VANDERBILT ASSESSMENT *of* LEADERSHIP *in* EDUCATION™

Technical Manual

Version 1.0

Andrew C. Porter, Joseph Murphy, Ellen Goldring, Stephen N. Elliott, Morgan S. Polikoff & Henry May

 https://doi.org/10.59656/EL-SB8588.001

Preface

Principal leadership is an essential element of successful schools. To date, much of the work on developing educational leadership for school improvement has focused on licensure, program accreditation, and professional development, including coaching and mentoring. The identification and development of effective leadership, however, has been significantly hampered by the paucity of technically sound tools for assessing and monitoring the performance of school leaders. Until the publication of the VAL-ED, there have been few school leadership assessment instruments that have undergone scientific, psychometric development.

With initial funding from the Wallace Foundation, a study team was assembled to address this problem. Members of the study team included both experts in the content of educational leadership and experts in testing and measurement. Andy Porter is a statistician and psychometrician whose research agenda has focused on assessment and accountability, the content and alignment of instruction, and the effects of curriculum policies. Joseph F. Murphy is a former school, district, and state administrator; he has written extensively on leadership and school improvement in over a dozen books and over two hundred book chapters and journal articles. Ellen Goldring has studied the roles of school leaders in changing organizational contexts. She has specifically focused on expertise in school leadership, new models for professional development for school leaders, and linking leading and learning. Stephen N. Elliott has studied assessment of children's social skills and academic competence, serving on the U.S. Department of Education's technical advisory panels for the National Assessment of Educational Progress and the National Alternate Assessment Study. He has also authored four widely used behavior rating scales. Morgan Polikoff is a graduate student in the Education Policy program at the University of

Pennsylvania and Henry may is a research assistant professor with specialties in applied statistics, psychometrics, and education policy, also with the University of Pennsylvania.

As research on the VAL-ED is ongoing, this Technical Manual is a "living document" that will be updated periodically with new information.

Acknowledgement

The authors acknowledge the encouragement and financial support in the form of a grant from the Wallace Foundation. Without this support, the VAL-ED would not exist. We also have benefited from the insights and assistance of a number of pre- and post-doctoral students: Xiu Cravens, Jennifer Frank, and Timothy Zeidner, in the design of the instrument. We thank James O'Toole for his skillful and persistent efforts to recruit a national sample of principals, their teachers, and supervisors to provide us the normative basis for interpreting VAL-ED scores. Finally, we thank members of our expert practitioner panel with whom we met twice and from whom we learned a great deal that improved both our instrument and reporting of results: Andy Cole, Steven Daeschner, Ann Duffy, Sandra Stein, Get Nichols, Veta Daley, Rose Fairweather-Clunie, Diana Larbi; and members of our researcher panel with whom we met once and corresponded with over the course of the development work: Linda Darling-Hammond, Ken Leithwood, Jim Spillane, Sue Bodilly, Rod Ogawa, Lynn Scott, Mike Knapp, Diana Pounder, Robert Linn.

Table of Contents

Chapter 1:
Introduction

The Vanderbilt Assessment of Leadership in Education (VAL-ED) is an evidenced-based, multi-rater rating scale that assesses principals' learning-centered leadership behaviors known to directly influence teachers' performance, and in turn students' learning. The VAL-ED measures critical learning-centered leadership behaviors for the purposes of diagnostic analyses, performance feedback, progress monitoring, and professional development planning.

Work began in 2006 to develop a 360 degree instrument to assess the effectiveness of school leaders as evaluated by teachers, supervisors, and principals themselves. The resulting assessment is available in paper or on-line and utilizes a multi-rater, evidence-based approach to measure the effectiveness of leadership behaviors. There are two parallel forms of the instrument to facilitate repeated assessments over time. The VAL-ED measures *core components* and *key processes*. Core components refer to characteristics of schools that support the learning of students and enhance the ability of teachers to teach. Key processes refer to how leaders create and manage those core components. Effective learning-centered leadership is at the intersection of the two dimensions: core components created through key processes. The outcomes of the assessment are profiles, interpretable from both norm-referenced and standards-referenced perspectives, and suggested clusters of behaviors for improvement.

A series of psychometric studies has established that when used as designed the VAL-ED (a) works well in a variety of settings and circumstances, (b) is unbiased, (c) is construct valid, (d) is reliable, (e) is feasible for widespread use (both on-line and paper-and-pencil versions), (f) provides accurate and useful reporting of results, (g) yields a diagnostic profile for formative purposes, and (h) can be used to measure progress over time in the development of leadership. In

the remainder of this manual, we provide detailed documentation about these developmental and initial validation studies for the VAL-ED. Because the validation of all tests and assessments is an ongoing process designed to refine technical features and understand long-term consequential aspects of the use of assessment, we encourage researchers and practitioners alike to consult our website (www.valed.com) for periodic updates about the VAL-ED. Persons interested in the administration and use of the VAL-ED are encouraged to consult a companion VAL-ED Users' Guide.

Chapter 2
The VAL-ED Conceptual Framework and Design Blueprint

The State of Leadership Assessment and the Rationale for Development of the VAL-ED

The core challenge facing America's schools, especially urban schools, is improving student achievement and decreasing the achievement gap. Research suggests that schools that cultivate particular in-school processes and conditions such as rigorous academic standards, high-quality instruction, and a culture of collective responsibility for students' academic success are best able to meet the needs of all students (Bryk & Driscoll, 1985; Newmann & Wehlage, 1995; Purkey & Smith, 1983). Principal leadership is widely recognized as important in promoting these in-school processes and conditions (Lieberman, Falk, & Alexander, 1994; Louis, Marks, & Kruse, 1996; Rosenholtz, 1989; Sheppard, 1996). Hence, meeting the excellence and equity challenges in urban schools depends on school leaders who effectively guide instructional improvement (Leithwood, 1994; Leithwood et al., 2004).

Finding practical ways to validly assess leaders can have an important impact on the quality of leadership and, through that, on the quality of education in our schools (Glasman & Heck, 1992; Thomas, Holdaway, & Ward, 2000). Leadership evaluation holds significant promise in providing educators with much-needed information that can be used for both improving leadership practices and for accountability purposes (Reeves, 2005; Waters & Grubb, 2004). There is, however, widespread criticism regarding the adequacy of leadership assessment instruments and the processes employed to evaluate school principals. As early as 1990, in a comprehensive review of the literature related to principal evaluation, Ginsberg and Berry (1990) found a wide array of practices reported with little systematic research to support one approach over another. In 1992 and 1993, the weakness of research on school leadership evaluation was the topic of two full issues of

The Peabody Journal of Education, in which Ginsberg and Thompson (1992) lamented "the state of research on principal evaluation emphasizes the lack of empirically supported information about best practices" (p.67). A nationwide survey by Reeves (2005) found that principals agreed that their evaluations were generally positive, accurate, and consistent with job expectations. However, few found the evaluation process relevant to enhancing their motivation and improving their performance.

There is also great variation in what is assessed. Assessments typically focus on job tasks or lists of responsibilities (Ginsberg & Thompson, 1992) and characteristics of the school leader, including traits, dispositions, and attributions. A recent comprehensive review concluded that the content of principal assessments in the field is weakly related to leadership behaviors that are associated with student learning (Goldring et al, 2007). The psychometric properties of the instruments are almost never reported. Only four assessments of 66 identified describe any psychometric properties. In addition, most instruments contain no information about how standards were set and none of the instruments provide norms for comparison purposes. The review concluded that the *Personnel Evaluation Standards* developed by the Joint Committee on Standards for Education Evaluation (1988) are not adhered to in the field of school leadership assessment. Recently, Portin and colleagues (2006) concluded "as valid and reliable assessments of leaders' work, these devices generally fall far short of accepted standards in the measurement field. … What is more, the assessments tend to be poorly aligned, if at all with priorities for educational practice and improvement. . . ."(pg. 2).

It is against this backdrop that, with funding from the Wallace Foundation, we began a three-year project to develop a set of instruments to assess the effectiveness of principal leadership. The focus is on the assessment of leadership behaviors. Our conception is aligned with a research-based

definition of educational leadership that is rooted in school improvement. We call this "learning-centered leadership" (Murphy et. al., 2006). The touchstones for this strand of leadership include the ability of leaders to (a) stay consistently focused on the core technology of schooling (learning, teaching, curriculum, and assessment) and (b) make all the other dimensions of schooling (e.g., administration, organization, finance) work in the service of a more robust core technology and improved student learning.

The Vanderbilt Assessment of Leadership in Education (VAL-ED) is markedly different from current leadership evaluation and assessment frameworks employed by states and districts throughout the United States. First, the VAL-ED is aligned to the Interstate School Leader Licensure Consortium (ISLLC) (2008) standards. Second, the VAL-ED uses 360 degree feedback, from teachers, principals, and supervisors. Third, the content of the assessment is *learning-centered* leadership behaviors, behaviors that are related to increases in student achievement. Fourth, the assessment is of leadership *behaviors*, not knowledge, dispositions, or personal characteristics of leaders. Fifth, the VAL-ED requires respondents to identify evidence on which they base their assessment of principal behaviors. Sixth, the psychometric properties are clearly documented. Information on norms, standards, and uses is available in this technical manual and a Users' Guide (Elliott, et. al., 2008). In short, the VAL-ED is conceptually and theoretically grounded and the resulting scores are reliable and valid for purposes of evaluating learning-focused school leadership (Goldring, et. al., in press).

Learning-Centered Leadership Framework: The Blueprint for VAL-ED

Our leadership assessment instrument is part of a comprehensive model of a leadership assessment system that captures in broad strokes how education leadership should be assessed. The model (see Figure 2.1) shows that leadership knowledge and skills, personal characteristics, and values and beliefs inform the actual leadership behaviors exhibited by individuals or teams in performing their leadership responsibilities. These leadership behaviors (the constructs measured in our assessment instrument and reviewed in detail below) lead to school performance on core components such as providing a rigorous curriculum and high-quality instruction. These school performances, in turn, lead to student success. Student success is defined as value-added, for example, improvements in student achievement, student attendance, student graduation rates, and college enrollment.

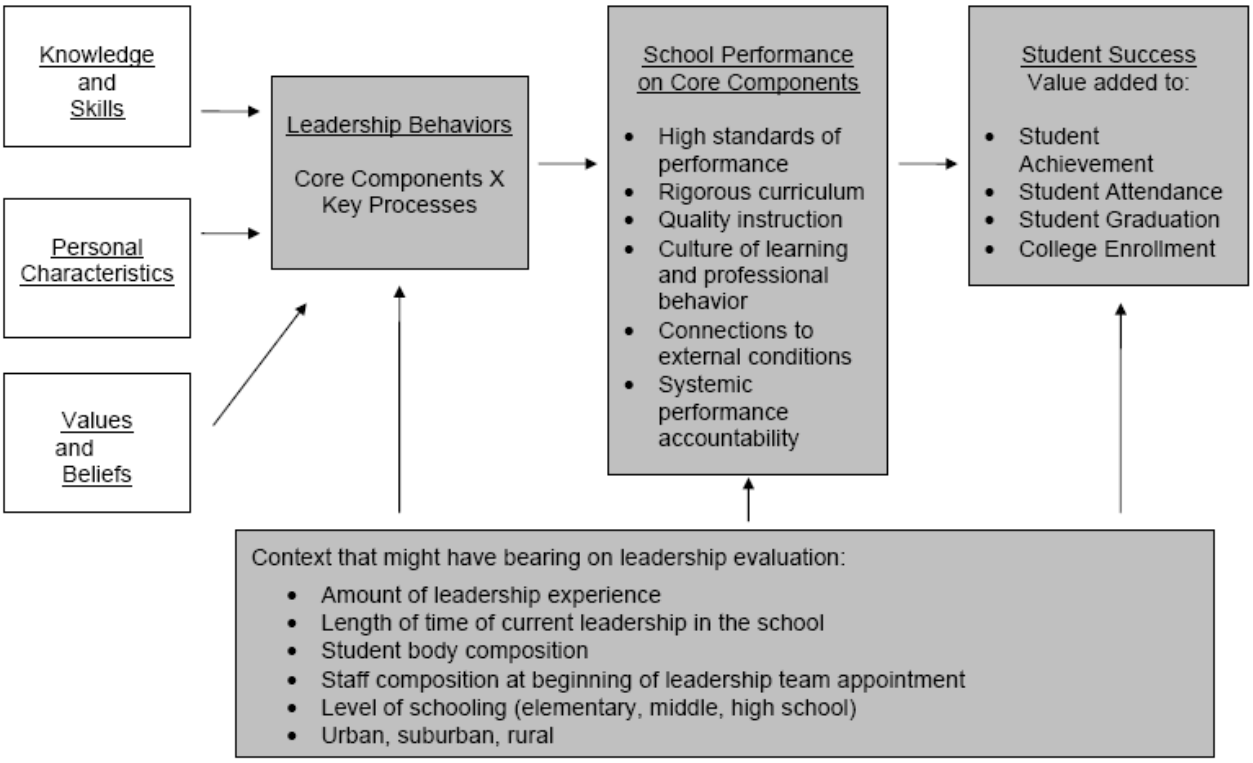


Figure 2.1. Model for Vanderbilt Assessment of Leadership in Education

Consistent with the empirical research (Hallinger & Heck, 1996; Heck & Hallinger, 1999; Leithwood et al., 2004), our assessment model does not envision direct effects of leadership behaviors on student success. Rather, the leadership behaviors lead to changes in school performance, which in turn lead to student success. Our leadership model also posits that there are aspects of the context within which leadership and schooling takes place that bear on leadership evaluation (Murphy & Meyers, 2008). Levels of experience, student body composition, staff composition, level of schooling, and geographic setting of the school can all have bearing on high-quality education leadership. The components of our leadership assessment system are highlighted in grey in Figure 2.1.

Inside this model, our proposed assessment instrument of principals' leadership behaviors is defined by the intersection of six core components of school performance and six key processes which together make up our conception of principal leadership (See Figure 2.2 below).

Key Processes						
Core Components	**Planning**	**Implementing**	**Supporting**	**Advocating**	**Communicating**	**Monitoring**
High Standards for Student Learning						
Rigorous Curriculum (content)						
Quality Instruction (pedagogy)						
Culture of Learning & Professional Behavior						
Connections to External Communities						
Performance Accountability						

Figure 2.2. The VAL-ED Constructs of Core Components and Key Processes

The framework states that school leadership assessments should measure the intersection of core components and key processes. Does the leadership in the school support teachers to develop a culture of learning and professional behavior? Does the leadership implement programs to ensure there is a culture of learning and professional behavior? Does the leadership communicate effectively about the culture of learning?

The VAL-ED assesses the intersection of *what* principals must accomplish to improve academic and social learning for all students (the core components), and *how* they create those core components (the key processes). A substantial research base supports the constructs of the core components and key processes (See Knapp et al, 2003; Leithwood et al., 2004; Murphy et al., 2007; Goldring et al, 2007 for recent reviews). Core components refer to characteristics of schools that

support the learning of students and enhance the ability of teachers to teach (Marks & Printy, 2003; Sebring & Bryk, 2000). Key processes are leadership behaviors, most notably aspects of transformational leadership associated with processes of leadership that raise organizational members' levels of commitment and shape organizational culture (Burns, 1978; Conley & Goldman, 1994; Leithwood, 1994).

Core Components

The six core components that represent the constructs of effective learning-centered instructional school leadership as grounded in the literature are:

High Standards for Student Learning. We defined high standards for student learning as the extent to which leadership ensures there are individual, team, and school goals for rigorous student academic and social learning. There is considerable evidence that a key function of effective school leadership concerns shaping the purpose of the school and articulating the school's mission (Hallinger & Heck, 2002; Knapp et al., 2003; Murphy et al., 2007). In our framework, we do not assess the mere presence of goals for student learning, but specifically emphasize the quality of the school goals, namely the extent to which there are high standards and rigorous learning goals. The research literature over the last quarter century has consistently supported the notion that having high expectations for all, including clear and public standards, is one key to closing the achievement gap between advantaged and less advantaged students, and for raising the overall academic achievement of all students (Betts & Grogger, 2003; Brookover & Lezotte, 1977; Newmann, 1997; Purkey & Smith, 1983).

