common vision and less on monitoring teaching and learning. This is not to say that any one instrument is going to be representative of one theory of leadership to the exclusion of all others, but rather that that a given instrument will lean one way or another. Individuals selecting an assessment instrument should ensure they understand its theoretical leanings and align its theoretical foundation with the needs and values of the organization. In the remainder of this section we underscore the key characteristics of five prominent conceptions of school leadership: instructional, transformational, shared, distributed, and socially just.

Instructional leadership. These instruments will assign increased weight to leaders who focus on their leadership on areas identified in research related to instructional leadership. As such, leaders who focus their energies on monitoring student progress, creating high expectations

for teachers and students, assessing instructional practices, and aligning curriculum (Barth, 1986; Hallinger & Murphy, 1987; Marks & Printy, 2003) would receive ratings or performance scores that indicate they are meeting their organizational responsibilities. Furthermore, as the paradigm of instructional leadership places the work of the principal at the center of effective schools, the assessment's underlying assumption would be that high quality leaders are engaged in these practices as opposed to sharing or delegating the responsibilities to their colleagues. Such delegation could result in a principal receiving lower ratings if other individuals within their schools engage in instructional leadership.

Transformational leadership. When notions of transformational leadership undergird an assessment of principals, privilege is assigned to work focused on vision and mission setting, identification of areas in need of improvement, development of school improvement plans, and encouraging broad participation from multiple stakeholders in the decision-making process (Avolio & Bass, 1995; Hallinger, 1992; Marks & Printy, 2003). This emphasis creates conditions where principals or schools who value or prioritize different forms of leadership are inherently disadvantaged on the assessment. Embedded in assessments of transformational leadership is the belief that effective schools are led by individuals who work to influence school culture and motivate others to engage the challenges of school improvement.

Shared leadership. Assessments grounded in ideas of shared leadership will center on the work of the principal, but also recognize that an important aspect of the principal's job is to share leadership responsibility with other members of the school community (Lambert, 2002; Pearce & Conger, 2002; Printy & Marks, 2006). Within the realm of shared leadership focus can vary between instructional leadership and transformational leadership; the important components are the recognition that the principal is not the sole hub of leadership and that the principal shares leadership work with members of the organization. Due to these underlying beliefs, indicator systems adhering to shared leadership would positively assess leaders who effectively share their work with members of their school. In other words, the assessment includes how effective the leader is at bringing stakeholders into the world of educational leadership. Shared leadership does not necessarily favor instructional leadership or transformational leadership. However, there is a growing body of research related to shared or distributed instructional leadership (see Bredeson, 2013; Kelley & Salisbury, 2013; Klar, 2013; and Printy & Marks, 2006, for examples).

Distributed leadership. While often thought of as similar to shared leadership, distributed leadership is a separate understanding of leadership within organizations. Whereas shared leadership focuses on the intentional distribution of leadership across individuals, distributed leadership advances a perspective that leadership is inherently distributed or stretched across an organization (Gronn, 2002; Spillane & Diamond, 2007; Spillane, Halverson, & Diamond, 2004). The important underlying assumption in distributed leadership is that regardless of intent, the work of leadership is distributed throughout an organization; hence, to understand leadership we cannot focus on individuals (Spillane, Halverson, & Diamond, 2004). Instead, investigations of school leadership need to focus on the tasks or activities in which leaders engage. This

understanding of leadership shifts the focus of an assessment of leadership from the individual principal, or leadership team, to the collective work of the school in engaging in requisite leadership activity. Leadership assessments centering distributed leadership will shift the focus from the principal to the organizational capacity to engage in leadership as the theory centers the work of leadership over the leader.

Socially just leadership. A burgeoning field within educational leadership is centered on how principals are able to advance socially just practices within schools. Socially just leadership has been theorized in multiple ways (see Brown, 2004; Dantley & Tillman, 2006; Bogotch, Beachum, Blount, Brooks, & English, 2008; Jean-Marie, Normore, & Brooks, 2009; Shields; 2004; Theoharis, 2007), but common throughout all is the idea that socially just educational leaders focus on ameliorating inequities in opportunities and outcomes for traditionally minoritized populations. Traditionally minoritized populations includes groups historically oppressed within our society such as people of color, individuals identified as (dis)abled, individuals whose first language is not English, individuals who do not identify as heterosexual, or individuals who are economically disadvantaged. Performance indicators assuming a leadership for socially just theoretical stance will focus on the work of school leaders to create equitable learning opportunities for minoritized populations via their work to promote culturally relevant practices, minimize overrepresentations in programs like special education, reduce the overrepresentation of students of color affected by school discipline, and eliminate tracking practices. Leadership assessments from a socially just perspective will score leaders poorly who do not intentionally focus on improving educational opportunities for all students.

While for the purposes of this paper we have presented various theories of leadership as distinct categories, many times these models are merged in some fashion. For example distributed leadership and leadership for social justice have been brought together into a model of distributed leadership for social justice (Brooks, 2012; Brooks, Jean-Marie, Normore, & Hodgins, 2008) or the previously mentioned work bringing instructional leadership and shared leadership together. However, architects of an assessment of leadership will foreground certain ideals of leadership, which are intrinsically linked to their theoretical assumptions of leadership. This approach can be seen as the result of epistemological forces working in tandem with assessment system constraints. Assessments can be only so long—we cannot ask school stakeholders to take a 12-hour survey—so content that designers feel is most important to the work of leadership is included while other material is excluded.

The end result is that schools, policy makers, and researchers need to select an instrument with a theoretical stance that aligns with their goals. For example, if a school district is interested in understanding a principal's impact on the instructional core, then they should select an assessment that foregrounds instructional leadership. But, if that district is interested in how effective a principal is at addressing racial inequities, than a tool focused on social justice may be the best candidate. Individuals or organizations using assessment instruments related to school leadership need to understand the theories supporting different systems to ensure consequential

validity and ensure meaningful dialogues can occur based on the information provided from the assessment.

Feasibility of Employing the Assessment Instrument

In many ways feasibility is the simplest decision point for schools. Feasibility includes multiple technical issues, such as time commitment, cost, mode of data collection, data source(s), and turn-around time. The two most basic issues of feasibility are how long the assessment takes to complete and the overall cost of employing the assessment. Schools may not want to use an assessment that takes two hours per person to finish or diverts needed funds from areas of need. The medium of the assessment also impacts feasibility: School staff may prefer to complete a web-based instrument as it provides increased flexibility, but this approach requires access to computers and adequate internet bandwidth and speed, which still represent notable constraints in some isolated, rural districts. Sources of data become a feasibility consideration for multi-source leadership instruments: Are teachers the lone source of data, or are students, parents, community members, district leaders, or school board members included? While increasing the variety of individuals providing data increases the robustness of the assessment, it also increases the degree of difficulty in collecting the data. The frequency of assessment also determines a tool's feasibility; ideally institutions, researchers, or policy makers would not rely on a single snapshot of a principal's performance in making decisions. As a result, they need to understand how often data will be collected. And finally, the turn-around time related to results is important for schools; for assessment information to maintain its consequential validity schools need to have access to it in a timely manner. While feasibility is a simple yes/no decision point, it is essential for users to think through whether a instrument is feasible for their district prior to selection; otherwise it becomes possible to imagine a district selecting an assessment instrument that takes more time to implement than they are willing to give, which could influence the validity of the results.

Actionability of the Feedback

The final decision point for school districts, researchers, and policy makers in selecting an instrument to assess school leadership is to look at the actionability of the data provided. Actionability refers to the ways in which the feedback or data provided to schools or other users can be used toward intended ends. From this perspective, feedback constitutes information or data that drives future processes (Senge, 2006; Greve, 2003). If the feedback provided by an instrument does not provide actionable data to an organization, it becomes useless or misinformative as schools and their leaders work toward organizational improvement (Halverson, 2010; Senge, 2006). Users of leadership assessments need to be able to engage in a processes of reflection on feedback as it relates to existing beliefs, knowledge, and experiences, otherwise known as sense-making (Spillane, Reiser, & Reimer, 2002). Feedback from the assessment tool needs to allow users to create actuation spaces (Halverson, 2010) that will enable the school to clarify organizational goals, cultivate steps to achieve those goals, and improve specific leadership actions and behaviors toward achieving those goals. If an indicator system is unable to achieve this level of actionability, than it fails to meet a basic intended outcome an

assessment system. Value-added measures have been criticized as a performance measure with low actionability (Goldring et al., 2015) since they provide scant direction to educators to improve their practice.

Actionability will look different for different organizations; much of the difference will be based on internal capacity to work with various forms of data. This situation means that user of assessment systems need to understand the type of feedback provided and their organization's capacity to interpret and act upon that feedback. For example, if an assessment of school leadership combines quantitative and qualitative feedback, the users need to have the ability to synthesize those two types of data into actionable steps toward organizational improvement. If the assessment system provides strictly quantitative data in the forms of means and variations, school staff need to be able to sift through those data, and create an action plan for leader and organizational improvement. As part of assessing the actionability of potential leadership assessment systems, schools need to investigate system supports for helping users engage in sense-making activities as well as actuation.

Finally, while no assessment of school leadership will be entirely tailored to the existing agenda of leadership within a school, assessments must measure a principal's effectiveness at working toward existing goals. Failure to do so will result in data being provided to a school, researcher, or policy maker that lacks actionability due to missing contextual relevancy necessary to making informed improvement decisions. Poorly contextualized data could result in schools and leaders scrapping existing initiatives due to mistakenly believing a lack of information about initiatives equates to poorly designed or implemented initiatives. Such behavior would be an unfortunate artifact of an assessment tool that was not designed to capture their work in a given area. Schools changing directions every few years rather than allowing their existing agendas to take hold and making intended changes is a well-documented phenomena (Newmann, King, & Youngs, 2000).

Thus far we have described a matrix of decision points that schools, policy makers, or researchers can use in selecting an appropriate assessment system of school leadership. This decision matrix helps potential users investigate the assessment's psychometric properties, its theoretical stance on school leadership, the feasibility of using it, and the actionability of the feedback the assessment provides. We believe that if schools evaluate potential leadership assessments along these four dimensions, the information schools gather will have heightened relevance, guide their improvement discussions in meaningful ways, and help combat the potentially pernicious effect of indicator systems within school improvement discussions. Throughout the remainder of this manuscript we demonstrate how our matrix could be employed to evaluate two commercially available and research-based assessments of school leadership—the Comprehensive Assessment of Leadership for Learning (CALL) and the Vanderbilt Assessment of Leadership in Education (VAL-ED).

Methods

As researchers, schools, districts, or states engage in a formal process of instrument selection, we advocate they consider following four domains: psychometric properties, the underlying theory of leadership, feasibility of implementation, and actionability. In the following sections we contrast two measures of school leadership—CALL and VAL-ED. We chose these two for their prominence in the field and familiarity to the research team. Two of us have been involved with the development of the CALL survey (Blitz, Salisbury, & Kelley, 2014), while the other co-author had experience with VAL-ED through a study examining the role of coaching and feedback on leadership practices (Goff, Goldring, & Bickman, 2014; Goff, Guthrie, Goldring, & Bickman, 2014). Our collective experience allows us to present a detailed juxtaposition of these two measures through the decision matrix outlined above.

To investigate the relationship between VAL-ED and CALL, we collected data from both surveys as part of the validation of the CALL instrument in 2012. As part of this work, we correlated CALL data to other indicators of school effectiveness including individual principal performance as measured by VAL-ED. We administered CALL in 100 schools and VAL-ED in 30 of those schools, resulting in approximately 900 participants (teachers and school leaders) who took both the VAL-ED survey and the CALL survey. Each survey is scored on a five-point scale with "5" being the highest possible score.

We present correlational and descriptive findings from a comparison of these two surveys administered in a sample of 30 schools. Although such a detailed comparison is beyond the immediate scope of our decision matrix (we realize few organizations will be able to pilot both instruments concurrently), we feel that the national prominence of these measures merits the additional empirical comparison.

Sample

As part of the validation of CALL in 2011, researchers sought to identify correlations between measurements of CALL and measurements of individual principal leadership. Of the 100 schools in 2011 that piloted the CALL survey, 30 schools also administered the VAL-ED survey. These schools were either mostly rural schools in Mississippi or suburban, rural, or urban schools in Wisconsin. Schools that did not opt to administer the VAL-ED survey in addition to the CALL survey cited time constraints and survey over-saturation as the primary reasons.