Rigorous Curriculum. We define a rigorous curriculum as the content of instruction, as opposed to the pedagogy of instruction, which is dealt with in the following section. Rigorous curriculum is defined as ambitious academic content provided to all students in core academic

subjects. School leaders play a crucial role in setting high standards for student performance in their schools. These high standards, however, must be translated into ambitious academic content represented in the curriculum students experience. Murphy and colleagues (2007) argue that school leaders in productive schools are knowledgeable about and deeply involved in the school's curricular program. These leaders work with colleagues to ensure that the school is defined by a rigorous curriculum program in general and that each student's program, in particular, is of high quality (Newmann, 1997; Ogden & Germinario, 1995). Learning-centered leaders ensure that each student has an adequate opportunity to learn rigorous content in all academic subjects (Boyer, 1983).

Quality Instruction. A rigorous curriculum (i.e., ambitious academic content) is insufficient to ensure substantial gains in student learning; quality instruction (i.e., effective pedagogy) is also required (Leithwood et al., 2004). Quality instruction is defined as effective instructional practices that maximize student academic and social learning. This component reflects research findings over the course of the past few decades about how people learn (National Research Council, 1999). That work makes clear that teachers' pedagogical practices must draw out and work with the pre-existing understanding that students bring to the classroom. Effective instructional leaders understand the properties of quality instruction and find ways to ensure that quality instruction is experienced by all students in their schools. They spend time on the instructional program, often through providing feedback to teachers and supporting teachers to improve their instruction (Wellisch et al., 1978; Marzano et al., 2005).

Culture of Learning and Professional Behavior. Another core component is leadership that ensures there are integrated communities of professional practice in the service of student academic and social learning—that is, a healthy school environment in which student learning is the central focus. Research has demonstrated that schools organized as communities, rather than bureaucracies,

are more likely to exhibit academic success (Bryk & Driscoll, 1988; Lee, Smith, & Croninger, 1995; Louis & Miles, 1990). Further, research supports the notion that effective professional communities are deeply rooted in the academic and social learning goals of the schools (Little, 1982; Rosenholtz, 1989). Often termed teacher professional communities, these collaborative cultures are defined by elements such as shared goals and values, focus on student learning, shared work, deprivatized practice, and reflective dialogue (Louis, Marks, & Kruse, 1996). School leadership plays a central role in the extent to which a school exhibits a culture of learning and professional behavior and includes integrated professional communities (Bryk, Camburn, and Louis, 1999; Louis, Marks, and Kruse, 1996)

Connections to External Communities. Leading a school with high expectations and academic achievement for all students requires robust connections to the external community. There is a substantial research base that has reported positive relationships between family involvement and social and academic benefits for students (Henderson & Mapp, 2002). A study of standards-based reform practices, for instance, found that teacher outreach to parents of low-performing students was related to improved student achievement (Westat and Policy Studies Associates, 2001). Similarly, schools with well-defined parent partnership programs show achievement gains over schools with less robust partnerships (Shaver & Walls, 1998). Learning-centered leaders play a key role in both establishing and supporting parental involvement and community partnerships.

Performance Accountability. There is individual and collective responsibility among the leadership, faculty, students, and the community for achieving the rigorous student academic and social learning goals. Accountability stems from both external and internal accountability systems (Adams & Kirst, 1999). External accountability refers to performance expectations that emerge from outside the school and the local community. Simultaneously, schools and districts have internal

accountability systems with local expectations and individual responsibilities. Internal goals comprise the practical steps that schools must take to reach their targets. Schools with higher levels of internal accountability are more successful within external accountability systems, and they are more skillful in areas such as making curricular decisions, addressing instructional issues, and responding to various performance measures (Bryk & Schneider, 2002; Elmore, 2005). Learning-centered leaders integrate internal and external accountability systems by holding their staff accountable for implementing strategies that align teaching and learning with achievement goals and targets set by policy.

Key Processes

Our conceptual framework features six key process constructs. Following a systems view of organizations, we acknowledge that the processes are interconnected, recursive, and reactive to one another, but for purposes of our assessment and descriptive analysis we review each individually.

Planning. An essential process of leadership is planning. We define planning as articulating shared direction and coherent policies, practices, and procedures for realizing high standards of student performance. Planning helps leadership focus resources, tasks, and people. Learning-centered leaders do not see planning as a ritual or as overly bureaucratic. They engage in planning as a mechanism to realize the core components of the school. Effective principals are highly skilled planners and in fact, they are proactive in their planning work (Leithwood & Montgomery, 1982). Planning is needed in each of the core components; it is an engine of school improvement that builds common purpose and shared culture (Goldring & Hausman, 2001; Teddlie, Stringfield, Wimpleberg, & Kirby, 1989).

Implementing. After planning, leaders implement; for example, they put into practice the activities necessary to realize high standards for student performance. In a comprehensive review of

the research on implementation of curriculum and instruction, Fullan and Pomfret (1977) concluded that "implementation is not simply an extension of planning… it is a phenomenon in its own right" (p. 336). Effective leaders take the initiative to implement and are proactive in pursuing their school goals (Manasse, 1985). Learning-centered leaders are directly involved in implementing policies and practices that further the core components in their schools (Knapp et al., 2003). For example, effective leaders implement joint planning time for teachers and other structures as mechanisms to develop a culture of learning and professional behavior (Murphy, 2005). Similarly, they implement programs that build productive parent and community relations as a way to achieve connections to external communities (Leithwood & Jantzi, 2005).

Supporting. Leaders create enabling conditions; they secure and use the financial, political, technological, and human resources necessary to promote academic and social learning. Supporting is a key process ensuring that the resources necessary to achieve the core components are available and used well. This notion is closely related to the transformational leadership behaviors associated with helping people be successful (Leithwood & Jantzi, 2005). The literature is clear that learning-centered leaders devote considerable time to supporting teachers, for example, in their efforts to strengthen the quality of instruction (Conley, 1991; Leithwood & Jantzi, 1990). This support takes varied forms. Leaders demonstrate personal interest in staff and make themselves available to them (Marzano et al., 2005). Leaders also provide support for high-quality instruction by ensuring that teachers have guidance as they work to integrate skills learned during professional development into their instructional behaviors (Murphy et al., 2007).

Advocating. Leaders promote the diverse needs of students within and beyond the school. Advocating for the best interests and needs of all children is a key process of learning-centered leadership (Murphy et al., 2007). Learning-centered leaders advocate for a rigorous instructional

program for all students. They ensure that policies in the school do not prevent or create barriers for certain students to participate in classes that are deemed gateways to further learning, such as algebra. They ensure that special needs students receive content-rich instruction. Similarly, effective leadership ensures that all students are exposed to high-quality instruction; they manage the parental pressures that often create favoritism in placing students in particular classes. Both the instruction and content of the school's educational programs honor diversity (Ogden & Germinario, 1995; Roueche & Baker, 1986). Through advocacy, learning-centered leadership works with teachers and other professional staff to ensure that the school's culture both models and supports respect for diversity. (Butty, LaPoint, Thomas, & Thompson, 2001; Goldring & Hausman, 2001).

Communicating. Leaders develop, utilize, and maintain systems of exchange among members of the school and with the school's external communities. In studying school change, Loucks and colleagues (1982) found that "principals played major communication roles, both with and among school staff, and with others in the district and in the community" (p. 42). Learning-centered leaders communicate unambiguously to all the stakeholders and constituencies both in and outside the school about the high standards of student performance (Leithwood & Montgomery, 1982; Knapp et al., 2003). Leaders communicate regularly and through multiple channels with families and community members, including businesses, social service agencies, and faith-based organizations (Edmonds & Frederiksen, 1978; Garibaldi, 1993; Marzano et al., 2005). Through ongoing communication, schools and the community serve as resources for one another that inform, promote, and link key institutions in support of student academic and social learning.

Monitoring. Monitoring is defined by leaders systematically collecting and analyzing data to make judgments that guide decisions and actions for continuous improvement. Early on, the effective schools literature identified monitoring school progress in terms of setting goals, assessing

the curriculum, and evaluating instruction as a key role of instructional leadership (Hallinger & Murphy, 1985; Purkey & Smith, 1983). Learning-centered leaders monitor the school's curriculum, assuring alignment between rigorous academic standards and curriculum coverage (Eubanks & Levine, 1983). They monitor students' programs of study to ensure that all students have adequate opportunity to learn rigorous content in all academic subjects (Boyer, 1983; Hallinger & Murphy, 1985). Learning-centered leadership also undertakes an array of activities to monitor the quality of instruction, such as ongoing classroom observations (Heck, 1992). Monitoring student achievement is central to maintaining systemic performance accountability.

Development of a Technically Sound Assessment of Leadership

To measure principals' behaviors at the intersection of core components and key processes, we developed the VAL-ED, a multi-respondent (principal, teacher, and supervisor) rating scale that requires respondents to make judgments about a principal's leadership behaviors that influence teachers' performance and students' learning. Our six-by-six, 36-cell conceptual model of leadership provides the framework for writing items that describe leaders' behaviors represented by the cell. Each cluster of items in each cell serves as indicators of our construct of leadership (see Figure 2.2).

The resulting VAL-ED 360 assessment consists of 72 items on each of two forms, A and C. Items were randomly assigned to a form within each of the 36 cells. For each respondent group (principal, supervisor, teachers), the forms are virtually identical. The assessment can be completed in paper and pencil format or online. For each of the 72 items, the respondent rates the effectiveness of the principal's behavior in "ensuring the school…," for example, "plans rigorous growth targets in learning for all students." The effectiveness scale has five options from outstandingly effective (5) to ineffective (1). Before rating the effectiveness of each of the 72

principal behavior items, the respondent is to check the sources of evidence on which the effectiveness rating is to be based. There are five options for sources of evidence: reports from others, personal observations, school documents, school projects or activities, other sources, or no evidence (See Appendix A for a copy of the instrument).

Principals' effectiveness is reported in terms of a total score across all 72 items (mean item response), a score on each of the core component subscales, and a score on each of the key process subscales. Since the core components and key processes subscales are based on the same set of 72 items, they are based on the same information organized in two different ways. Total score and each of the twelve sub-scores are reported once for each respondent group and once aggregated across respondent groups where each respondent group is weighted equally (i.e. an average is first formed across teachers and then the average is taken across the three respondent groups). Results are reported in the effectiveness scale metric. In addition, percentile ranks are available as well as proficiency standards of distinguished, proficient, basic, and below basic. See Appendix B for a sample VAL-ED principal report.

The design of the VAL-ED is directly influenced by technical standards for high-quality assessments (*Standards for Educational and Psychological Testing*, AERA, APA, & NCME, 1999, Personnel Evaluation Standards Joint Committee for Education Evaluation, 1988), and time-tested practices of item and test development (Haladyna, Downing, & Rodriguez, 2002). Collectively, these professional documents and the published research on test development provide strong guidelines for designing a high-quality and successful assessment program for school leaders.

In the next chapter, we describe the development of the VAL-ED. We focus on the content and construct validity of the instrument, and how validity was supported through a number of qualitative and quantitative studies of the instrument. In chapter 4, we present results from a

national field trial of the VAL-ED, designed to provide data for calculating initial norms, setting performance standards, and determining the validity and reliability of the VAL-ED.

Chapter 3:
The Development of the VAL-ED

The development of a technically sound assessment is an ongoing process that begins with the conceptualization of the instrument and continues well after the instrument is published. The development of the VAL-ED was guided by a comprehensive plan that involved: (1) specifying the purposes of the assessment, (2) defining content assessed, (3) writing items, (4) designing instructions and response format, (5) piloting test forms, (6) designing scoring and interpretation frameworks, (7) conducting studies that yield evidence for the reliability and validity of the scores, (8) refining items, format, and score interpretation procedures, (9) field-testing forms with a representative sample, (10) developing norms and standards to guide interpretation of results, and (11) writing a technical manual that summarizes technical characteristics and sound uses of the assessment. This chapter describes the instrument development phase from conception through finalization of forms. Chapter 4 describes the national field trial and resulting psychometric evidence for the validity and reliability of the VAL-ED.

Instrument development

The first phase of test development began with a thorough examination of the research literature and creation of the conceptual framework. From the 36-cell conceptual framework pictured in Figure 2.2, the process of item writing began. For each cell in the framework, one of the test's authors first wrote a set of leadership behaviors intended to be exhaustive. Another researcher examined several extant principal leadership evaluations to cull additional items that fit into specific cells in the framework. From the first comprehensive list of items, both original and assembled from

other instruments, item editing continued with the goal of developing a census of all important leadership behaviors in each cell.

Items were then examined by the full team for redundancy within cell, within core component, and within key process. Where necessary, items were moved to more appropriate cells. Items were evaluated for their level of detail, so that items that were too global (not anchored in specific behaviors) or too specific were removed from the list. Next, the list of items was examined by the research team to identify any important missing items, which were added to the list. An appropriate set of verbs was defined for items in each key process (e.g., for advocating: advocates, represents, challenges, promotes), and each item was modified to include an appropriate verb.

Next, items in each core component were assigned to one research team member for extended scrutiny. The items were evaluated for the explicitness of the link to the core component—for those not linked closely enough, they were modified to fit more closely or deleted if modification was impossible. Also, all team members read and evaluated each item and rated each item on a 3 point scale: 1) unique and important, 2) unique and marginally important, and 3) redundant with some other item. At team meetings, every item that did not score all 1's was discussed by the team and improved or removed. Finally, a list of 294 items was subjected to an inspection within core components and within key processes for redundancy.

The item-writing process took place over a span of seven months and produced what we believe to be an item set with several important characteristics. First, every item is at an appropriate grain-size—neither too broad nor too narrow. Second, the items are intended as a census of the possible items that fit into the two-dimensional framework, with no redundancies. Third, every item fits clearly into a specific cell based on the definitions of the core component and key process that correspond to that cell.

With the item writing done and other key decisions made, the first complete draft of the instrument and items was ready for examination. Parallel forms A and C were constructed so schools can use the instrument in consecutive years without seeing the same items twice. The goal is to focus attention on the domains of behavior represented by each of the 36 cells in the conceptual framework, not the specific sample of behaviors in one form of the instrument. There were enough items in each cell to allow for random sampling of items from each cell domain to create the forms, initially three items per cell but ultimately two per cell.

The item construction and test development phase is the most important step in establishing instrument validity. Again, items were specifically written for each of the 36 cells in the conceptual framework. Items were repeatedly revised by researchers and corrected for grain size, redundancy, clarity, and cell fit. If no additional validation work had been done, at this point the items and directions would have comprised a content-valid instrument.

Sorting Study (Study 1)

A sorting study served as a first step in empirically testing the validity of our measure. The study sought to establish whether the items within the instrument measure the domains that they were constructed to measure. The guiding question was to determine whether school principals could accurately place items into the 36 cells defined by the intersection of the six core components and six key processes. Nine principals were recruited to the task. Each was provided with the definitions of each core component and each key process and the 36 cell matrix in Figure 2.2. To make the task manageable, the pool of 294 items was divided into three random sets stratified by cell. Each set of 98 items was independently sorted by three principals. Items were presented in a random order with no identification as to core component or key process. Principals completed the task off site and on their own timeline.

Results

Eighty-six percent of the classifications into cells of the 294 items resulted in the correct cell identified by at least one of the three principals assigned the item. Fifty-nine percent of the classifications of items were in the exact correct cell by two of three principals assigned the item. Placement in the correct cell is a demanding criterion. When the criterion for classification was relaxed to ask whether the principal identified the item's correct core component, 75% of the placements were correct. For key processes, 76% of the placements were correct.

Table 3.1 provides detailed results on the percent of accurate classifications at the cell, core component, and key process levels for the items in each of the 36 cells in the conceptual framework. Results reveal that some core components and some key processes were easier to classify accurately than were others. High percent accurate classifications were found for some specific cells: Advocating High Standards, Planning and Communicating Rigorous Curriculum, Monitoring Quality Instruction, Communicating Culture of Learning and Professional Behavior, and Supporting Connections to External Communities. Each of these five cells had 70% or greater correct classification for the items in that cell. At the other extreme, Implementing Quality Instruction (37%) and Implementing Performance Accountability (36%) were more difficult combinations of key processes by core components to classify. Comparing the first entry in each cell (i.e. percent of accurate classification at the cell level) to each of the other two entries in the cell identifies whether it was primarily the core component or the key process that created a difficulty in accurate classification. For example, in Implementing Rigorous Curriculum there was 46% accurate classification at the cell level, 92% accurate classification for the core component and only 46% accurate classification for the key process. Clearly, it was the key process of Implementing that principals had difficulty detecting.

Table 3.1 : Principals' classification of items in the conceptual framework

Key Processes							
Core Components	**Planning**	**Implementing**	**Supporting**	**Advocating**	**Communicating**	**Monitoring**	**Marginals for Core Components**
High Standards for Student Learning	42% 63% 67%	44% 83% 61%	63% 79% 79%	80% 87% 93%	57% 57% 100%	37% Total 52% CC 74% KP	68%
Rigorous Curriculum (content)	71% 92% 79%	46% 92% 46%	67% 67% 87%	60% 67% 93%	78% 78% 100%	89% Total 89% CC 100% KP	83%
Quality Instruction (pedagogy)	52% 57% 81%	37% 63% 57%	58% 79% 71%	47% 47% 93%	67% 94% 67%	71% Total 81% CC 86% KP	71%
Culture of Learning & Professional Behavior	53% 83% 60%	47% 80% 60%	63% 89% 74%	58% 67% 88%	76% 90% 86%	86% Total 86% CC 100% KP	82%
Connections to External Communities	67% 90% 71%	41% 82% 44%	71% 90% 76%	63% 85% 78%	50% 62% 81%	76% Total 95% CC 81% KP	81%
Performance Accountability	50% 63% 75%	36% 74% 41%	63% 83% 75%	58% 67% 75%	62% 71% 86%	64% Total 72% CC 92% KP	72%
Marginals for Key Processes	72%	51%	76%	86%	85%	88%	

*Note: "Total" indicates the percentage of items in each cell that matched our precise placement within the cell. "CC" indicates the percentage of items in the cell that matched our placement for the core component. "KP" indicates the percentage of items in the cell that matched our placement for the key process. Marginals represent the results across an entire key process or core component.

For the key processes, averaged across all core components, Planning had 72% accurate classification; Implementing, 51%; Supporting, 76%; Advocating, 86%; Communicating, 85%; and Monitoring, 88%. Similarly, for core components averaging across key processes, High Standards for Student Learning had 68%; Rigorous Curriculum, 83%; Quality Instruction, 71%; Culture of Learning and Professional Behavior, 82%; Connection to External Communities, 81%; and Performance Accountability, 72%. Overall, the results of the sorting study indicate that, at least for school principals, the behaviors captured by the 294 items were generally content valid when judged against the conceptual framework of core components by key processes against which the items were written.

Nevertheless, the items were revised in several ways as a result of the sorting study to improve the content validity of the instrument. The respondents had particular difficulty sorting Implementing items correctly, often sorting them into Planning or Supporting. To address this problem, all Planning items were edited to include the words "plan" or "planning." Additionally,

each core component was assigned to a study team member to examine items with significant sorting issues (0 or 1 respondent placed the item in the correct cell). The team member suggested appropriate changes to ensure better fit to the target cell. If no appropriate remedy could be reached, the item was deleted or reworded substantially and assigned to another cell. The full study team signed off on each item change.

Cognitive Labs (Study 2)

Next, two rounds of cognitive labs were conducted. Cognitive labs are helpful in addressing the most common threats to survey validity, including the complexity of the phenomena under question, the possibility of socially desirable responses, and the likelihood of unintentionally misleading responses (Biemer, Groves, Lyberg, Mathiowetz, & Sudman, 1991; Desimone & LeFloch, 2004). There are two stages to a cognitive lab. In the first stage, respondents are encouraged to "think aloud" as they answer questions or read directions. Here, respondents are asked to describe their thought process as it occurs, providing as much detail as possible. Whether the item is clear or ambiguous, respondents are to say whatever is on their mind. In the second stage, interviewers ask specific questions of respondents about item or response choice interpretation (for a full description of the cognitive lab methodology, see Desimone & LeFloch, 2004).

Two rounds of cognitive labs with three sets of interviews each were conducted. In the first round, the interviews were conducted in one school each in three urban districts, including a middle school, a high school, and an elementary school. In each district, there were three respondents: a principal, one of the principal's teachers, and a supervisor of a principal. Each city's respondents were paired with a form of the leadership assessment instrument.

For the first round of cognitive labs, subjects were introduced to the cognitive lab "think aloud" methodology with an example. Next, respondents were asked to read aloud and examine the

assessment's cover page and directions and comment on language, aesthetics, and clarity. They were also asked probing questions about particular phrases and words the investigators anticipated being problematic. In the next step, respondents read the assessment aloud, item-by-item, describing their thought process as they identified sources of evidence and checked effectiveness ratings. (Respondents were encouraged to complete the assessment as if they were actually evaluating a principal.) Periodically, the researcher would stop the respondent after answering a particular item to ask the respondent questions about his or her interpretation of key words and phrases. Finally, at the end of the interview, respondents were asked several overall questions about their opinions of the instrument and their beliefs about its utility in the field. After the first round of cognitive labs, the research team met and examined cognitive lab data to make improvements to instructions, formatting, and individual item wording.

The second round of cognitive labs was conducted with three respondents each in three districts—two urban and one suburban. Again, an elementary, middle, and high school was included. Here, subjects first completed the assessment on their own, making notes by items they wanted to discuss. A change in the format of the survey items was also included in the interviews. The initial format was that of 108 items in random order, three items for each cell in the framework. The alternative format organized the 108 items by core component and, within core component, key process. Respondents completed the form without interruption. When respondents were done with the survey, the researcher probed them on key words and phrases that still seemed potentially unclear after the first round of edits. This modified interview methodology was used to give the research team a better idea of whether respondents could successfully complete the instrument without additional support.

Results

Results from the cognitive labs helped provide additional evidence as to the content validity of the instrument. Respondents in the first round of cognitive labs seemed to periodically forget the stem, "The principal ensures the school…" focusing instead on whether the principal performed the behavior directly. This problem encouraged us to add the stem to each item in the second round to ensure respondent understanding. Another respondent problem in both rounds was with the term "leaders" in items such as "The principal ensures the school allocates leaders' time to support a system that holds students accountable for their learning." Some respondents thought that administrators were leaders, while one principal thought that every teacher in the school was a leader. Forceful terms such as "ensure" and "cause" often created problems for respondents across rounds, who felt that, for instance, nothing could "ensure students would meet high standards." This concern indicated a need to soften the language in order to more closely approximate the intended meaning of the item.

A particular challenge that resulted in both cognitive lab rounds was the propensity of the interviewee to defer to outcomes when determining an effectiveness rating. For example, when an item indicated that the principal ensures the school plans a rigorous curriculum, the rating was given on whether or not a rigorous curriculum existed, which reveals some combination of good planning and good implementation, two separate key processes. This "bleeding" of processes caused us to analyze and revise items in many instances to more fully distinguish between the key processes. Such "bleeding" of categories also occurred at times between the core component of performance accountability and the key process of monitoring crossed with other core components. Once again, potential revisions were debated and changes were implemented at the conclusion of the cognitive labs.

At the end of the cognitive labs, we asked each respondent a series of questions about the feasibility and validity of the instrument. When interviewees were asked if anything was missing from the survey, a few suggested that having data on some traditional outcomes, such as student achievement, would have helped them with their ratings. This indicated a tendency by some individuals to defer to outcomes regardless of the core component and key process the item was seeking to highlight. Overall, however, the response was that the instrument was inclusive, sometimes even redundant, and that it seemed to capture key principal leadership behaviors. In short, though the cognitive labs raised concerns about certain key words and phrases in certain items, the overall response was that the VAL-ED was measuring the key leadership behaviors that mattered to the respondents. Most importantly, the cognitive labs indicated specific ways in which the instrument could be and was improved.

Item Bias Study (Study 3)

A fairness review of the VAL-ED instructions and items was conducted to identify and remove aspects of test questions or directions that might hinder respondents from various groups from completing the instrument as intended and could lead to inappropriate inferences about their relevant knowledge and skills. The fairness review was based on the test fairness guidelines published and used by ETS (ETS, 2000):

Guideline 1. Treat people with respect in test materials.
Guideline 2. Minimize the effects of construct-irrelevant knowledge or skills.
Guideline 3. Avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting.
Guideline 4. Use appropriate terminology.
Guideline 5. Avoid stereotypes.
Guideline 6. Represent diversity in depictions of people.

The fairness review was conducted via individual electronic surveys to each panelist followed by a webex conference after all surveys were returned. Nine individuals with knowledge of

testing and rating scale methods were selected to participate on the panel. Of the nine members, six were female and three male, and all but one currently worked in public schools as either a teacher, behavior specialist, or administrator. The non-school based person worked in the testing industry as an editor. The panel self-identified themselves as four Caucasians, two Hispanics or Latinos, two African Americans, and one Asian-American. Three respondents had PhDs, two had Masters, three had a Bachelors degree, and one had a high school degree. Collectively, the panel members represented six regions of the country.

The respondents were trained on the six ETS Fairness Guidelines using a 21-slide Powerpoint show. The Powerpoint was reviewed independently by all individuals, then reviewed and discussed briefly by the group on a conference call. At the end of this training phase, all panel members reported they understood the Fairness Guidelines and felt confident they could apply them to the review of rating scale items.

Finally, panel members were asked to independently review the VAL-ED Principal's Forms A and C and circle any words or items that they believed violated a Fairness Guideline. Each reviewer was asked to note which guideline was a concern for any item or word circled. At the conclusion of the session, the set of all challenged items was identified and discussed by the group of panelists to determine if a revision could be made to resolve the fairness challenge.

Results

The panelists worked independently through both forms of the VAL-ED and recorded Fairness Guideline violations for the instrument's instructions and each item. The aggregated results of all nine panelists indicated no fairness concerns with the VAL-ED instructions or introductory content. With regard to Form A, two or more panelists identified 13 items that raised a fairness concern and possible violation. On Form C, the panelists identified 14 items that raised a fairness

concern and possible violation. From this total pool of 27 items, four items were perceived to be a serious concern for three or more panelists (see Figure 3.1). These items and the identified type of violations were discussion on a conference call with all panelists. The end result of the discussion was suggested revisions for each of these items. These revisions are documented in Figure 2.1 in bold type face. A review of the items indicates three of them concerned the leadership behavior process of advocating. The subtle, but meaningful suggested changes for these items emphasized person-first language. The authors of the VAL-ED reviewed the panelists' suggested item revisions and accepted them. After the bias study, the VAL-ED's instructions and items were seen as meeting or exceeding widely accepted fairness criteria for tests and assessments used to characterize the skills of intended respondents.

ANALYSIS OF FORM A ITEMS

Item Number	Item Content (original and **suggested revision)**	# Panelists Who Indicated Problems	Fairness Guideline Cited
8	challenges low expectations for special needs students. **challenges low expectations for students with special needs.**	7	3,4,5
56	challenges teachers to work with community agencies to support students at risk. **challenges teachers to work with community agencies to support students' needs.**	3	1,4,5
67	challenges faculty who blame others for student failure. **challenges faculty who attribute student failure to others.**	5	1,2,3,4

ANALYSIS OF FORM C ITEMS

17	supports teachers to participate in professional development that deepens their understanding of the rigorous curriculum **supports participation in professional development that deepens teachers understanding of a rigorous curriculum.**	3	1,2,4

Figure 3.1. VAL-ED Items Identified as Potentially Unfair and Suggested Revisions

Nine-School Pilot Test (Study 4)

With revisions to the instrument made, and potential concerns about bias mitigated, the next step in the validation of the instrument was a small pilot test. An urban district was recruited to participate in the pilot study in the spring of 2007. A total of nine schools were recruited, three at each level—elementary, middle, and high. Five of the schools were randomly assigned to use form A and four to use form C. Each form contained 108 items, 3 items randomly selected from each of the 36 cells in the conceptual framework with no overlap between forms.

The district was recruited through contact with the district's Wallace Foundation LEAD coordinator. All contact with schools was coordinated through the LEAD liaison. Survey forms were sent to the liaison and she sent them to each school to be completed. No instructions were given as to the setting in which the assessment was to be completed. Members of the VAL-ED research team traveled to the schools to collect the forms two to three days after the schools received the forms. Respondents were also provided with postage paid envelopes if they wished to mail back additional completed forms. In each participating school, the principal, his/her supervisor, and all teachers in the school were requested to participate. Teachers were assured of confidentiality. To encourage high response rates, a graded system of incentives was implemented. Schools received $500 for participating, but the incentive increased to $750 for 75% teacher response rate and $1000 for 90% teacher response rate.

An important issue that arose in the pilot study related to the supervisor's ratings. Only one supervisor evaluated the principals from each level of school. The elementary school supervisor rated each of his/her three principals as "highly effective" on all items, for an overall highest possible rating of 5.00. These data suggest that the supervisor did not take the exercise of rating the principals seriously. This may be due to the fact that the pilot study was not taken under "high stakes" conditions. That is, no accountability was associated with the ratings provided, so supervisors (and other respondents) may not have given the same ratings they would actually give under conditions of regular use.

Results

Feasibility. The first element of feasibility simply asks whether respondents completed the assessment. Response rates from the pilot study suggest that teachers and supervisors are willing to complete the VAL-ED. Of the nine schools, two had 100% teacher response rates and three others had greater than 90% teacher response rates. One school had between 75% and 90% response rate, and the remaining three schools had response rates of 39%, 41%, and 58%. The overall teacher response rates were 70% for form A and 75% for form C (72.5% overall). Response rates were 70% or greater for each level of school. Nine of nine supervisor forms were completed, and eight of nine principal forms were completed. A total of 319 teacher responses were collected: 153 on form A and 166 on form C.

A second element of feasibility concerns whether respondents completed individual items. There are two ways in which respondents could choose to not rate a principal: they could leave an item blank (missing data), or they could select the "Don't know" option. Principals did not have the option of selecting "Don't know," and they left 0% of items blank. Supervisors selected "Don't know" 4% of the time and left no items blank. For teachers, 1.7% of items were left blank and 6.1%

of items were marked "Don't know." As shown in Table 3.2, all twelve scales had low missing data rates. However, two core components—Connections to External Communities and Performance Accountability—and two key processes—advocating and monitoring—had higher proportions of "Don't know" ratings, with proportions greater than 10% on one or both forms. At the item level, no items had more than 6% missing data. Six items on form A and one item on form C had greater than 25% "Don't know" ratings. The majority of items on both forms had below 10% "Don't know" ratings. In short, missing and "Don't know" frequencies were not a problem at the item, scale, or form level for any respondent group.

Table 3.2: Teacher missing data and "Don't know" by scale

Teacher rating distribution by form, 9-school pilot, Spring 2007		Don't know	Missing/not entered
Form C	High standards	1.3%	1.3%
	Rigorous curriculum	3.0%	2.0%
	Quality instruction	3.4%	2.4%
	Culture of learning	2.2%	2.1%
	Connections to communities	11.5%	1.9%
	Performance Accountability	7.9%	1.7%
	Planning	3.7%	2.3%
	Implementing	4.0%	2.1%
	Supporting	2.9%	2.3%
	Advocating	6.2%	1.4%
	Communicating	4.2%	1.5%
	Monitoring	8.2%	1.8%
	Total	4.9%	1.9%
Form A	High standards	3.1%	1.7%
	Rigorous curriculum	3.7%	0.7%
	Quality instruction	4.8%	1.5%
	Culture of learning	4.7%	2.4%
	Connections to communities	15.9%	2.6%
	Performance Accountability	12.4%	2.5%
	Planning	6.0%	1.4%
	Implementing	4.3%	1.9%
	Supporting	4.4%	1.9%
	Advocating	10.1%	2.2%
	Communicating	6.7%	1.7%
	Monitoring	13.2%	2.4%
	Total	7.4%	1.9%

Analysis of the sources of evidence used in this pilot study is provided in Table 3.3. Results show that all respondent groups were most likely to indicate "Personal Observation;" roughly 70% of items had "Personal Observation" as evidence. Principals and supervisors selected more sources

of evidence than teachers, especially "School Documents." Additionally, elementary and middle school respondents marked more sources of evidence than high school respondents. The patterns of evidence fit expectations and suggest that evidence is a feasible component of the VAL-ED assessment.