Findings

Psychometric Properties

The market claims of an instrument's validity are easily made, but they can be challenging to substantiate. CALL and VAL-ED are presented as research-based instruments that are valid and reliable. Both measures have engaged in pilot testing and cognitive interviews with participants to clarify language and improve reliability. Researchers have undertaken several studies to validate VAL-ED (Porter et al., 2010; Polikoff et al., 2010) and CALL (Kelley et al., 2012; Blitz,

Salisbury, & Kelley, 2014), showing these instruments to be robust measures of their intended constructs across multiple contexts. One element that appears to be missing from validity studies is a rigorous inquiry into the predictive validity of these instruments: When principals move the needle on one of the VAL-ED core components or CALL domains, do we see a subsequent change in teacher practice or student outcomes?

Both CALL and VAL-ED comprise subscales to ensure content validity and to facilitate feedback interpretation. The five subscales of the CALL are referred to as leadership domains; the six subscales of VAL-ED are referred to as core components. In our empirical comparison of the instruments' reliability measures Cronbach's alpha was 0.95 or higher for all six VAL-ED core components; reliability for CALL domains ranged from 0.75 to 0.89, all above the established minimum of 0.70.

Our final consideration regarding reliability pertains to the structure of response options used throughout the survey. In recent work on rubric design Humphry and Heldsinger (2014) argue that that there is no *a priori* reason why rubrics should all have the same number of response categories. The authors find that this design feature (matrix rubrics) induces raters to give more similar scores across items, rather than selecting the category that best describes the individual's response performance. In short, when each prompt/item has the same number of response options and when these are worded in the same or similar manner, it is more likely that an assessment will receive more of one particular value (e.g., mostly 2s).

Although a survey response set is distinct from a rubric, when surveys are used to evaluate a performance task, the distinction between rubrics and survey responses begins to blur. The response options for VAL-ED are consistent across all 72 items—all responses vary from "Ineffective" to "Outstandingly Effective." In contrast, the CALL survey also uses Likert-style response options, however the number of response options vary by item, ranging from three to five. The structure of these items also varies to reflect the specific practice in the question stem.

If we view each survey item as a task and the response options as a small, condensed rubric that describe a particular leadership performance task, then the finding that varied response options increase rating validity (Humphry & Heldsinger, 2014) suggests an advantage of CALL over VAL-ED. One result of the uniform VAL-ED item response structure may be an artificial inflation of internal consistency (Cronbach's alpha), and some of the information that might be gleaned from individual items is muted. CALL, on the other hand, will have a lower reliability as compared to VAL-ED, perhaps because the item response structure of CALL is better suited to picking up item-level variation. This difference in item construction remains an important and distinguishing feature, and we will return to this point when we examine actionability. In the next section we turn examine the theoretical foundations of the two instruments.

Underlying Theory of Leadership

VAL-ED's theory of school leadership. In developing VAL-ED, Goldring, Porter, Murphy, Elliot, and Cravens (2009) created an orthogonal conceptual framework consisting of *Core*

Components and *Key Processes* to assess the effectiveness of an individual school principal. To identify the appropriate constructs on which to assess school leadership, the VAL-ED researchers referred to seminal works on instructional leadership (i.e., Hallinger & Heck, 1996; Heck & Hallinger, 1999). The VAL-ED instrument is a 360-degree survey in which in a given school the teachers, the principal, and the principal's supervisor answer questions about the principal's instructional leadership capacity.

Drawing from the work of Goldring and colleagues (2009), we present here the six core components that serve as the constructs for measuring the effectiveness of the primary school leader in VAL-ED. Each of these six core components is measured across six key processes (planning, implementing, supporting, advocating, communicating, and monitoring).

High standards for student learning. This component emphasizes school leaders setting clear goals for student learning. In addition, these goals need to be high in quality and marked by high standards. School leaders need to maintain high expectations for all students to achieve. They must effectively communicate these expectations across the school in order for teachers to share those expectations and maintain those standards.

Rigorous curriculum. High standards and expectations must be accompanied by a curriculum that is ambitious in content across all subjects. This core component emphasizes the importance of instruction in enriching student learning and conveying a rich curriculum. Effective school leaders work closely with teachers in this area and position themselves as curriculum experts (Murphy, Elliott, Goldring, & Porter, 2006). Working with teachers, school leaders ensure instruction is rigorous and aligned with the high standards identified in the previous core component.

Quality instruction. While definitions and ideas of *instructional leadership* vary, it is this area where school leaders work most directly with the art of instruction. Goldring and colleagues (2009) define *quality instruction* as "effective instructional practices that maximize student academic and social learning" (p. 10). Teachers should be expected to communicate clearly, hold high expectations for all students, and monitor progress of student learning. Moreover, effective school leaders, to support quality instruction, must provide useful and specific feedback to teachers.

Culture of learning and professional behavior. This core component emphasizes the cultivation of communities of professional practice. Effective school leaders develop professional learning communities that focus on teaching and learning specifically and regularly. To be sure, professional learning communities have become more commonplace in schools, but they also may deviate from their intended purpose. Effective learning communities share goals, focus on student learning, and engage in reflective conversation around teaching practice.

Connections to external communities. This core component aligns with the research base that has reported significant benefits between family involvement in school and the social and academic benefits for students (Henderson & Mapp, 2002). Moving beyond passive

participation, effective community involvement in this area includes procuring social services that are linked to the school community, support for parents, and initiatives to organize the school community at large (Mediratta & Fruchter, 2001). Moreover Goldring and colleagues (2009) identify effective leaders in this area as those who develop relationships with the business, political, and religious leaders in the school community.

Systemic performance accountability. Last, this core component reflects current education policy and the presence of high-stakes accountability. This construct recognizes external accountability and its impact on leadership. At the same time, accountability exists in the form of local expectations. Therefore, it is incumbent on school leaders to balance the various forms of accountability by ensuring school staff implement the initiatives that promote student learning and comply with mandated policy.

CALL's theory of school leadership. While VAL-ED emphasizes the instructional role of the school principal, other school leadership scholars have sought to accomplish a similar task through a different theory of action in developing CALL. Similar to VAL-ED, CALL utilizes a multi-source survey to measure leadership (Kelley et al., 2012). However, rather than focus on the individual school leader, CALL utilizes a distributed leadership framework (Spillane, Halverson, & Diamond, 2004).

One challenge of utilizing and discussing research on distributed leadership is isolating the exact usage of such a framework. Researchers and practitioners alike have widely adopted distributed leadership. The term itself is accessible and supports sensibilities of collaboration and employee empowerment (Harris, 2008). CALL researchers adopted a distributed leadership model conceptualized and promoted by Spillane, Halverson, and Diamond (2001, 2004). According to these scholars, distributed leadership provides a lens with which to understand and analyze leadership rather than support a specific approach to leadership. This model moves away from a leader-centric model. Furthermore, Spillane (2005) does not promote a singular leadership style to accompany a distributed leadership perspective: "a distributed perspective allows for leadership that can be democratic or autocratic" (p. 149).

Given that CALL utilizes a distributed leadership lens that focuses on leadership as a function, we will now look at the five core domains wherein the leadership practices reside. It is worth noting explicitly that CALL does not measure the extent of leadership distribution within a school, but rather operates on the assumption that leadership *is* distributed. CALL seeks to measure the spectrum of leadership practices that function across the school. Halverson, Kelley, and Shaw (2014) describe CALL's five core domains in greater detail in their work. We introduce them here:

Focus on learning. This domain focuses on the work of school leaders to regularly engage the school community in conversations around instruction and student learning. In this area, school leaders seek to address problems with teaching and learning through a collaborative

process, promote a clear vision for student learning outcomes, and prioritize supporting the learning of students who traditionally struggle.

Monitoring teaching and learning. Within this area of leadership practice, school leaders emphasize continuous, formative assessment of student learning that influences instruction. In addition, the school utilizes summative data to inform decision-making, while ensuring that high-stakes standardized testing preparation and results complement the larger educational program. In monitoring teaching, school leaders provide ongoing feedback to teachers to support professional growth. Also, the formal evaluation of teachers needs to be more than for compliance purposes: the generated feedback needs to be specific and useful.

Building nested learning communities. This domain recognizes various ways to cultivate professional learning communities. Effective school leaders build in opportunities for teacher collaboration around instructional issues. Also, within this area, teachers and teacher leaders participate in school improvement planning and decision-making processes. Lastly, utilizing peer coaches and mentors lends to the cultivation of a professional learning community in which expert teachers support their colleagues.

Acquiring and allocating resources. Within this domain, a resource takes on various forms. Effective school leaders supply *time* for teachers to plan and address student learning issues together. Also, in effective schools, teachers know that school leaders consistently work to procure *funding* to support teacher-based innovation. School leaders also utilize the *expertise* that exists in and outside of the school building for professional development. Finally, effective school leaders recognize the school *community* at large as a resource to support students and to cultivate a positive school culture.

Maintaining and safe and effective learning environment. Taking into consideration Maslow's (1943) hierarchy of needs, this domain emphasizes the importance of a learning environment in which students and teachers feel safe and secure to engage in learning. This domain focuses on the school discipline policy and the extent to which it is fair and effective and the extent to which those who apply the policy do so indiscriminately. In addition, this area of leadership practice focuses on supporting those students who traditionally struggle and who may need more instructional and social support systems.

Feasibility of Implementation

The first element of feasibility pertains to duration: How long does it take to complete the survey? VAL-ED comprises 72 items, not including optional ancillary questions, such as those on teaching position and experience. In contrast, 210 items constitute the core leadership items for CALL. Because of the uniform structure across all the VAL-ED response options, the per-item completion rate is higher for VAL-ED than for CALL. This same phenomenon may lead to somewhat diminished response variation and inflated reliability. This psychometric flaw becomes a feasibility feature as the VAL-ED can be completed in substantially less time,

approximately 20 minutes as compared to 45 minutes for CALL. Both instruments have fast turn-around times, providing comprehensive feedback reports within days of the survey closure.

Administration of VAL-ED and CALL is exclusively online. The web-based feedback systems typically have slightly lower response rates and thus provide diminished representation of perspectives across the faculty. In our experience shifting from written responses to online administration drops response rates from 89% to 78%, although this trend varied somewhat across schools. We found no systematic differences in response groups by teaching domain, demographics, or years of teaching experience.

In a study using VAL-ED (Goldring, Mavrogordato, & Haynes, 2014), teachers and principals alike emphasized the need for the principal to clearly communicate the important need for teachers to complete the surveys and to express how the feedback would be used. In schools where this communication did not take place, response rates were lower and responses were more homogeneous. The point we hope to convey here is that principals must express their intentions with regard to the survey and the survey results clearly to teachers *at the time that they are requesting the survey be completed.* This is true for written administrations, and particularly true for electronic versions, where communication regarding the survey may be indirect.

While cost structures for the two measures are somewhat fluid, shifting with contract duration and size, the financial implementation costs of CALL and VAL-ED are of a similar order of magnitude. While neither measure is free, the annual costs for both are well within the constraints of most school budgets; for most schools the total for employing either instrument would be less than $1,000. This places both instruments in a space where they are affordable for the majority of school districts in the United States, especially when considering the wealth of information that both tools to provide to users.

Both the CALL and VAL-ED surveys require significant time to administer. With a strong alignment between the needs and values of the instrument and the organization, the data collected are more likely to be useful and to be seen as useful. This utility—the ways in which data can be translated into action—is the topic of our next section.

Actionability of Feedback

Actionability characterizes features of the instruments that facilitate a transformation from survey data into action. When considering how actionable survey results might be, we underscore the importance of item construction and feedback reports. Our rationale for this focus is that the structure of the items in many ways determines the types of feedback that can be collected, and the nature of the feedback determines how the information can be translated into action.

Item construction focuses on the aim and structure of the individual survey items. We focus here on the response emphasis and the response options. For example, a survey question may focus on a leaders' observing instruction, emphasizing the quality of that activity. Another item

may be constructed to emphasize the frequency of observations. A third way of constructing this item may ask respondents to consider how observations are typically conducted. Many surveys adopt a particular approach to item construction and apply that approach uniformly throughout the survey. The utility of survey items to the end user often depends on the response emphasis. Continuing the example from above, schools that employ a uniform, scripted observation protocol may be less interested in the *quality* of observations (since these are standardized) and more interested in the *quantity* of observations. Conversely, schools that endorse an observation strategy that relies on professional judgment, such as a coaching model, may find that items constructed with a quality focus lend themselves to greater actionability.