Table 3.3: Sources of evidence used

Mean % of items with each kind of evidence by respondent, form, school type, 9-school pilot, Spring 2007

	Teacher	Supervisor	Principal	Form A	Form C	Elementary	Middle	High	Overall
Reports from others	28.0%	48.1%	43.5%	28.7%	28.8%	26.4%	27.2%	31.8%	**28.8%**
Personal observation	71.4%	69.7%	73.6%	70.1%	71.7%	71.4%	74.1%	67.3%	**70.9%**
School documents	45.3%	79.9%	74.5%	45.8%	47.4%	49.1%	52.8%	38.6%	**46.6%**
School projects or activities	32.7%	42.9%	42.7%	36.3%	29.9%	36.3%	34.6%	29.3%	**33.0%**
Other sources	11.4%	1.4%	28.6%	10.1%	12.7%	15.8%	11.2%	9.1%	**11.5%**
No evidence	2.7%	2.7%	0.7%	3.1%	2.2%	2.5%	2.1%	3.3%	**2.6%**
Average # of sources	**1.91**	**2.45**	**2.64**	**1.94**	**1.93**	**2.02**	**2.02**	**1.79**	**1.93**

The core of the VAL-ED assessment is the ratings given to principals. Data on the distribution of ratings also provide evidence about the feasibility of the response scale. The percentages shown in Table 3.4 reveal that ratings were high. On the initial 0 to 5 scale (later revised to 1 to 5), with 0 being "Not done" and 5 being "Highly effective," most scales had roughly 80% of teacher ratings at the 4 or 5 levels. Overall, roughly 30% of items were rated a 4 and 47% of items were rated a 5. Approximately 10% of items were rated a 3, and 3% of items were rated a 0, 1, or 2. Teacher item-level means had a roughly normal distribution, with a mean item response of approximately 4.4 on both forms. Except for one outlier item, item means ranged from 3.9 to 4.7. Teacher item standard deviations ranged from .6 to 1.3, with a mean item standard deviation of .95. The item distribution results suggest either that the principals were extremely effective or that the VAL-ED forms used in this pilot study experienced a common issue with behavior rating scales—the tendency of respondents to give high ratings overall.

Table 3.4: Teacher rating distributions by scale/total

Teacher rating distribution by form, 9-school pilot, Spring 2007		0	1	2	3	4	5
Form C	High standards	0.3%	0.7%	2.3%	10.1%	28.5%	55.5%
	Rigorous curriculum	0.2%	0.9%	2.0%	11.7%	28.1%	52.1%
	Quality instruction	0.1%	0.7%	1.8%	10.6%	28.0%	52.9%
	Culture of learning	0.4%	0.4%	2.1%	10.3%	29.5%	53.1%
	Connections to communities	0.4%	0.7%	2.9%	14.6%	27.0%	40.9%
	Performance Accountability	0.5%	0.6%	2.4%	12.2%	26.9%	47.9%
	Planning	0.3%	0.6%	2.2%	11.5%	28.9%	50.4%
	Implementing	0.1%	0.6%	2.4%	11.9%	27.8%	51.1%
	Supporting	0.2%	0.7%	2.6%	10.1%	28.6%	52.6%
	Advocating	0.4%	0.7%	2.5%	13.0%	29.2%	46.5%
	Communicating	0.1%	0.8%	1.9%	11.6%	27.4%	52.5%
	Monitoring	0.7%	0.7%	2.0%	11.2%	26.2%	49.3%
	Total	0.3%	0.7%	2.3%	11.6%	28.0%	50.4%
Form A	High standards	0.3%	0.4%	1.0%	9.5%	33.0%	50.9%
	Rigorous curriculum	0.1%	0.4%	2.0%	10.3%	34.0%	48.8%
	Quality instruction	0.5%	0.6%	2.0%	9.7%	29.1%	51.7%
	Culture of learning	0.6%	0.7%	1.8%	9.6%	27.7%	52.4%
	Connections to communities	1.5%	0.6%	2.2%	12.4%	28.3%	36.5%
	Performance Accountability	1.2%	0.8%	2.0%	11.6%	29.1%	40.5%
	Planning	0.7%	0.8%	1.7%	10.2%	32.0%	47.2%
	Implementing	0.7%	0.5%	1.9%	10.4%	31.2%	49.1%
	Supporting	0.4%	0.5%	1.5%	8.9%	29.5%	52.9%
	Advocating	1.0%	0.7%	2.1%	12.5%	29.1%	42.3%
	Communicating	0.3%	0.3%	2.1%	10.3%	30.3%	48.2%
	Monitoring	1.1%	0.7%	1.6%	10.5%	28.8%	41.8%
	Total	0.7%	0.6%	1.8%	10.5%	30.2%	46.9%

Note: Percentages going across add to 100% when missing and "Don't know" values from Table 3.2 are added. Percentages may not add to 100% due to rounding.

The fifth and final component of feasibility to be discussed is the respondents' reactions to questions about the VAL-ED's feasibility. Respondents were asked to answer six items on the final page of the assessment, with response categories of 1, strongly disagree; 2, disagree; 3, agree; and 4, strongly agree. Results appear in Table 3.5. The three most important items from a feasibility and validity standpoint are items 1, 2, and 6. Teachers and supervisors leaned toward agreement that the response form was easy to use, while principals were neutral. Teachers, principals, and supervisors also leaned toward agreement that (a) the items focused on important leadership behaviors and (b) they understood the items. All three respondent groups were neutral in their views of (a) using the

instrument every year and (b) supporting the instrument's use in their district. As for checking sources of evidence, all three groups were just above neutral. Additional space on the form was left for respondent comments. Ninety-nine respondents left comments; 81 suggested the form was too long or too repetitive. Given the complaints about time required for completion, the neutral assessment of use is surprisingly positive.

Table 3.5: Responses to feasibility questions

Responses to final questions, by respondent, 9-school pilot, Spring 2007						
	Teachers		Principals		Supervisors	
	Mean	SD	Mean	SD	Mean	SD
I found this response form easy to use.	2.82	0.77	2.50	0.53	3.00	0.00
I believe the vast majority of items focused on important leadership behaviors.	3.15	0.61	3.13	0.35	3.33	0.50
I would not object to completing this assessment of my principal every year.	2.55	0.90	2.29	0.76	2.33	0.50
I believe checking the sources of evidence for my ratings was useful.	2.73	0.75	2.63	0.52	3.00	0.00
Based on my experience today, I would support use of this assessment to evaluate school principals in my district.	2.70	0.81	2.13	0.64	2.33	0.50
I understood the vast majority of items.	3.19	0.63	3.25	0.46	3.67	0.50
	1 = strongly disagree, 4 = strongly agree					

Reliability. An important component of any assessment instrument is its reliability. There are many forms of reliability; in the pilot study, only internal-consistency reliability could be estimated. Reliabilities for both forms and all scales were high. Cronbach's Alpha reliabilities for teacher scores are presented in Table 3.6. For all twelve scales on both forms, reliabilities were near perfect. For the total score, reliabilities were greater than .98 on both forms. Reliabilities tended to be somewhat higher for core components than for key processes.

Table 3.6: Estimates of internal consistency reliability

Internal consistency for scales and total, nine school pilot, Spring 2007		
Cronbach's α	Form A	Form C
High standards for student learning	0.95	0.97
Quality instruction	0.94	0.95
Rigorous curriculum	0.95	0.97
Culture of learning	0.93	0.96
Connections to external community	0.95	0.97
Performance accountability	0.95	0.97
Planning	0.92	0.95
Implementing	0.94	0.95
Supporting	0.93	0.96
Advocating	0.94	0.96
Communicating	0.94	0.97
Monitoring	0.93	0.96
Total	0.99	0.99

As will be discussed later, after the nine-school pilot study, the decision was made to randomly delete one item from each of the 36 cells to shorten the forms because feedback indicated that the assessment was too long and reliability coefficients with a small number of items continued to be high. Thus, the number of items was reduced to 72 for each form and 12 for each scale for all remaining studies. Still, reliabilities for both forms and all scales remained nearly as high. Scale reliabilities for teacher scores were all above or near .9 and total score reliabilities still near perfect on these shortened forms.

Validity. Confirmatory factor analysis using teacher data from the nine-school pilot was done to investigate data fit to our conceptual model. The factor analytic model was designed to parallel the conceptual framework for the VAL-ED by incorporating higher-order factors for core components, key processes, and an overall score. Thus, the hierarchical factor analytic model had

four levels. The first level involved the 108 individual items, which were endogenous to latent factors for the 36 cells representing six core components crossed with six key processes at the second level. At the third level were latent factors for the six core components or key processes. At the fourth level was a single latent trait representing overall principal leadership (i.e., the total score). Because each item contributed to both a core component and a key process, the factor analytic model was split into two separate analyses: one on core components and the other on key processes. To gauge agreement between the two sets of models, factor scores for the overall leadership score were estimated for both models and the correlation between them was estimated.

The CFA models were fit using PROC CALIS as implemented in SAS 9.1.3. Results from the confirmatory factor analyses reveal that both the core components and the key processes models fit the data well, having goodness of fit indices between .96 and .99 for both the GFI and the Adjusted GFI. Root mean square error was .04 for form A and .03 for form C. Even after adjusting for model complexity, the parsimonious goodness of fit indices (PGFI, Mulaik et al., 1989) were still high, ranging from .93 to .96. All of the item factor loadings were salient, ranging from .41 to .94 with a median loading of .82. The second order factor loadings were also salient, ranging from .60 to 1.00 with a median loading of .92. The third order factor loadings were salient, ranging from .89 to 1.00 with a median loading of .98. Lastly, the correlation between overall leadership factor scores from the Core Components and Key Processes CFA models was .99. The increase in saliency across levels, the consistently high loadings at level 3, and the high correlation between overall scores from the two models suggests that the core components and the key processes have similar degrees of influence on the total score. In other words, the six core components and six key processes all contribute to the overall measure of principal leadership.

A second piece of validity evidence was obtained by examining the relationship of teacher ratings and principal ratings. A scatter plot of teacher and principal ratings is found in Figure 3.2. There are only eight data points because one principal did not complete the assessment. The scatter plot suggests that principals and teachers tended to give similar ratings of principals' effectiveness. For example, the principal who gave him/herself the lowest score was also rated the lowest by his/her teachers. The correlation of principal and teacher ratings in these eight data points is a moderate .47. This finding suggests the between principal variance for both teacher and principal data is measuring something in common, a form of concurrent validity. Scatter plots of supervisors' ratings are not included because of the supervisor who gave uniformly perfect effectiveness ratings across principals.

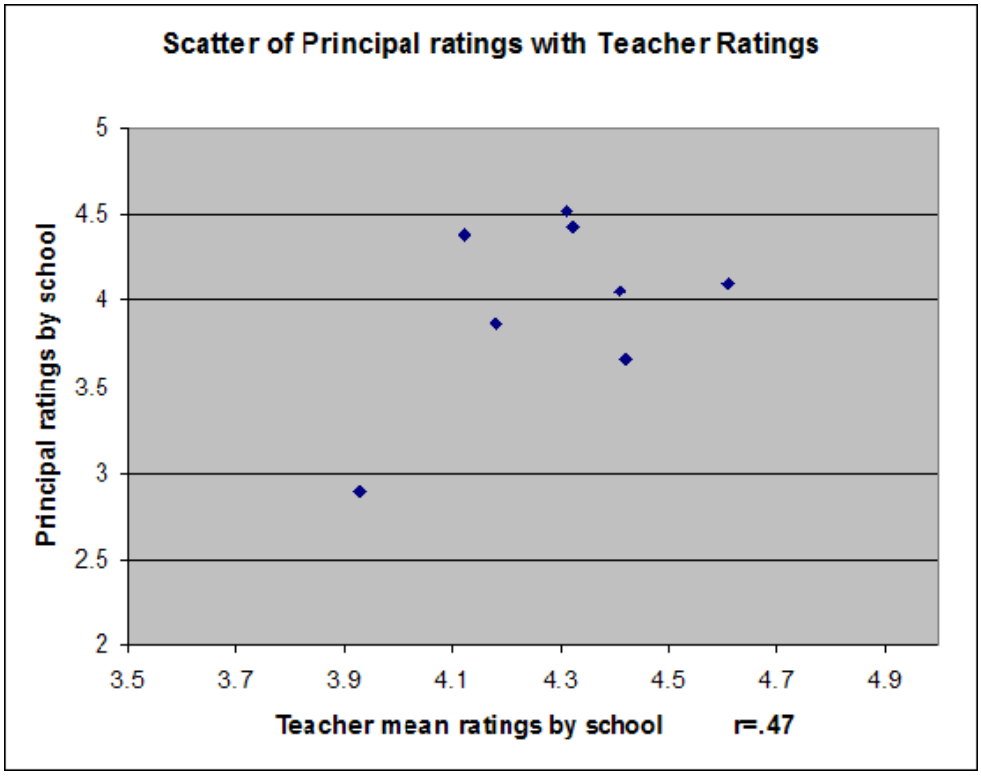


Figure 3.2: Scatterplot of principal ratings with mean teacher ratings for nine-school pilot

A third source of validity evidence is the core component and key process intercorrelations, provided in Tables 3.7 and 3.8. The correlations were high, both for core components and for key

processes, though they appear somewhat higher for key processes. For core components, correlations ranged from a low of .73 (Connections to External Communities and High Standards for Student Learning) to a high of .90 (Quality Instruction and High Standards for Student Learning). For key processes, correlations ranged from a low of .89 (Supporting and Monitoring) to a high of .94 (Monitoring and Communicating). Correlations of core components and key processes with total score were all high, with none lower than .9. These high intercorrelations, along with the factor analysis results described above, suggest that the instrument is measuring a strong underlying construct, principal leadership.

Table 3.7: Intercorrelations of core components

Correlations among core components, all schools, 9-school pilot, Spring 2007						
	High standards	Instruction	Curriculum	Culture	Connections	Performance Accountability
High standards	1.00					
Instruction	0.91	1.00				
Curriculum	0.84	0.90	1.00			
Culture	0.81	0.84	0.85	1.00		
Connections	0.73	0.78	0.81	0.81	1.00	
Performance Accountability	0.79	0.83	0.84	0.79	0.83	1.00
Total	0.91	0.95	0.94	0.92	0.90	0.92

Table 3.8: Intercorrelations of key processes

Correlations among key processes, all schools, 9-school pilot, Spring 2007						
	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1.00					
Implementing	0.93	1.00				
Supporting	0.91	0.92	1.00			
Advocating	0.91	0.93	0.90	1.00		
Communicating	0.92	0.92	0.92	0.92	1.00	
Monitoring	0.91	0.91	0.89	0.91	0.94	1.00
Total	0.96	0.97	0.96	0.96	0.97	0.96

Parallel forms. The data support the parallel nature of the two forms. Of course, the forms were created by stratified random assignment of items. For item-level mean ratings, a comparison of the two forms reveals that the distributions, except for one outlier on form A, were very similarly

shaped with similar ranges. Table 3.9 shows teacher mean ratings on scales and total score by form. The results reveal similar scores—all scale means are within .04 except one core component and two key processes. Mean scores on form A are 4.31, and mean scores on form C are 4.33. While these are not definitive data because there were only five schools for form A and four schools for form C, they suggest that teacher ratings on the two forms were roughly equal.