The VAL-ED is structured such that each of the 72 items has the same five-item Likert-style effectiveness response scale (minimally effective to outstandingly effective). The VAL-ED items are global—that is, they speak to the leadership behavior broadly and do not specify the exact task that may be used in any given school. For example, one VAL-ED item asks "How effective is your principal at ensuring the school … supports teachers in meeting school goals?" Another asks "How effective is your principal at ensuring the school … encourages students to successfully achieve rigorous goals for student learning?"

The 210-item CALL survey also uses Likert-scale response options, however the number of response options vary by item, ranging from three to five. The structure of these items also varies to reflect the specific practice in the question stem. An example is provided below:

> Which of the following <u>best</u> describes the development of a common language regarding instruction between you and your colleagues?
>
> a) We do not typically talk to each other about instruction.
>
> b) We typically talk about instruction, but have not developed a common language.
>
> c) We have developed a common language to talk about instruction <u>within</u> our subject area or specialty.
>
> d) We have developed a common language to talk about instruction <u>across</u> subject areas or specialties.
>
> e) We have <u>fully</u> developed a common language to talk about instruction <u>across</u> subject areas or specialties.

These differing approaches to item construction have inherent benefits and trade-offs. The global approach adopted by VAL-ED allows items to cover a broad swath of behaviors, without having to specify the individual tasks. This strategy allows more information to be gleaned from fewer items, resulting in a shorter survey. By not specifying the exact behavior, the items are more likely to include a wide range of leadership practices. The cost of this approach is that respondents need to have a clear and shared understanding of what it means for a principal to be

outstandingly effective at "creating expectations that faculty maintain high standards for student learning." The recipient of the feedback needs to be aware how respondents understand these global items within their local context. It may be the case that "outstandingly effective" means something different to a veteran teacher, who has worked for and experienced many leadership behaviors, than it does for a novice teacher. Research on reference dependence suggests that individuals almost always have a reference group in mind when weighing various options (e.g., Kőszegi & Rabin, 2006; Lurie & Mason, 2007). If the reference group is not known, then feedback can become ambiguous (Is the principal effective relative to other principals in the district? Relative to other principals the teacher has worked with? Relative to a predefined standard of excellence?). Clearly, such ambiguity inhibits the actionability of feedback, yet this potential liability can become an asset if the survey administration is coupled with professional development and discussion.

In contrast to VAL-ED's global approach, the items on the CALL survey are specific and context dependent. One cost to this approach manifests in the expanded size of the survey—nearly three times as long as VAL-ED. This approach is also limited in that a list of specific behaviors or policies can never be fully exhaustive and logistically feasible. The CALL research team has had to select the most prominent behaviors for inclusion in the final survey, and still runs the risk that principals supporting a similar set of related behaviors may be unnoticed. The benefit to clearly articulating the spectrum of behaviors for each item is twofold. First, items focused on specific leadership practices are less likely to suffer from heterogeneous or ambiguous reference dependence, making them easier to interpret. A second, and related, benefit to using specific rather than global items is that feedback from specific items can be more readily made actionable.

In this section we have applied our decision matrix to the CALL and VAL-ED surveys, examining psychometric properties, conceptions of leadership, feasibility of implementation, and actionability. We find reason to believe that both instruments are valid and reliable, although the ultimate evidence—does improvement on these measures lead to improved outcomes for students—is lacking for both. Both instruments are anchored in an instruction-focused approach, with VAL-ED's learning-centered approach focused primarily on the principal, while the CALL portrays leadership to be distributed across individuals and embedded in organizational structure. Feasibility for measures is similar as costs appear to be comparable; VAL-ED appears most useful when additional time has been invested to establish a collective vision for the interpretation of the broad, global items and this additional time investment is earned back with shorter administration durations. This paradigm is also reflected in the actionability of surveys as the descriptive, specific nature of items in the CALL survey generates feedback that lends itself readily to interpretation and discussion.

Correlation Analysis

To better understand relationships between CALL and VAL-ED, we conducted a correlation analysis among the five CALL Domains and the six VAL-ED Core Components. Table 1 presents the results of that analysis. We see that CALL Domain 1 (Focus on Learning) and

Domain 3 (Building Nested Learning Communities) correlate most highly with VAL-ED Core Components. Conversely, the VAL-ED components are fairly consistent in correlations to each of the CALL Domains. We do see that CALL Domain 5, *Maintaining a Safe and Effective Learning Environment,* yields a low correlation to each of the VAL-ED components.

Table 1. **Correlations between VAL-ED Core Components and CALL Domains**

		VAL-ED Core Components					
		High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning and Professional Behavior	Connections to External Communities	Systemic Performance Accountability
CALL Domains	Focus on Learning	0.580**	0.555**	0.643**	0.554**	0.517**	0.636**
	Monitoring Teaching and Learning	0.564**	0.410	0.402	0.445**	0.460**	0.507**
	Building Nested Learning Communities	0.652**	0.491**	0.672**	0.599**	0.550**	0.607**
	Acquiring and Allocating Resources	0.513**	0.424	0.483**	0.446**	0.454**	0.476**
	Maintaining a Safe and Effective Learning Environment	0.460**	0.356	0.292	0.287	0.315	0.354
Stars indicate statistically significant correlations at the 0.05 (*), 0.01 (**), and 0.001 (***) levels.							

Four or five subdomains reside within each of the five core domains of CALL, and we conducted a correlation analysis of each subdomain with the VAL-ED Core Components. One finding that stood out was the relationship among CALL subdomain 1.2, *Formal leaders are recognized as instructional leaders,* and VAL-ED Core Components (see Table 2). This subdomain is the only one in CALL that focuses specifically on the principal of the school, thereby aligning more closely with the VAL-ED theory of school leadership. This finding is notable, and we present it because it provides additional evidence of the representative theoretical foundations of each instrument.

Table 2. **Correlations among the sub-domains of CALL Domain 1 (Focus on Learning) and VAL-ED Core Components**

		VAL-ED Core Components					
		High Standards for Student Learning	Rigorous Curriculum	Quality Instruction	Culture of Learning and Professional Behavior	Connections to External Communities	Performance Accountability
CALL Domain 1 Subdomains	1.1: Maintaining a School-wide Focus on Learning	0.356	0.363	0.500**	0.342	0.294	0.443**
	1.2: Formal Leaders are Recognized as Instructional Leaders	0.782**	0.636**	0.766**	0.756**	0.681**	0.752**
	1.3: Collaborative Design of Integrated Learning Plan	0.419	0.450**	0.540**	0.431	0.406	0.523
	1.4: Providing Appropriate Services for Students who Traditionally Struggle	0.327	0.356	0.292	0.287	0.315	0.354

Stars indicate statistically significant correlations at the 0.05 (*), 0.01 (**), and 0.001 (***) levels.

Discussion

Indicative of the education policy era in which we reside, *assessment* and in turn *evaluation* are common terms used in conversations about school and district practice. State legislatures across the country are mandating implementation of educator effectiveness systems; states are receiving waivers for flexibility from the federal No Child Left Behind Act of 2001; and Title I schools are required to demonstrate a plan of action to address the gaps that resulted in their designation of "priority." These situations for public schools, combined with the growing desire to assess leadership effectiveness, leads to increased attention on how best approach leadership assessment and evaluation systems. A clear nexus has emerged at the intersection of professional growth and accountability. Through the development of the framework within this paper, we present opportunities for schools, local education agencies, and state education agencies to engage in leadership assessment and evaluation that not only complies with state and/or federal mandates, but also provides valuable feedback and data to the subject of the evaluation.

Of course, "value" is inherently subjective. Education leaders need to consider their own context when implementing an evaluation system. The two tools highlighted in this paper, VAL-

ED and CALL, meet the criteria outlined and explained in this paper. While CALL and VAL-ED meet these criteria independently, they differ from one another in notable ways. Our decision matrix has allowed us to identify several of these differences, such as theoretical underpinnings and item specificity. VAL-ED primarily focuses on the actions of the principal, thereby utilizing an individual-based approach to assessing leadership. This approach seems logical given the singular nature of leadership roles: while many schools incorporate formal sub-leadership roles (i.e. associate principal), the role of principal is the most common reference in leadership rubrics, standards, and legislation.

CALL utilizes a distributed leadership framework to measure leadership, thereby moving away from the individual leader in a school. The CALL survey items focus on specific tasks in a school, but do not usually position the role of the principal as the focus of the question regarding these tasks. In an individual-based approach, a survey may inquire about the effectiveness of the principal in a given task, such as implementing professional development. For an approach based on distributed leadership, the survey would inquire about the nature of task itself, in this case the effectiveness of the professional development activities. As a result, the response options for such a question consist of potential practices that describe said professional development activities. These response options affect data and feedback within the backend of the assessment in that a principal could examine an actual item for input on school improvement planning. The theoretical underpinnings underscore the clear differences between these two instruments.

These theoretical differences are manifest in the measures we collected, as evidenced by Table 1, where we can see that the two measures share a common foundation and each contribute unique information that the other does not.

We look to the correlation analysis between VAL-ED and CALL data to find further similarities. Table 1 reveals that CALL Domains 1-3 yield high and significant correlations to the VAL-ED Core Components. Domain 4 of CALL has a somewhat weaker relationship to the VAL-ED core components, and Domain 5 appears to provide information unique to CALL. This speaks to the research that influenced the development of these tools: The CALL constructs that begin to stray from the VAL-ED constructs are more commonly considered in a school climate survey. Having these constructs situated within a survey of school leadership reflects the CALL developers' perspective that school climate is a function of the distributed nature school leadership.

When examining the correlation between the CALL subdomains and the VAL-ED Core Components in Table 2, we see that the CALL subdomain that yielded the highest correlation to VAL-ED was the only area of CALL that focused on the primary individual leader. This finding tells us that the analysis has picked up on the differences between an individual-based assessment and a distributed leadership-based assessment, as the other distribute leadership-based subdomains are not as highly correlated to VAL-ED. This credible evidence shows the conceptual foundations of the respective measures are not simply academic—they are manifest in the instruments themselves.

The implementation of a leadership assessment tool is complex, to be sure, and when novice principals are under consideration, this process becomes that much more complicated. Focusing on professional growth opportunities, one should look to the actionability of an instrument to support a principal new to the position. In this case, providing a new principal individual-based leadership effectiveness data from her/his predecessor would not necessarily provide much guidance to the principal on what action to take moving forward. At the same time, novice principals may have a tougher time using CALL feedback due to its specificity, which may require a modicum of experience to put into action. On the other hand, the specificity of the CALL items may be more appealing to novice principals because it provides a clear and defined path to improvement. In either case, the novice and/or new principal would benefit from ongoing guidance from a mentor who can spend time with the feedback reports and data to facilitate sense making and action planning.

The quality of feedback to practitioners will ultimately determine the extent to which a leadership assessment tool has served its intended purpose. District leaders looking for tools that alleviate their responsibilities as principal supervisors who provide support to their charges will find this search to be futile. The data that result from CALL and VAL-ED are different, but they are similar in that school leaders require additional support and guidance as they interpret the information. For CALL, the resulting data may be actionable in that the information measures and reports on discrete practices occurring throughout the school and in the classrooms, but school leaders may find the data to be incongruent with personal experience. And, regarding the actionability of the data, there is a difference between being given information on where to focus for improvement and knowing specifically what to do in that area. District leaders are in a position to provide that support. They are also in the position to take data that are a bit more broad and work with principals and other leaders on distilling information to narrow the focus of their work going forward. VAL-ED offers a more global view of leadership practice that should promote dialogue among district and school leaders on what is happening in a given area and why they could do to make progress in that area.

While both tools measure and provide feedback on school leadership, there are clear differences between CALL and VAL-ED. However, this is not to say that one approach should be considered "better" or more valuable. In fact, a school or district could consider implementing both CALL and VAL-ED to measure school leadership, perhaps alternating by season or year. By adopting such an approach, a school leader would receive data on schoolwide practices in addition to more specific individual practices. This could in turn promote individual professional development plans, the active distribution of leadership, and more specific targeted areas for school improvement planning. Both VAL-ED and CALL measure instructional leadership practices, both highlight exemplary practices found in education leadership research, and both were developed through grant-funding from the Institute for Education Sciences.