Table 3.9: Teacher ratings by form

Comparison of ratings by form and respondent type, 9-school pilot, Spring 2007		
	Teacher mean	
	Form A	Form C
High standards	4.39	4.38
Rigorous curriculum	4.34	4.35
Quality instruction	4.36	4.38
Culture of learning	4.37	4.38
External communities	4.15	4.19
Performance accountability	4.21	4.30
Planning	4.31	4.33
Implementing	4.31	4.34
Supporting	4.40	4.36
Advocating	4.22	4.27
Communicating	4.34	4.36
Monitoring	4.26	4.33
Total	4.31	4.33

Other results also suggest parallel forms. Missing and "Don't know" data show that the four scales most likely to be marked "Don't know" were the same on the two forms. Though form C had slightly higher rates of "Don't know" responses, this could be due to the fact that a much larger percentage of form C respondents were from middle or high schools than in form A. Internal consistency estimates in Table 3.6 are similar across forms. Evidence sources in Table 3.3 are similar across forms, with the mean number of sources of evidence used differing by just .01 between forms. Though none of the data reported here affirm that the forms are parallel, there is

good evidence that this is the case, and there is no evidence to the contrary. Given that schools were randomly assigned to forms and that there were only nine schools, the data could hardly be more supportive.

The overall message from the nine-school pilot study was straightforward. VAL-ED's items were clear to respondents, and respondents were willing to complete them. Reliability was excellent. Most importantly, the great majority of respondents agreed that VAL-ED measures key leadership behaviors.

Changes to the Instrument

In light of the findings of the nine-school pilot, several changes were made to the instrument. One of the key issues that arose in the pilot study was overall high effectiveness ratings given to principals. While high ratings on behavior rating scales are common, the evidence suggested ways to improve the rating scale to increase between-principal variance. One change arose from evidence that respondents were not selecting the "Not done" option that corresponded to 0 at the bottom of the scale. The "Not done" option had been included to emphasize the conceptual difference between not doing a behavior and doing it ineffectively. However, two issues led to the removal of the "Not done" category. First, the cognitive interviews suggested that some respondents were cued to a measure of frequency by the words "Not done." The VAL-ED was designed to measure effectiveness of behaviors, not frequency. Second, an alternate interpretation of ineffective could include not doing a behavior. For these reasons, the "Not done" category was removed and the scale was changed to a 1 to 5 scale.

A second change made was providing labels for all five of the effectiveness ratings, rather than just three in the original model. Although respondents in the cognitive interviews showed they generally understood the meanings of the unlabeled effectiveness ratings, labels were added to

levels 2 and 4. Labels were included because it was thought that, in conjunction with relabeling the level 3 and 5 ratings, this change would result in increased spread of ratings.

A third change made was to re-label the level 3 and 5 ratings. The goal was to stretch the top end of the distribution, so "Highly effective" was moved to level 4, and level 5 was renamed "Outstandingly effective." Level 3 was renamed from "Moderately effective" to "Satisfactorily effective," and level 2 was named "Minimally effective." To further emphasize the exceptional principal behaviors to be rated "Outstandingly effective," a sentence describing outstandingly effective behaviors was added on the directions page. A sentence describing ineffective behaviors was also added. The set of changes to the rating scale described here was designed to have the effect of stretching the distribution to create more between-school variance in ratings of principal effectiveness.

Fourth, the number of items was reduced from 108 to 72. Respondents to the pilot study overwhelmingly indicated that the form was too long. Additionally, our contacts in districts and states suggested that VAL-ED would be more useful to schools and districts if it took less than 30 minutes to complete. There was concern that removing items might reduce the reliability of the instrument, but the findings above make clear this was not the case—reliabilities remained high even after removing 36 items. To accomplish the change, items were randomly selected one per cell to be removed from each form.

The only exception to the random removal of items was an outlier item, which read "The principal ensures the school uses data on parent involvement in teacher evaluations." The item had a mean teacher rating of 3.56, roughly .4 lower than any other item on either form. Also, without reviewing item ratings, the item was identified by a former school principal and superintendent as a

problematic item. The item was replaced with an item randomly selected from the remaining items in the pool of items for the cell (Connections to External Communities and Monitoring).

Fifth, in conjunction with the removal of "Not done" from the rating scale, an additional change was made to focus respondents on effectiveness rather than frequency. The item stem was changed from "The principal ensures the school …" to "How effective is the principal at ensuring the school …." This stem fits more appropriately with the response scale and adds "effective" to the stem, emphasizing that the instrument is measuring effectiveness.

Cognitive Labs of the On-line Version (Study 5).

To ensure that respondents would use the online version of the VAL-ED as intended, we conducted a round of cognitive labs with potential users. Three sets of cognitive labs were conducted. The interviews were conducted in one school in a rural district and two schools in a suburban district in two states. In total, there were seven participants—a principal and teacher from each of the schools, and one supervisor of a principal. Interviews were conducted by research assistants and extensive notes were taken. Audio recordings were also taken for verification purposes and in case exact quotes were needed.

Each cognitive lab consisted of the same format. First, subjects were introduced to the cognitive lab "think aloud" methodology. Next, respondents were asked to use their web browser to load the online prototype of the VAL-ED instrument and log in to the system. After reading aloud and examining the directions page, respondents were asked to comment on language, aesthetics, and clarity. They were also asked questions about the functionality of the online interface and several substantive changes to the directions that had been made based on previous studies. In the next step, respondents began the survey and were asked to examine the online interface for its functionality and ease of use. Respondents then read the survey aloud, item-by-item, describing their thought

process as they identified sources of evidence and checked effectiveness ratings. (Respondents were encouraged to complete the assessment as if they were actually evaluating a principal.) After the respondent completed the assessment for several core components, the respondent was asked to "log out" and evaluate the instrument by responding to summary questions.

The purpose of the cognitive labs was twofold. First, we sought to examine potential users' reactions to the online prototype of the VAL-ED. Second, we sought to examine several changes made to the instrument after the nine-school pilot.

With regard to the first purpose, respondents generally found the online instrument easy to use. There were several small concerns regarding screen resolution, the layout of the instructions, and the ability to click on terms to see definitions. However, respondents were easily able to navigate through the survey without guidance, and a majority preferred the online assessment to a potential pen-and-paper version.

With regard to the second purpose, respondents generally did not voice any concerns regarding the recent changes made to the instrument. Respondents felt comfortable using the full scale and understood the meaning of the different effectiveness ratings. Respondents also estimated a short amount of time required to complete the assessment, suggesting agreement with our decision to shorten the instrument from 108 to 72 items. Overall, there were no serious concerns about the changes made after the nine-school pilot.

The cognitive labs provided critical feedback for revising the online prototype of the VAL-ED instrument and approving recent modifications to the instrument in anticipation of further development and wide-scale use.

Eleven-School Pilot Test (Study 6)

After the substantial changes made to the instrument resulting from the nine-school pilot, an additional pilot study was conducted to examine the effects of the revisions. The methods for the study were identical to the nine school pilot, except that eleven schools in four districts in a second Midwest state participated in the study. The forms used in the study were the updated 72-item forms with the modified stem and response categories. The primary concerns for this pilot were the distributions of teacher, principal, and supervisor effectiveness ratings. We focus on results that bear on the changes made after the nine-school pilot.

Results

Results support that the changes made had the desired effects. Mean teacher responses for the eleven school pilot were 3.29 for total score, ranging from a low of 3.10 (Connections to External Communities) to a high of 3.37 (Culture of Learning & Professional Behavior). Comparing these results to the results in Table 9 for the nine-school pilot, we see that teacher scores were over a full point lower after the revisions. Principal (3.72) and supervisor (3.77) total scores were also lower on the eleven-school pilot than the nine-school pilot. Furthermore, results were more spread. In the nine-school pilot, school-level teacher means ranged from a low of 3.93 to a high of 4.61, a spread of less than .7 points on the 5-point effectiveness scale. In the eleven-school pilot, school-level teacher means were as low as 2.81 and as high as 3.90, a spread of over a full point. While it is possible that the principals in the eleven-school pilot were less effective and more variable in quality than those in the nine-school pilot, it is more likely that these results suggest the re-scaling was effective in lowering and increasing the spread of effectiveness ratings.

A second change to arise from the nine-school pilot was a reduction from 108 items to 72 items. This change was made because 81 of approximately 350 respondents expressed concerns about length in comments at the end of the forms. Furthermore, respondents in cognitive interviews

had argued that the form was too long at 108 items. In the eleven-school pilot, the reduction in items had little effect on reliability; principal and supervisor scale and total score reliabilities remained above .89, and teacher scale and total score reliabilities remained above .94. Also, the number of respondents commenting on the length of the instrument decreased to 30 out of over 500, suggesting that length was less of a concern.

The changes made to the instrument also had the effect of improving the correlations among the response groups. A scatterplot of teacher and principal mean effectiveness ratings by school is provided in Figure 3.3. The correlation was .79. For the individual scales, correlations ranged from .68 to .88. The correlation for teacher-supervisor on total score was .68, and the correlation for principal-supervisor was .51. Across all response groups, there was strong agreement about principal effectiveness.

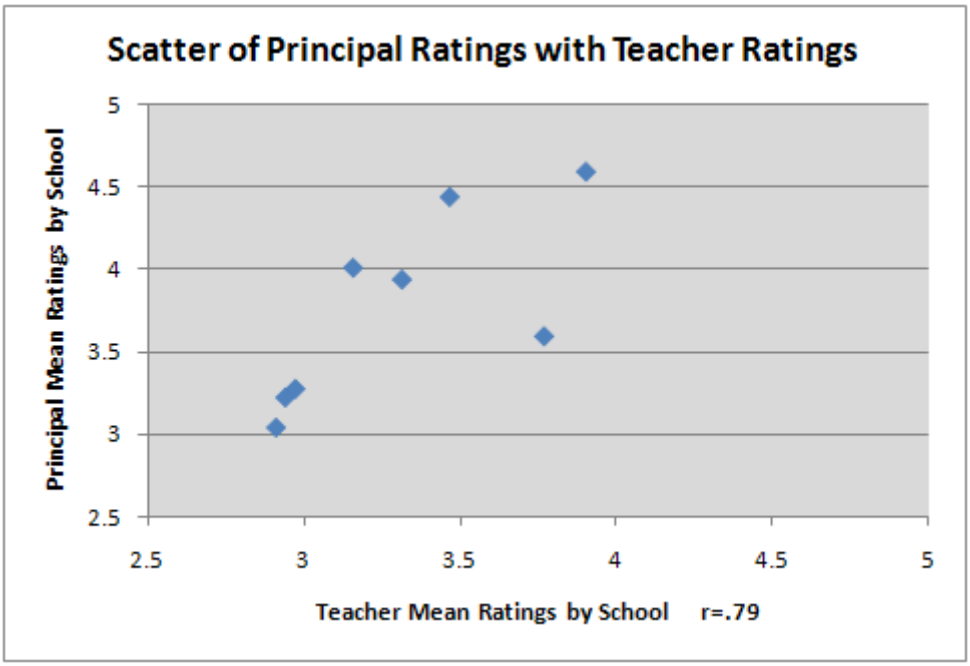


Figure 3.3: Scatterplot of principal ratings with mean teacher ratings for eleven-school pilot

Overall, the results of the eleven-school pilot suggested that the changes made after the nine-school pilot were successful. Effectiveness ratings were lower and more variable with higher levels of agreement, and there was far less concern about the assessment's length.

Summary of Development Work

The first, and most important step in establishing the content validity of the VAL-ED was the item development phase. In this phase, we used an iterative item-writing process based on the 36-cell conceptual framework. The team of researchers repeatedly examined and revised the items to insure a) proper fit to the framework and b) proper grain size. While item writing was ongoing, important decisions were also made about the format of the instrument, the item stem, and other key factors that contributed to the VAL-ED.

After initial item writing and instrument development, we conducted a series of studies to gather initial evidence as to the content and construct validity of the instrument, as well as to suggest necessary modifications to the instrument. These studies were a sorting study, a series of cognitive interviews of both the paper-and-pencil and online version, a bias review study, a nine-school pilot study, and an 11-school pilot study. After these studies and subsequent revisions, the VAL-ED was deemed ready for a large scale national field test.

Chapter 4:
National Standardization and Standard Setting

National Field Trial

In the spring of 2008, a national field trial of the VAL-ED was conducted. There were five primary purposes to be served by the field trial. First, although the VAL-ED's psychometric properties had been repeatedly investigated in a variety of studies and through two pilots, samples for the pilots were neither nationally representative nor large. The national field trial targeted a nationally representative sample of 300 schools on which to investigate psychometric properties. Second, in addition to reporting results on the VAL-ED in terms of mean item response on the effectiveness rating scale (ranging from 1.0 to 5.0), VAL-ED results are to be reported in terms of principals' percentile ranks against national norms and also in terms of performance standards. A second purpose of the national field trial was to establish initial norms for reporting a principal's percentile rank in overall effectiveness and on each of the twelve subscales, in terms of performance aggregated across respondent groups as well as by respondent group. As for performance standards, performance level descriptors were written for distinguished, proficient, basic, and below basic. These were used to guide a Bookmark standard setting process. The Bookmark procedure requires that items be ordered according to difficulty and that impact data be provided. Thus, a third purpose of the national field trial was to create the item-ordered booklets and to provide impact data on what percent of principals would be judged to be, for example, proficient if the standard was set at a particular point on the effectiveness scale.

Respondents to the VAL-ED assessment of principal leadership in the national field trial also completed a feasibility questionnaire, the same feasibility questionnaire that had been used in the

two pilot studies. Nine questions were asked in the survey to determine the extent to which respondents found the instrument understandable and easy to use, focused on important leadership behaviors, and appropriate for principal accountability purposes in their district. Thus, a fourth purpose of the national field trial was to investigate issues of feasibility of the VAL-ED.

Yet a fifth purpose to be served by the national field trial was to investigate possible differences in principal performance according to design factors. The national field trial involved a stratified random sample. The strata were: a) level of schooling (elementary, middle, or high school), b) geographic distribution (Northeast, South, Midwest, or West), c) locale (urban, suburban, or rural), and d) whether or not the school was in a Wallace Foundation-supported site. Investigating the extent to which principal effectiveness varied with these design variables does not bear directly on the psychometric properties of the instrument, but it does address questions of substantive interest to the funding source, the Wallace Foundation, as well as to others who study principal leadership.

Design of the Sample

For the nationally-representative field trial, 300 schools were targeted, to be selected from four regions of the United States (Northeast, South, Midwest, and West as defined by the US census) of which 100 were to be elementary, 100 middle, and 100 high schools. There were to be 150 urban schools, reflecting the Wallace Foundation's focus on urban education, 100 suburban schools, and 50 rural schools. In addition, the sample design called for the 150 urban schools to include 50 drawn from Wallace grantee districts, 50 from Wallace grantee states, and 50 urban schools drawn from non-Wallace grantee districts and states. With the exception of the Wallace districts, districts were sampled randomly with probability in proportion to student enrollment. Once a district had agreed to participate, schools within that district were selected using a simple

random sample. When a district declined to participate, a replacement district from that stratum was again randomly selected with probability in proportion to size. Similarly, when a school within a district refused to participate, another school was selected using a simple random sample.

Ninety-nine districts were contacted and 60 agreed to participate for a 61% "response" rate. Twenty-nine districts declined to participate and 10 additional districts agreed to participate but subsequently dropped out. By region, 18 of 23, or 78%, participated from the Northeast, 16 of 23, or 70%, from the South, 14 of 28, or 50% from the Midwest, and 12 of 25, or 48% of the West. Of suburban districts, 27 of 51 participated, or 53%, for rural, 20 of 32, or 63%, and for urban, 13 of 16, or 81%.

Across the 60 districts that agreed to participate, 109 elementary, 100 middle, and 100 high schools were recruited to participate of the 461 initially selected for a 67% participation rate.

The analysis file has data on principals from 235 schools, data on supervisors from 253 schools, and data on teachers from 245 schools, amounting to responses from 8,863 teachers (4,140 for Form A and 4,723 for Form C). There were 218 schools for which there were data from all three response groups, including data from 6,391 teachers, and 276 schools for which there were data from at least one respondent group (89%).