Conclusion

This paper has outlined and discussed a decision matrix for examining school leadership assessment tools. With the growing need to assess and evaluate school leaders, it is important for

decision makers to be aware of several criteria when adopting a tool. We put this decision matrix to work by examining two prominent leadership assessment tools: VAL-ED and CALL. By applying the above decision matrix we determined differences between instruments in several areas, notably the underlying theory of school leadership and item construction. The correlation analysis showed that these tools also have quite a bit in common. Therefore, a central factor to be considered when adopting a leadership assessment system is the organization's context and goals. To be sure, any tool that meets the criteria of this paper's outline would provide value to the school in which it is implemented. Of course, any instrument that is chosen should not be utilized as a stand-alone evaluation or development tool. Rather, it should be part of a larger comprehensive evaluation system—a system that supports the practitioner(s) that is being evaluated and that provides key information to the supervisor.

References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, D.C.

Avolio, B. J., & Bass, B. M. (1995). Individual consideration viewed at multiple levels of analysis: A multi-level framework for examining the diffusion of transformational leadership. *The Leadership Quarterly*, *6*(2), 199–218.

Barth, R. S. (1986). The principal and the profession of teaching. *The Elementary School Journal*, *86*(4), 471–492.

Blitz, M., Salisbury, J., & Kelley, C. (2014). The role of cognitive validity testing in the development of CALL, the comprehensive assessment of leadership for learning. *Journal of Educational Administration*, *52*(3), 358–378.

Bogotch, I., Beachum, F., Blount, J., Brooks, J., & English, F. (Eds.). (2008). *Radicalizing educational leadership: Dimensions of social justice*. Rotterdam, The Netherlands: Sense Publishers.

Bredeson, P. V. (2013). Distributed instructional leadership in urban high schools: Transforming the work of principals and department chairs through professional development. *Journal of School Leadership, 23*(2*)*, 362–388.

Brooks, J. S. (2012). *Black school, White school: Racism and educational (mis)leadership*. New York, NY: Teachers College Press.

Brooks, J. S., Jean-Marie, G., Normore, A. H., & Hodgins, D. W. (2008). Distributed leadership for social justice: Exploring how influence and equity are stretched over an urban high school. *Journal of School Leadership, 17*(4), 378–408.

Brown, K. M. (2004). Leadership for social justice and equity: Weaving a transformative framework and pedagogy. *Educational Administration Quarterly*, *40*(1), 77–108.

Bryk, A. S., & Hermanson, K. L. (1993). Educational indicator systems: Observations on their structure, interpretation, and use. *Review of Research in Education, 19*, 451–484.

Burnstein, L., McDonnell, L. M., Van Winkle, J., Ormseth, T., Mirocha, J., & Guitón, G. (1995). *Validating national curriculum indicators*. Santa Monica, CA: Rand.

Burnstein, L., Oakes, J., & Guiton, G. (1992). Education indicators. *Encyclopedia of Educational Research*, *2*, 410–418.

Camburn, E. M., Huff, J. T., Goldring, E. B., & May, H. (2010). Assessing the validity of an annual survey for measuring principal leadership practice. *The Elementary School Journal*, *111*(2), 314–335.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*(1), 98–104.

Dantley, M. E., & Tillman, L. C. (2006). Social justice and moral transformative leadership. In C. Marshall & M. Oliva (Eds.), *Leadership for social justice: Making revolutions in education* (pp. 16–30). Boston, MA: Allyn & Bacon.

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.

Diamond, J. B. (2007). Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction. *Sociology of Education*, *80*(4), 285–313.

Diamond, J. B. (2012). Accountability policy, school organization, and classroom practice: Partial recoupling and educational opportunity. *Education and Urban Society*, *44*(2) 151–182.

Elmore, R. F. (2000). *Building a new structure for school leadership*. Washington, D.C: Albert Shanker Institute.

Goff, P. (2013, November). *Response distribution in multi-source feedback*. Paper presented at University Council for Educational Administration, Indianapolis, Indiana.

Goff, P., Goldring, E., & Bickman, L. (2014). Predicting the gap: perceptual congruence between American principals and their teachers' ratings of leadership effectiveness. *Educational Assessment, Evaluation and Accountability*, *26*(4), 333–359.

Goff, P., Guthrie, J., Goldring, E., & Bickman, L. (2014). Changing principals' leadership through feedback and coaching. *Journal of Educational Administration*, *52*(5), 682–704.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*(2), 96–104.

Goldring, E. B., Mavrogordato, M., & Haynes, K. T. (2014). Multisource principal evaluation data principals' orientations and reactions to teacher feedback regarding their leadership effectiveness. *Educational Administration Quarterly*, *51*(4), 572–599.

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), 1–36.

Greve, H. R. (2003). *Organizational learning from performance feedback: A behavioral perspective on innovation and change*. New York, NY: Cambridge University Press.

Grissom, J. A., Kalogrides, D., & Loeb, S. (2015). Using student test scores to measure principal performance. *Educational Evaluation and Policy Analysis, 37*(1), 3–28.

Gronn, P. (2002). Distributed leadership as a unit of analysis. *The Leadership Quarterly*, *13*(4), 423–451.

Guthrie, J. W. (1990). The evolving political economy of education and the implications for educational evaluation. *Educational Review*, *42*(2), 109–131.

Hallett, T. (2010). The myth incarnate recoupling processes, turmoil, and inhabited institutions in an urban elementary school. *American Sociological Review*, *75*(1), 52–74.

Hallinger, P. (1992). The evolving role of American principals: From managerial to instructional to transformational leaders. *Journal of Educational Administration*, *30*(3), 35–48.

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.

Hallinger, P., & Murphy, J. (1987). Assessing and developing principal instructional leadership. *Educational Leadership*, *45*(1), 54–61.

Halverson, R. (2010). School formative feedback systems. *Peabody Journal of Education*, *85*(2), 130–146.

Halverson, R., Grigg, J., Prichett, R., & Thomas, C. (2007). The new instructional leadership: Creating data-driven instructional systems in school. *Journal of School Leadership*, *17*(2), 159–193.

Halverson, R., Kelley, C., & Shaw, J. (2014). A CALL for improved school leadership. *Phi Delta Kappan*, *95*(6), 57–60.

Harris, A. (2008). Distributed leadership through the looking glass. *Journal of Educational Administration*, *46*(2), 172–188.