Of the 218 schools with data from all three response groups, 39% were elementary schools, 32% middle, and 28% high schools. Twenty-three percent of the schools were from the West, 30% from the South, 22% from the Midwest, and 25% from the Northeast. There were 39% urban schools, 39% suburban schools, and 22% rural schools. Twenty-nine percent of the schools were from Wallace-funded sites. Thus, the obtained sample, in terms of its design parameters and composition, matches well the intended sample, with the exception that urban schools were

underrepresented at 39% in comparison to their target of 50%, while both suburban and rural schools were slightly more represented than targeted.

Based on the 245 schools from which at least some teacher data were returned, teacher response was variable with the median across participating schools of teacher response equal to 68%. Twenty-five percent of the schools had a response rate of 78% or better and 75% of the schools had a response rate of 54% or better.

Forms A and C were randomly assigned to districts in equal number as districts were recruited. In the obtained sample of 218 schools with data from all three response groups, 115 schools used Form C and 103 used Form A. For the field trial, only the paper version of the VAL-ED was used.

Design Factors

As described above, the national field trial sample was stratified by level of schooling, locale, and region of the country. The data allow investigation of the extent to which there were significant differences in principal effectiveness as measured by the VAL-ED.

Respondent Groups

In what follows, design factors are investigated first for the aggregate sample across all respondent groups and then for each of the respondent groups. Before going into those results, it is useful to ask to what extent one respondent group differed from another for the entire sample of schools for which data were available for the respondent group. As seen in Table 4.1, there were 235 principals, 253 supervisors, and 245 sets of teachers. Supervisors gave principals the most positive evaluations, 3.68, while principals gave themselves the least positive evaluations, 3.53, and teachers were in between, 3.59. The difference between supervisors and principals is statistically significant at the .05 level, using Tukey post-hoc pairwise comparisons.

Table 4.1 Mean Effectiveness Rating by Respondent Group*	
Principal	3.53 (235)
Supervisor	3.68 (253)
Teacher	3.59 (245)
*Supervisor ratings are significantly higher than principal	

The contrast of all supervisors to all principals and all teachers does not hold the sample of schools/principals completely constant. For the 218 schools for which data was available on all three respondent groups, the means were rank-ordered the same, with supervisors most positive, 3.72, followed by teachers, 3.60, and principals, 3.52. For the 218 schools, supervisors were significantly more positive than both teachers and principals. Apparently, principals interpreted the task of self-evaluation in roughly the same way as did their supervisors and their teachers.

Level of Schooling

Table 4.2 provides mean effectiveness ratings by level of schooling for the aggregated sample as well as each of the three respondent groups. Sample sizes are given in parentheses. Table 4.2 should be inspected row by row because the featured contrasts are levels of schooling, not respondent groups. There are no significant differences among levels of schooling, either for the aggregate sample or for each of the three response groups. While high school, middle school, and elementary school principals are seen as equally effective by the three respondent groups, in each analysis high school principals had the lowest effectiveness ratings.

Table 4.2 Mean Effectiveness Rating by Level of Schooling			
	Elementary	Middle	High School
Overall	3.65 (86)	3.65 (70)	3.52 (62)
Principal	3.56 (92)	3.56 (73)	3.45 (70)
Supervisor	3.69 (95)	3.73 (83)	3.60 (75)
Teacher	3.66 (94)	3.59 (75)	3.52 (76)

The lack of significant differences among levels of schooling was reassuring because, among other things and as will be seen elsewhere, we set performance standards (distinguished, proficient, basic, and below basic) regardless of level of schooling.

Effectiveness Ratings by Locale

Table 4.3 shows the comparisons among suburban, urban, and rural schools. On the aggregate sample, principals in suburban schools are judged to be on average more effective, 3.66, than principals in rural schools, 3.50, with urban schools in between, 3.63. The standard deviation is approximately .35. The difference between suburban and rural schools is statistically significant at the .05 level using Tukey post-hoc pairwise comparisons. The difference of .16 on the 5-point mean item response effectiveness scale is large in comparison to the types of differences seen for other design factors.

Table 4.3 Mean Effectiveness Rating by Locale

	Suburban	Urban	Rural
Overall	3.66 (84)	3.63 (85)	3.50 (49)
Principal	3.55 (91)	3.52 (95)	3.49 (49)
Supervisor	3.77 (96)	3.69 (102)	3.49 (55)
Teacher	3.64 (96)	3.59 (99)	3.50 (50)

When one looks by respondent group, the differences between suburban, urban, and rural were not statistically significantly different for both principals and teachers, although the rank order from suburban to urban to rural was maintained. For supervisors, again the rank order of mean effectiveness ratings was maintained from suburban to urban to rural, and suburban principals were rated significantly more effective than rural principals.

Effectiveness Ratings by Geographic Region

The results contrasting geographic regions in terms of principal effectiveness are found in Table 4.4. Principals in the Northeast, 3.68, were given a slightly higher effectiveness rating than

principals in the South, 3.67, and principals in those two geographic regions were given more effective ratings than principals in the Midwest, 3.60, or West, 3.48. The difference between Northeast and South was not statistically significant at the .05 level using Tukey post-hoc pairwise contrasts. Neither was the difference between Midwest and West, but principals in the Northeast and South were significantly more effective than principals in the West, again at the .05 level using Tukey post-hoc pairwise comparisons. For principals, the results were similar. The Northeast and South were rated most effective and the Midwest and West least effective. In both the South and Northeast, principal effectiveness ratings were statistically significantly higher than principal effectiveness ratings in the West at the .05 level using Tukey contrasts. For supervisors, there were no statistically significant differences among geographic regions, but for teachers, again effectiveness ratings were rank-ordered with the Northeast and South at the top and Midwest and West at the bottom. In the teacher data, principals were rated significantly more effective in the South than in the West. None of the other pairwise comparisons were statistically significant among geographic regions for the teacher data. When interpreting this design factor, it may be important to remember that not only were principals in the West and Midwest seen as less effective but the participation rates were lower for those two regions as well. Thus, the reasons for differences are not clear.

Table 4.4 Mean Effectiveness Rating by Geographic Region

	Northeast	South	Midwest	West
Overall	3.68 (54)	3.67 (66)	3.60 (47)	3.48 (51)
Principal	3.62 (58)	3.63 (77)	3.47 (49)	3.31 (51)
Supervisor	3.81 (60)	3.64 (80)	3.68 (54)	3.59 (59)
Teacher	3.60 (60)	3.68 (79)	3.57 (52)	3.48 (54)

Effectiveness Ratings by Wallace Support

The Wallace Foundation has made major investments in supporting improvement in principal effectiveness. Some of their programs are with specific districts and some are state-wide. Districts with Wallace Support and districts in states with Wallace support were oversampled in the national field trial. In fact, of the 150 urban schools (Wallace's support for building principal effectiveness is targeted on urban schools), 100 of them either came from Wallace-supported districts or districts in Wallace-supported states. In Table 4.5, principal effectiveness in schools in districts and states with Wallace support was contrasted with principal effectiveness in schools in districts and states without Wallace support. None of the differences are statistically significant at the .05 level and all of the differences are small.

Table 4.5 Mean Effectiveness Rating by Wallace Support

	Wallace	Not
Overall	3.63 (63)	3.60 (155)
Principal	3.55 (68)	3.52 (167)
Supervisor	3.64 (78)	3.70 (175)
Teacher	3.62 (71)	3.58 (174)

The finding of no difference between Wallace and non-Wallace supported sites does not mean that the Wallace investment has had no effect. These data do not take into account whether the principals actually received Wallace training, or if they did, for how long and how long they had been in their current school. There are many other factors that are not controlled in the contrast. The ideal study would randomly assign principals to receive Wallace support or not; the national field trial was not designed for purposes of teasing out causal effects of Wallace support.

Nevertheless, we took one closer look by contrasting principals in districts or states with Wallace support to principals in districts or states without Wallace support, but holding constant that

the schools were in an urban locale. Again, no statistically significant differences were found. The Wallace mean was overall (the aggregate variable) 3.63 and the non-Wallace mean was 3.61.

Mean Effectiveness Ratings by Form

Table 4.6 contrasts Form A to Form C for the aggregate sample as well as each respondent group. None of the differences between the two forms are statistically significant at the .05 level. All of the differences are .04 points on the effectiveness scale or smaller. These findings support that forms A and C were not only built to be parallel, but they are operating as parallel forms.

Table 4.6 Mean Effectiveness Rating by Form

	Form A	Form C
Overall	3.62 (103)	3.61 (115)
Principal	3.51 (106)	3.54 (129)
Supervisor	3.66 (126)	3.70 (127)
Teacher	3.61 (114)	3.57 (131)

Further Analyses of Design Factors

We used regression analyses to see if design variables and other variables were significant predictors of principal effectiveness (Table 4.7). These analyses were done on the aggregate sample as well as the sample for each response group. The difference in the analyses reported above and these regression analyses is that many more variables were included. Specifically, teachers' years of experience in the school was included, as was the number of teachers in the school as an indicator of school size. The analyses also included the nine questions in the feasibility survey the respondents completed after having taken the VAL-ED. Results for the feasibility questions will be reported elsewhere. Here, however, two results from these analyses are relevant. First, while the "number of teachers" variable was not significant in any of the four analyses, nor did it approach significance, principals' years of experience in the school was statistically significant at the .05 level for the aggregate sample, but not for any of the three respondent groups by themselves (though the P-value

was .06 for supervisors). Principals with more years of experience in the school were rated more highly when data were aggregated across teachers, supervisor, and principal. The other finding is that even in the regression analyses where all variables entered into the equation serve as controls for the others, the design factor of principals in schools in the West having lower effectiveness was replicated. There was also a design factor of principals in high schools having lower effectiveness ratings, consistent with the earlier rank ordering of means that was not statistically significant. The design factor for school locale (urban, suburban, rural) was not replicated and the lack of the design factor for Wallace-supported schools was replicated.

Table 4.7 Design Factors Regression Results for Total Score Aggregated Sample		
Design Factors	**Coefficient**	**Std. Error**
Number of Teachers	0.00005	0.001
Wallace	0.07	0.09
Midwest	-0.08	0.07
West	-0.21**	0.07
South	-0.05	0.07
Suburban	0.08	0.09
Rural	0.02	0.10
Form C	0.03	0.05
Middle	-0.07	0.06
High	-0.17*	0.07
Teacher Response Rate	0.002	0.001
Years Principal	0.008	0.004
Principal Feasibility		
I found this response form easy to use.	-0.06	0.04
I understood the vast majority of items.	0.03	0.05
I believe the vast majority of items focus on important leadership behaviors.	0.04	0.06
I do not believe the items are biased against any race or gender of a principal being assessed.	-0.004	0.04
This assessment is appropriate for use at the elementary, middle, and high school levels.	-0.03	0.04
I would prefer a web-based format for this assessment over the paper-and-pencil version I just completed.	-0.02	0.03
Teachers should have input into the assessment of their principal's leadership.	0.02	0.04
I would support the use of this assessment instrument to hold principals accountable in my district.	0.02	0.03
The amount of time required to complete this instrument is reasonable.	0.07	0.04
Supervisor Feasibility		
I found this response form easy to use.	0.15**	0.05

I understood the vast majority of items.	-0.003	0.06
I believe the vast majority of items focus on important leadership behaviors.	0.04	0.06
I do not believe the items are biased against any race or gender of a principal being assessed.	-0.006	0.05
This assessment is appropriate for use at the elementary, middle, and high school levels.	0.04	0.05
I would prefer a web-based format for this assessment over the paper-and-pencil version I just completed.	-0.01	0.03
Teachers should have input into the assessment of their principal's leadership.	0.02	0.04
I would support the use of this assessment instrument to hold principals accountable in my district.	-0.01	0.04
The amount of time required to complete this instrument is reasonable.	-0.11*	0.05
Teacher Feasibility		
I found this response form easy to use.	0.16	0.17
I understood the vast majority of items.	0.39	0.21
I believe the vast majority of items focus on important leadership behaviors.	0.16	0.15
I do not believe the items are biased against any race or gender of a principal being assessed.	0.12	0.19
This assessment is appropriate for use at the elementary, middle, and high school levels.	0.05	0.19
I would prefer a web-based format for this assessment over the paper-and-pencil version I just completed.	-0.04	0.09
Teachers should have input into the assessment of their principal's leadership.	0.03	0.18
I would support the use of this assessment instrument to hold principals accountable in my district.	-0.14	0.13
The amount of time required to complete this instrument is reasonable.	-0.18	0.15
*p<.05, **p<.01, p<***.001		

Reliability

Internal Consistency

Data from the national field trial were used to estimate the internal consistency reliability of the VAL-ED for a total score and for each of the six core components and six key processes subscales, separately for Form A and Form C. Table 4.8 provides the results. In all cases, the internal consistency reliabilities estimated with Cronbach's Alpha using pairwise deletion for missing data were high for total score across respondent groups and forms. The internal consistency reliability ranged from .98 to .99, nearly perfect. The pattern of reliabilities between the two forms

of the VAL-ED was virtually the same, again supporting that we have parallel forms that can be used interchangeably. Internal consistencies were slightly lower for principals than for supervisors or teachers and the reliabilities for key processes based on principal data were slightly lower than the reliabilities for core components. The conclusion is that there is strong internal consistency reliability for both total score and across the twelve subscales, as based on national field trial data. These results replicate the promising results on internal consistency reliability from each of the two previous pilot studies.

Table 4.8 Internal Consistency Reliability

Respondent Group	Principal		Supervisor		Teachers	
Form	A	C	A	C	A	C
Total Score	0.98	0.98	0.99	0.99	0.99	0.99
Core Components						
High Standards	0.89	0.91	0.96	0.95	0.96	0.95
Rigorous Curriculum	0.90	0.92	0.96	0.95	0.96	0.95
Quality Instruction	0.90	0.89	0.95	0.94	0.95	0.94
Culture of Learning	0.90	0.93	0.96	0.95	0.96	0.96
External Community	0.92	0.91	0.96	0.93	0.97	0.96
Performance Accountability	0.91	0.92	0.97	0.95	0.97	0.97
Key Processes						
Planning	0.88	0.90	0.96	0.93	0.96	0.95
Implementing	0.88	0.89	0.96	0.93	0.96	0.95
Supporting	0.90	0.88	0.96	0.94	0.95	0.95
Advocating	0.87	0.89	0.95	0.93	0.95	0.95
Communicating	0.87	0.91	0.95	0.94	0.95	0.95
Monitoring	0.89	0.91	0.95	0.94	0.96	0.95

From Table 4.8, the lowest internal consistency reliability coefficient for total score is .98 and that is for principal data. The supervisor data and teacher data both yielded internal consistency estimates of .99 for total score. When results are aggregated across respondent groups, the reliability of the aggregate total score is certainly at least .98 reliable. The standard deviation of the aggregate total score is .35. Thus, the standard error for the aggregate total score is at least as small as .05. For example, the principal's aggregated total score effectiveness would have a

68% confidence of falling within a range plus or minus .05 from the reported score and a confidence of 95% of falling within a range plus or minus .10. Standard errors of measurement for disaggregated data in the subscales vary, as is shown in Table 4.9. Using the reliability of .9 and the standard deviation of .55, a standard error of measurement is .17. Most of the subscales are .9 reliable or greater and as seen in Table 4.24 and Table 4.26, most of the standard deviations are .55 or lower. The standard deviations for supervisor data, as reported in Table 4.25, however, range from .6 to .8. Thus, all standard errors of measurement across subscales, forms, and respondents are .20 or less.

Table 4.9 Standard Error of Measurement

Respondent Group	Principal		Supervisor		Teachers	
Form	A	C	A	C	A	C
Total Score	0.07	0.07	0.07	0.06	0.04	0.04
Core Components						
High Standards	0.17	0.16	0.15	0.14	0.09	0.09
Rigorous Curriculum	0.17	0.17	0.15	0.15	0.08	0.09
Quality Instruction	0.18	0.19	0.17	0.15	0.10	0.10
Culture of Learning	0.18	0.15	0.14	0.14	0.09	0.09
External Community	0.18	0.18	0.15	0.17	0.08	0.08
Performance Accountability	0.17	0.17	0.13	0.15	0.08	0.07
Key Processes						
Planning	0.18	0.17	0.15	0.16	0.09	0.09
Implementing	0.17	0.17	0.15	0.17	0.09	0.09
Supporting	0.17	0.18	0.14	0.15	0.10	0.10
Advocating	0.18	0.18	0.17	0.15	0.09	0.09
Communicating	0.20	0.17	0.16	0.15	0.10	0.10
Monitoring	0.18	0.18	0.17	0.17	0.09	0.10

Reliability of Differences between Subscales. The intention is to report principal effectiveness not only on the total score but also by core component and by key process. As suggested by the high internal consistency reliabilities for total score, and as will be seen in another section of this report, the intercorrelations among core components and the intercorrelations among key processes are high. Of course, there is a redundancy between either set of six subscales and the

total score since they are based on the same 72 items. Further, there is a complete redundancy between the set of core components and the set of key processes because again, they are based on the same 72 items.

The conceptual framework of the VAL-ED, which is six core components by six key processes identifying 36 domains of principal behavior, makes investigation of the factor structure difficult. Ideally, there would be 36 factors, one for each of the 36 cells in the six-by-six conceptual framework. In the instrument however, each of the 36 cells is measured by only two items, making the finding of the 36-factor structure highly unlikely. Neither is it likely to find a 12-factor structure, one for each of the six core components, plus one for each of the six key processes because one set of six is totally redundant with the other set of six. While the analyses presented in the following sections provide information about factor structure, the most important questions are whether the core components and key processes can be distinguished from the overall score, whether one core component can be reliably distinguished from another, and whether one key process can be reliably distinguished from another.

The first way to examine the reliability of the difference between subscales is to use the classical approach (Stanley, 1967). His formula for the reliability of the difference is

$$\frac{\rho_{11'}\sigma_1^2 + \rho_{22'}\sigma_2^2 - 2\rho_{12'}\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho_{12}\sigma_1\sigma_2}.$$

Tables 4.10, 4.11, and 4.12 show the reliability of the difference between scales for principals, teachers, and supervisors, respectively. Each table includes the results for Form A at the top and Form C at the bottom. In a few cases there are negative values for the reliability of the difference between two scales. These are due to the use of Cronbach's alpha for reliability, which is only the true lower bound of reliability under the assumption of uncorrelated errors (Raykov, 1997).

The first comparison that can be made is between core components/key processes and the total score. This comparison is shown in the bottom row of each of the four matrices in each of the three tables. For core components, Culture of Learning, Connections to External Communities, and Performance Accountability can all be reliably distinguished from the total score for all respondents. The other three core components have somewhat reliable difference scores for principals and teachers, but not very reliable difference scores for supervisors. There are a few instances for supervisors where the reliability of the difference score with total score is zero.

The results are less positive for comparing key processes with total score. With the exception of Advocating for supervisors on Form C, there are no key processes that can be reliably distinguished from the total score for principals or supervisors. For teachers, Monitoring and, to a lesser extent, Advocating can be reliably distinguished from the total score. In part, these somewhat poor results are a function of the very high correlations between subscales and the total score for key processes.

We can also compare core components with each other and key processes with each other. In terms of differentiating core components from one another, the reliability of the difference calculations indicate that Culture of Learning, Connections to External Communities, and Performance Accountability are highly distinguished from each other and from the other core components across respondent groups. For these three core components, all but two of the comparisons with other core components have reliability greater than .50 across the six form-by-respondent analyses. For teachers in particular the reliabilities are high for these three core components, with all reliabilities above .68 except for the comparison between Culture of Learning and Quality Instruction on both forms. Connections to External Communities is the best differentiated from the other core components across respondents and forms. The other three core

components, Rigorous Curriculum, High Standards for Student Learning, and Quality Instruction, are well differentiated from one another for teachers but not as well differentiated for principals and supervisors. Overall, however, the results suggest that the core components can be reliably distinguished from one another, especially for teachers.

The results are more mixed for comparisons of key processes to one another. For principals and supervisors, there are few reliabilities greater than .50, with the exception of four comparisons for supervisors on Form C. Planning, Implementing, and Supporting are especially poorly differentiated from one another for these two respondent groups, with no reliability greater than .33. For teachers, the results are moderate to positive. The large majority of reliabilities are between .40 and .60. There are a few particularly strong comparisons, such as the comparison between Advocating and Monitoring and the comparison between Supporting and Monitoring. Nevertheless, there are no comparisons with reliabilities greater than .70 for any respondent group for any key process. Still, given the traditionally low reliability associated with difference scores, the overall results from the traditional reliability of the difference analysis are mostly positive. There is good support for the differentiability of core components and mixed but positive support for key processes.

Table 4.10. Reliability of the Difference Between Subscales, Principal						
FORM A	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	0.43	1				
Quality Instruction	0.46	0.40	1			
Culture of Learning	0.60	0.69	0.54	1		
External Community	0.72	0.73	0.71	0.74	1	
Performance Accountability	0.63	0.62	0.60	0.72	0.74	1
Total Score	0.32	0.38	0.25	0.55	0.68	0.56

FORM A	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	0.14	1				
Supporting	0.17	0.17	1			
Advocating	0.25	0.25	0.22	1		
Communicating	0.38	0.26	0.36	0.24	1	
Monitoring	0.34	0.08	0.35	0.32	0.04	1
Total Score	0.00	-0.23	-0.01	-0.04	-0.01	-0.05

FORM C	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	0.36	1				
Quality Instruction	0.57	0.49	1			
Culture of Learning	0.64	0.69	0.59	1		
External Community	0.75	0.73	0.64	0.72	1	
Performance Accountability	0.72	0.72	0.66	0.67	0.58	1
Total Score	0.49	0.47	0.31	0.54	0.61	0.57

FORM C	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	0.05	1				
Supporting	0.06	-0.10	1			
Advocating	0.35	0.12	0.17	1		
Communicating	0.44	0.39	0.24	0.43	1	
Monitoring	0.51	0.49	0.43	0.55	0.32	1
Total Score	0.06	-0.20	-0.43	0.14	0.14	0.37

Table 4.11. Reliability of the Difference Between Subscales, Teacher						
FORM A	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	0.60	1				
Quality Instruction	0.64	0.58	1			
Culture of Learning	0.70	0.70	0.58	1		
External Community	0.84	0.84	0.82	0.79	1	
Performance Accountability	0.74	0.73	0.69	0.70	0.82	1
Total Score	0.58	0.56	0.41	0.48	0.84	0.66

FORM A	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	0.26	1				
Supporting	0.48	0.42	1			
Advocating	0.48	0.50	0.55	1		
Communicating	0.45	0.47	0.52	0.45	1	
Monitoring	0.54	0.58	0.60	0.60	0.52	1
Total Score	0.15	0.14	0.32	0.37	0.25	0.52

FORM C	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	0.58	1				
Quality Instruction	0.66	0.50	1			
Culture of Learning	0.69	0.69	0.64	1		
External Community	0.80	0.79	0.75	0.76	1	
Performance Accountability	0.79	0.76	0.72	0.77	0.81	1
Total Score	0.57	0.48	0.42	0.55	0.77	0.74

FORM C	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	0.20	1				
Supporting	0.40	0.25	1			
Advocating	0.49	0.48	0.57	1		
Communicating	0.46	0.45	0.46	0.58	1	
Monitoring	0.54	0.55	0.59	0.66	0.47	1
Total Score	0.11	-0.02	0.17	0.50	0.25	0.51

Table 4.12. Reliability of the Difference Between Subscales, Supervisor						
FORM A	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					total
Rigorous Curriculum	0.29	1				
Quality Instruction	0.16	0.38	1			
Culture of Learning	0.59	0.63	0.41	1		
External Community	0.75	0.78	0.77	0.72	1	
Performance Accountability	0.49	0.62	0.60	0.73	0.78	1
Total Score	-0.01	0.35	0.14	0.48	0.77	0.57

FORM A	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	0.26	1				
Supporting	0.17	0.07	1			
Advocating	0.37	0.38	0.36	1		
Communicating	0.14	0.00	-0.40	0.10	1	
Monitoring	0.07	0.31	0.03	0.41	0.02	1
Total Score	-0.05	-0.05	-0.51	0.22	-0.82	-0.05

FORM C	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	0.42	1				
Quality Instruction	0.48	0.37	1			
Culture of Learning	0.70	0.69	0.59	1		
External Community	0.80	0.78	0.70	0.75	1	
Performance Accountability	0.68	0.71	0.64	0.73	0.79	1
Total Score	0.48	0.42	0.07	0.58	0.74	0.65

FORM C	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	-0.13	1				
Supporting	0.34	0.19	1			
Advocating	0.50	0.50	0.62	1		
Communicating	0.36	0.26	0.33	0.57	1	
Monitoring	0.40	0.31	0.64	0.64	0.39	1
Total Score	-0.17	-0.40	0.30	0.53	0.07	0.38

To augment the traditional analysis, we conducted a generalizability analysis of the data. We use generalizability theory within a hierarchically nested linear model framework to partition variability among VAL-ED item responses into component sources of information and error. Specifically, we partitioned variability into the following five components: 1) overall effectiveness, 2) relative effectiveness for core components, 3) relative effectiveness for key processes, 4) individual teacher rater effects, and 5) item-specific variance. This partitioning of item variance into five components allows us to test the hypothesis that variability in core component and/or key process subscales may or may not be distinguished from variability in the overall score. Thus, the G-Theory analysis addresses a slightly different question than the difference score reliabilities. Instead of asking whether the subscales can be distinguished from each other, the G-Theory model addresses the equally important question of whether the subscales can be distinguished from the overall score.

The mathematical form of the G-Theory HLM model for the teacher response data is:

$$Y_{pckti} = \theta_p + \gamma_{pc} + \lambda_{pk} + \eta_{pt} + \varepsilon_{pckti}$$

Where Y is the response to item *i* from teacher *t* on key process *k* and core component *c* for principal *p*. The term θ_p represents the overall effectiveness of principal *p*, γ_{pc} is the relative effectiveness of principal *p* on core component *c* (expressed as a deviation from the principal's overall effectiveness), λ_{pk} is the relative effectiveness of principal *p* on key process *k* (expressed as a deviation from the principal's overall effectiveness), η_{pt} is the overall rater effect of teacher *t* for principal *p*, and ε_{pckti} is the residual item error term. The model applied to VAL-ED principal and supervisor response data is identical, minus the teacher rater effect component.

The models were estimated using restricted maximum likelihood (REML) as implemented in PROC MIXED in SAS version 9.1.3. Overall scores, subscale scores, and teacher rater effects were

obtained by estimating each random effect and its associated standard error. Overall and subscale score G-Theory reliabilities were estimated via the following formula,

$$r = 1 - \frac{[median(SE)]^2}{\tau^2}$$

where SE is the standard error for a set of random effects (i.e., the median error variance), and τ^2 is the random effects variance component for that set of random effects (i.e., the total variance). Results are reported separately by respondent group and form.

Table 4.13 shows the G-theory variance decomposition for each of the three respondent groups and, within each of the respondent groups, for each of the two forms. As might be expected, a major proportion of the variance is connected with the residual (error) shown in the bottom row of the table and principal and teacher, as shown at the top of the table. For each respondent group and form, the first entry (to the left) is the estimate of the variance component and the second entry (to the right) is the P-value indicating whether or not that variance component is statistically greater than zero. Using the convention of testing significance at the .05 level, all P-values smaller than .05 indicate that the variance component is statistically greater than zero. The desired result is for each core component and each key process to have unique variance greater than zero. For example, under principal Form A, the unique variance for the High Standards core component is .010 and the P-value is .12. The conclusion is there is no variance for High Standards unique from the other core components or key processes.

Table 4.13 Generalizability Variance Partitioning												
	Principal				Supervisor				Teacher			
Form	A		C		A		C		A		C	
	unique variance	P-value	unique variance	P-value	unique variance	P-value	unique variance	P-value	unique variance	P-value	unique variance	P-value
Principal	0.216	0.00	0.243	0.00	0.507	0.00	0.358	0.00	0.172	0.00	0.141	0.00
Teacher									0.439	0.00	0.444	0.00
Core Components												
High Standards	0.010	0.12	0.051	0.00	0.003	0.26	0.032	0.00	0.003	0.00	0.019	0.00
Rigorous Curriculum	0.034	0.00	0.046	0.00	0.016	0.01	0.026	0.00	0.004	0.00	0.010	0.00
Quality Instruction	0.053	0.00	0.032	0.00	0.030	0.00	0.004	0.22	0.017	0.00	0.006	0.00
Culture of Learning	0.105	0.00	0.077	0.00	0.061	0.00	0.072	0.00	0.024	0.00	0.028	0.00
External Community	0.330	0.00	0.184	0.00	0.127	0.00	0.130	0.00	0.043	0.00	0.019	0.00
Performance Accountability	0.098	0.00	0.168	0.00	0.046	0.00	0.084	0.00	0.010	0.00	0.020	0.00
Key Processes												
Planning	0.014	0.03	0.009	0.06	0.008	0.04	0.003	0.26	0.002	0.00	0.002	0.00
Implementing	0.000	1.00	0.000	1.00	0.011	0.02	0.004	0.17	0.001	0.01	0.002	0.00
Supporting	0.072	0.00	0.013	0.03	0.016	0.00	0.025	0.00	0.027	0.01	0.011	0.00
Advocating	0.018	0.02	0.023	0.00	0.048	0.00	0.052	0.00	0.006	0.00	0.009	0.00
Communicating	0.023	0.01	0.025	0.00	0.000	1.00	0.019	0.00	0.007	0.00	0.011	0.00
Monitoring	0.015	0.03	0.059	0.00	0.000	1.00	0.038	0.00	0.005	0.00	0.016	0.00
Residual	0.378	0.00	0.350	0.00	0.242	0.00	0.242	0.00	0.362	0.00	0.387	0.00

In virtually all cases, the unique variances are significant (62 out of 72). Exceptions for key processes are Implementing for principals on Forms A and C and supervisors on Form C. Communicating and Monitoring were not significantly different from zero for supervisors for Form A, where the variance component could not be estimated and the best approximation is zero variance. For the key process of Planning, the variance component was not significantly greater than zero for principals on Form A or supervisors on Form C. In addition, for the core component of High Standards, the variance component was not significantly greater than zero for principals Form A (as already seen) or supervisors Form A. For Quality Instruction, the variance component as estimated on supervisor data for Form C was not statistically greater than zero. All the other variance components estimates for core components and key processes were significantly greater than zero, indicating that, despite the high co-linearity among core components and key processes, each has some unique variance. Generally speaking, the variance component estimates were larger for core components than key processes, indicating that core components are better distinguished one from another than are key processes. Within core components, for Culture of Learning and Professional Behavior and Connections to External Communities the variance components tended to be larger across forms and respondents. Further, the variance components for core components and key processes tended to be a bit larger for the principal data than for the supervisor data and a bit larger for the supervisor data than the teacher data. Many of the variance components are larger than 5% and all of those are statistically significant which seems both reliable and meaningful.

The variance component data can be used to estimate the reliability of differences. Because there are so many pairwise contrasts possible, reliabilities of the difference for each core component from total score and each key process from total score are reported in Table 4.14. In addition,

generalizability theory can be used to estimate reliabilities for total score, which can be compared to the internal consistency reliabilities reported above.

Table 4.14 Generalizability Reliability Estimates						
	Principal		Supervisor		Teacher	
	Form A	Form C	Form A	Form C	Form A	Form C
Total Score Reliability	0.92	0.92	0.97	0.96	0.88	0.86
Subscale Difference Score Reliability						
Core Components						
High Standards	0.17	0.51	0.09	0.51	0.40	0.81
Rigorous Curriculum	0.41	0.49	0.35	0.46	0.51	0.68
Quality Instruction	0.53	0.39	0.49	0.11	0.85	0.53
Culture of Learning	0.69	0.62	0.67	0.70	0.88	0.87
External Community	0.88	0.80	0.79	0.80	0.92	0.80
Performance Accountability	0.68	0.79	0.60	0.73	0.75	0.81
Key Processes						
Planning	0.26	0.20	0.23	0.08	0.47	0.38
Implementing	0.00	0.00	0.27	0.13	0.31	0.43
Supporting	0.64	0.25	0.37	0.48	0.92	0.79
Advocating	0.31	0.38	0.63	0.66	0.69	0.74
Communicating	0.36	0.40	0.00	0.41	0.74	0.80
Monitoring	0.27	0.61	0.00	0.57	0.62	0.84

Reliability of difference scores are notoriously low because of co-linearity among subscales (Cronbach, 1990; Feldt, 1995). That is the finding here, though the reliabilities are not as low as one might have expected, and certainly no lower than one finds, for example, among sub-scores on a student achievement test (e.g., within Mathematics, Algebra versus Geometry versus Measurement). To some extent, the relatively good reliabilities of difference scores may be a function of the high internal consistency reliabilities for each subscale. Still, as the pattern of variance components would suggest, some subscales cannot be reliably distinguished from total score and therefore, neither can they be distinguished from any of the other subscales.