Heck, R., & Hallinger, P. (1999). Conceptual models, methodology, and methods for studying school leadership. *The 2nd handbook of research in educational administration.* San Francisco, CA: McCutchan.

Henderson, A. T., & Mapp, K. L. (2002). *A new wave of evidence: The impact of school, family, and community connections on student achievement. Annual synthesis 2002*. National Center for Family and Community Connections with Schools.

Hilton, M. E. (1989). A comparison of a prospective diary and two summary recall techniques for recording alcohol consumption. *British Journal of Addiction, 84*(9), 1085–1092.

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253–263.

Jean-Marie, G., Normore, A. H., & Brooks, J. S. (2009). Leadership for social justice: Preparing 21st century school leaders for a new social order. *Journal of Research on Leadership Education, 4*(1), 1–31.

Kelley, C., Halverson, R., Camburn, E., Blitz, M., Salisbury, J., & Dikkers, S. (2012). The design and validation of the Comprehensive Assessment of Leadership for Learning (CALL) formative school leader assessment. In American Educational Research Association Conference, Vancouver, British Columbia, April (pp. 13–17).

Kelley, C., & Salisbury, J. (2013). Defining and activating the role of department chair as instructional leader. *Journal of School Leadership, 23*(2), 287–323.

Klar, H. W. (2013). Principals fostering the instructional leadership capacities of department chairs: A strategy for urban high school reform. *Journal of School Leadership, 23*(2), 324–361.

Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics, 121*(4), 1133–1165.

Lambert, L. (2002). A framework for shared leadership. *Educational Leadership, 59*(8), 37–40.

Leithwood, K., & Seashore-Louis, K. (2011). *Linking leadership to student learning*. San Francisco, CA: Jossey-Bass.

Levine, R. E., Chambers, J. G., Ixtlac, E., & Hikido, C. S. (1998). *Improving the measurement of staffing resources at the school level: The development of recommendations for NCES for the Schools and Staffing Surveys (SASS)*. Washington, D.C.: American Institutes for Research.

Lurie, N. H., & Mason, C. H. (2007). Visual representation: Implications for decision making. *Journal of Marketing, 71*(1), 160–177.

Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational Administration Quarterly, 39*(3), 370–397.

Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works: From research to results*. Alexandria, VA: Association for Supervision and Curriculum Development.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, *50*(4), 370–396.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, *21*(1), 29–45.

Mediratta, K., & Fruchter, N. (2001). *Mapping the field of organizing for school improvement: A report on education organizing in Baltimore, Chicago, Los Angeles, the Mississippi Delta, New York City, Philadelphia, San Francisco, and Washington D.C.* New York: Institute for Education and Social Policy.

Murphy, J., Elliott, S. N., Goldring, E., & Porter, A. C. (2006). *Learning-centered leadership: A conceptual foundation*. Learning Sciences Institute, Vanderbilt University.

Newmann, F. M., King, M. B., & Youngs, P. (2000). Professional development that addresses school capacity: Lessons from urban elementary schools. *American Journal of Education*, *108*(4), 259–299.

Oakes, J. (1986). *Educational indicators: A guide for policymakers*. New Brunswick, NJ: Center for Policy Research in Education, Rutgers University

Ogawa, R. T., & Collom, E. (2000). Using performance indicators to hold schools accountable: Implicit assumptions and inherent tensions. *Peabody Journal of Education*, *75*(4), 200–215.

Pearce, C. L., & Conger, J. A. (2002). *Shared leadership: Reframing the hows and whys of leadership*. Thousand Oaks, CA: Sage Publications.

Polikoff, M. S., May, H., Porter, A. C., Elliott, S. N., Goldring, E., & Murphy, J. (2010). An examination of differential item functioning in the Vanderbilt Assessment of Leadership in Education. *Journal of School Leadership*, *19*(6), 661–679.

Porter, A. C., Polikoff, M. S., Goldring, E. B., Murphy, J., Elliott, S. N., & May, H. (2010). Investigating the validity and reliability of the Vanderbilt Assessment of Leadership in Education. *The Elementary School Journal*, *111*(2), 282–313.

Printy, S. M., & Marks, H. M. (2006). Shared leadership for teacher and student learning. *Theory into Practice*, *45*(2), 125–132.

Robinson, V. M., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly*, *44*(5), 635–674.

Rubin, D. C., & Baddeley, A. D. (1989). Telescoping is not time compression: A model. *Memory & Cognition, 17*(6), 653–661.

Selden, R.W. (1994). How indicators have been used in the USA. In K. A. Riley & D. L. Nutta; (Eds.), *Measuring quality: Education indicators–United Kingdom and international perspectives* (pp. 41–48). Bristol, PA: Farmer.

Senge, P. M. (2006). *The fifth discipline: The art and practice of the learning organization*. New York, NY: Broadway Business.

Shields, C. M. (2004). Dialogic leadership for social justice: Overcoming pathologies of silence. *Educational Administration Quarterly, 40*(1), 109–132.

Spillane, J. P. (2005, June). Distributed leadership. *The Educational Forum, 69*(2), 143–150.

Spillane, J. P., & Burch, P. (2006). The institutional environment and instructional practice: Changing patterns of guidance and control in public education. In H. D. Meyer & B. Rowan (Eds.), *The new institutionalism in education* (87–102). Albany, NY: State University of New York Press.

Spillane, J. P., & Diamond, J. B. (2007). Taking a distributed perspective. In J. P. Spillane & J. B. Diamond (Eds.), *Distributed leadership in practice* (pp. 1–15). New York, NY: Teachers College Press.

Spillane, J., Diamond, J., Hallett, T., Halverson, R., & Burch, P. (2002). Managing in the middle: School leaders and the enactment of accountability policy. *Educational Policy, 16*(5), 731–762.

Spillane, J. P., Halverson, R., & Diamond, J. B. (2001). Investigating school leadership practice: A distributed perspective. *Educational Researcher, 30*(3), 23–28.

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.

Spillane, J. P., Reiser, B. J., & Reimer, T. (2002). Policy implementation and cognition: Reframing and refocusing implementation research. *Review of Educational Research, 72*(3), 387–431.

Theoharis, G. (2007). Social justice educational leaders and resistance: Toward a theory of social justice leadership. *Educational Administration Quarterly, 43*(2), 221–258.