The first thing to recognize in Table 4.14 is that the total score reliabilities are high, regardless of respondent group or form, though for principals and teachers the reliabilities are slightly lower than found when using Cronbach's Alpha. The reliability of total score for principal data was .92 regardless of form; for teachers, it was .88 for Form A and .86 for form C. In contrast, the reliability of total score based on supervisor data was .97 for Form A and .96 for Form C. Again, these reliabilities are uniformly high, showing once again that the total score is a reliable measure of principal effectiveness.

Within Table 4.14, there are substantial differences among reliabilities of differences as one would have inferred from the variance components results. Generally, reliability of difference scores were greater for core components than key processes. In fact, the reliability of contrasting Culture of Learning, External Communities, and Performance Accountability, each with total score, were uniformly excellent, ranging from a low of .60 to a high of .92. Again, these are reliabilities of difference scores and reliabilities of difference scores are notoriously low. The reliability of distinguishing Rigorous Curriculum from Total Score was uniformly good, ranging from a low of .35 to a high of .68 across respondent groups and forms. The reliability of distinguishing Quality of Instruction from Total Score was quite good as well, with the exception of a reliability of only .11 based on supervisor data for Form C. The reliability distinguishing High Standards from Total Score was more variable across respondent groups and forms, ranging from a high of .81 for Form C teacher data to a low of .09 for Form A supervisor data.

Turning to key processes, the two key processes most reliably distinguished from Total Score are Supporting and Advocating, with a reliability of differences for Supporting ranging from .25 to .92 and for Advocating from .31 to .74. For the other key processes, the pattern of reliability across respondent groups and forms was variable, with some reliability of differences quite high (as

high as .84), but some, as we learned from the estimate of variance components, were zero. Generally, Planning and Implementing are not well distinguished from Total Score while Communicating and Monitoring are sometimes well distinguished from total score and sometimes not.

Combined with the results from the traditional reliability of difference scores analysis, we take these reliability of difference score findings as support for our conceptual framework and support for reporting scores at the subscale level, as well as total score. Additional support for the conceptual framework follows from factor analysis and mean item differences among core components, key processes, and even individual cells in the six by six conceptual framework. Clearly, contrasts among core components are more reliable than contrasts among key processes.

Construct Validity & Factor Structure

It is clear that conceptual structure on which the VAL-ED was designed (see Figure 2.1 and Figure 2.2) is strong. This would be true regardless of the findings of the reliability of difference and G-theory analyses. Nevertheless, the reliability of the differences and G-theory results indicate that to some extent the core components can be distinguished one from another and each has unique variance greater than zero; the same is generally true for key processes, but with some exceptions and with the variance components slightly smaller. The reliabilities of the contrasts between subscales and total score for core components were quite strong; for key processes, the reliabilities were strong in two cases and variable for the other four. Here, the factor structure of the data set is investigated to evaluate construct validity of the VAL-ED.

To place the generalizability analyses above and the factor structure investigations which follow in context, the intercorrelations of core components and key processes are presented in Table 4.15 for principal data, Table 4.16 for supervisor data, and Table 4.17 for teacher data. As is seen in

each of those tables, intercorrelations among subscales are high for teacher data, still high but somewhat lower for supervisor data, and still high but quite a bit lower for principal data. As stated above, given these relatively large correlations among subscales, the reliability of the difference scores above is somewhat surprising and undoubtedly a function of the high internal consistency reliability of each subscale.

Table 4.15 Intercorrelations Along Core Components and Key Processes - Supervisor Data

Scale	Overall Effectiveness	Core Components	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Communities	Performance Accountability	Key Processes	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Overall Effectiveness	1.00		.96	.96	.96	.94	.87	.94		.97	.97	.97	.94	.97	.96
Core Components															
High Standards	.96		1.00	.92	.92	.86	.78	.89		.93	.93	.92	.88	.93	.93
Rigorous Curriculum	.96		.92	1.00	.92	.86	.78	.88		.93	.92	.92	.91	.92	.91
Quality Instruction	.96		.92	.92	1.00	.89	.80	.89		.94	.93	.95	.89	.93	.93
Culture of Learning	.94		.86	.86	.89	1.00	.82	.85		.90	.92	.92	.89	.93	.88
External Communities	.87		.78	.78	.80	.82	1.00	.79		.84	.86	.83	.83	.85	.81
Performance Accountability	.94		.89	.88	.89	.85	.79	1.00		.92	.92	.90	.86	.91	.92
Key Processes															
Planning	.97		.93	.93	.94	.90	.84	.92		1.00	.94	.93	.89	.93	.93
Implementing	.97		.93	.92	.93	.92	.86	.92		.94	1.00	.94	.89	.94	.92
Supporting	.97		.92	.92	.95	.92	.83	.90		.93	.94	1.00	.87	.95	.90
Advocating	.94		.88	.91	.89	.89	.83	.86		.89	.89	.87	1.00	.90	.87
Communicating	.97		.93	.92	.93	.93	.85	.91		.93	.94	.95	.90	1.00	.93
Monitoring	.96		.93	.91	.93	.88	.81	.92		.93	.92	.90	.87	.93	1.00

Table 4.16 Intercorrelations Along Core Components and Key Processes - Teacher Data

Scale	Overall Effectiveness	Core Components	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Communities	Performance Accountability	Key Processes	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Overall Effectiveness	1.00		.98	.97	.97	.97	.95	.98		.99	.99	.98	.97	.98	.98
Core Components															
High Standards	.98		1.00	.94	.93	.94	.91	.95		.97	.97	.95	.94	.97	.94
Rigorous Curriculum	.97		.94	1.00	.96	.91	.88	.94		.96	.95	.94	.95	.94	.95
Quality Instruction	.97		.93	.96	1.00	.93	.88	.95		.96	.96	.96	.94	.95	.95
Culture of Learning	.97		.94	.91	.93	1.00	.90	.94		.96	.96	.96	.94	.95	.94
External Communities	.95		.91	.88	.88	.90	1.00	.92		.93	.93	.90	.94	.93	.91
Performance Accountability	.98		.95	.94	.95	.94	.92	1.00		.96	.96	.95	.95	.97	.96
Key Processes															
Planning	.99		.97	.96	.96	.96	.93	.96		1.00	.97	.95	.96	.96	.96
Implementing	.99		.97	.95	.96	.96	.93	.96		.97	1.00	.97	.95	.96	.94
Supporting	.98		.95	.94	.96	.96	.90	.95		.95	.97	1.00	.93	.95	.94
Advocating	.97		.94	.95	.94	.94	.94	.95		.96	.95	.93	1.00	.95	.95
Communicating	.98		.97	.94	.95	.95	.93	.97		.96	.96	.95	.95	1.00	.95
Monitoring	.98		.94	.95	.95	.94	.91	.96		.96	.94	.94	.95	.95	1.00

Table 4.17 Generalizability Variance Partitioning

Form	Principal				Supervisor				Teacher			
	A		C		A		C		A		C	
	unique variance	P-value	unique variance	P-value	unique variance	P-value	unique variance	P-value	unique variance	P-value	unique variance	P-value
Principal	0.216	0.00	0.243	0.00	0.507	0.00	0.358	0.00	0.172	0.00	0.141	0.00
Teacher									0.439	0.00	0.444	0.00
Core Components												
High Standards	0.010	0.12	0.051	0.00	0.003	0.26	0.032	0.00	0.003	0.00	0.019	0.00
Rigorous Curriculum	0.034	0.00	0.046	0.00	0.016	0.01	0.026	0.00	0.004	0.00	0.010	0.00
Quality Instruction	0.053	0.00	0.032	0.00	0.030	0.00	0.004	0.22	0.017	0.00	0.006	0.00
Culture of Learning	0.105	0.00	0.077	0.00	0.061	0.00	0.072	0.00	0.024	0.00	0.028	0.00
External Community	0.330	0.00	0.184	0.00	0.127	0.00	0.130	0.00	0.043	0.00	0.019	0.00
Performance Accountability	0.098	0.00	0.168	0.00	0.046	0.00	0.084	0.00	0.010	0.00	0.020	0.00
Key Processes												
Planning	0.014	0.03	0.009	0.06	0.008	0.04	0.003	0.26	0.002	0.00	0.002	0.00
Implementing	0.000	1.00	0.000	1.00	0.011	0.02	0.004	0.17	0.001	0.01	0.002	0.00
Supporting	0.072	0.00	0.013	0.03	0.016	0.00	0.025	0.00	0.027	0.01	0.011	0.00
Advocating	0.018	0.02	0.023	0.00	0.048	0.00	0.052	0.00	0.006	0.00	0.009	0.00
Communicating	0.023	0.01	0.025	0.00	0.000	1.00	0.019	0.00	0.007	0.00	0.011	0.00
Monitoring	0.015	0.03	0.059	0.00	0.000	1.00	0.038	0.00	0.005	0.00	0.016	0.00
Residual	0.378	0.00	0.350	0.00	0.242	0.00	0.242	0.00	0.362	0.00	0.387	0.00

Clearly, these results show that field trial data indicate high correlation among the VAL-ED subscales. Whether these high correlations will remain when the VAL-ED is used in practice is not

clear. To the extent that there is not correlation, however, it is useful to consider the extent to which the non-correlation supports the conceptual framework described in Chapter 2. Furthermore, because we are committed to reporting results at the level of the core component and key process subscale, it is useful to examine whether the data support the existence of unique factors attributable to the subscales.

The field trial data allow us to investigate construct validity in several ways. Here, we present results from three distinct analyses designed to investigate the construct validity of the VAL-ED. The first is an exploratory factor analysis, designed to examine whether item responses tend to cluster in ways that indicate the presence or absence of core components or key processes. Next, we test a confirmatory factor analysis model that tests whether the data fit the framework off which the instrument was created. Finally, we use ANOVAs to test the extent to which means for core components, key processes, and individual cells are different, one from another. Each statistical analysis is designed to provide additional evidence as to the overall fit of the data to the conceptual framework.

Exploratory Factor Analysis

To conduct an exploratory factor analysis, we used a PROMAX approach to oblique rotations. For Form A, eight eigenvalues had values equal to or greater than 1.0 and for Form C, there were nine. Nevertheless, we tried a six-, an eight-, and a twelve-factor solution for both Form A and Form C. The eight-factor solution was driven by the rule of thumb that there are as many factors as there are eigenvalues equal to or greater than 1.0. Of course, we're looking for the same factor solution for each form, and thus, we decided on an eight-factor solution for each form. The six-factor solution was predicated on the notion that perhaps a factor structure supporting core components or a factor structure supporting key processes might emerge and there are six of each.

The twelve-factor solution was driven by the hypothesis that perhaps there would be six factors for the core components and an additional six factors for the key processes.

For all three solutions and for each form, there was strong evidence in support of a factor for Connections to External Community and a factor for Performance Accountability. The six-factor solution had difficulty obtaining simple structure (i.e. the goal of having a factor matrix once rotated that has one large factor loading per row (item) with the rest of the factor loadings for that item near zero). For the six-factor solution, too many of the variables were complex, having more than one large loading across factors. For the twelve-factor solutions, the last three factors were generally non-interpretable. Thus, we focused on the eight-factor solution for Form A and for Form C.

Table 4.18 contains the factor matrix for Form A and Table 4.19 the factor matrix for Form C. As is always the case in exploratory factor analysis, there is some art in naming factors. The more easily interpreted factor matrix was for Form C.

Table 4.18 Factor Matrix - Form A

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
High Standards								
1	0.07	0.06	0.24	0.26	-0.06	0.48	0.15	0.18
2	0.05	0.18	0.31	0.25	-0.02	0.22	0.14	0.31
3	0.28	0.09	-0.07	0.31	0.17	0.28	0.07	0.19
4	0.27	-0.03	0.26	0.14	0.17	0.29	0.12	0.30
5	-0.02	0.14	-0.07	0.13	0.22	0.44	0.26	0.24
6	0.44	0.11	-0.03	0.09	0.07	0.20	0.06	0.41
7	-0.02	0.01	0.01	0.06	0.12	0.21	0.60	0.22
8	0.02	0.03	-0.03	-0.02	0.18	0.30	0.61	0.17
9	0.17	0.11	0.27	0.21	0.04	0.37	0.15	0.16
10	0.21	0.26	0.16	-0.10	0.31	0.46	-0.06	0.05
11	0.21	0.18	0.31	0.17	0.06	0.21	0.11	0.19
12	0.25	0.04	0.41	0.34	0.04	0.09	0.14	-0.06
Rigorous Curriculum								
13	0.15	0.17	0.09	0.19	-0.11	0.42	0.22	0.20

14	0.10	0.13	0.13	0.01	0.09	0.03	0.62	0.09
15	0.00	0.21	0.01	0.26	-0.04	0.33	0.16	0.39
16	0.08	0.14	0.08	0.26	-0.09	0.38	0.17	0.29
17	0.57	0.08	0.12	0.05	0.06	0.20	0.04	0.09
18	0.42	0.07	0.39	0.12	-0.07	0.00	0.16	0.23
19	0.37	0.04	0.13	0.03	0.18	0.43	0.29	-0.20
20	0.25	0.15	0.20	0.18	-0.03	0.31	0.36	-0.02
21	0.00	0.23	0.54	-0.01	-0.12	0.00	0.39	0.06
22	0.09	-0.08	0.17	0.32	0.16	0.03	0.57	-0.07
23	0.08	0.23	0.16	0.29	-0.06	0.28	0.27	0.13
24	0.03	0.09	0.14	0.34	0.12	0.42	0.09	0.19
Quality Instruction								
25	0.28	0.07	0.00	0.21	0.18	-0.06	0.55	0.01
26	0.28	0.09	0.25	0.10	0.03	0.01	0.07	0.48
27	0.35	-0.12	0.32	0.19	0.04	0.18	0.25	0.25
28	0.31	0.03	0.16	-0.08	0.16	0.22	0.26	0.16
29	0.59	0.05	0.18	0.05	0.13	-0.01	0.13	0.18
30	0.64	0.02	0.06	-0.01	0.05	0.17	0.09	0.23
31	0.24	0.09	0.17	-0.10	0.23	0.31	0.31	0.14
32	-0.01	0.22	0.16	0.14	0.14	0.18	0.12	0.18
33	0.09	0.11	0.66	0.07	0.00	0.26	-0.02	-0.06
34	0.15	0.17	0.25	0.15	0.25	0.25	0.11	0.20
35	0.08	-0.06	0.59	0.13	0.29	0.02	0.11	0.19
36	0.11	-0.03	0.54	0.08	0.21	0.20	0.04	0.24
Culture of Learning								
37	0.45	-0.11	0.04	0.23	0.09	-0.01	0.08	0.42
38	0.49	-0.05	0.10	0.04	0.22	0.08	0.05	0.44
39	0.36	0.04	-0.01	0.00	0.41	0.21	0.06	0.32
40	0.19	0.04	0.08	0.20	0.20	0.38	-0.01	0.35
41	0.39	0.12	0.11	0.17	0.24	0.14	0.10	0.22
42	0.68	0.06	0.02	0.20	0.05	0.02	0.04	0.09
43	0.28	-0.14	-0.03	-0.01	0.42	0.29	0.39	0.11
44	0.06	0.18	-0.02	-0.06	0.37	0.09	0.04	0.58
45	0.07	0.03	0.04	0.04	0.61	0.27	0.00	0.24
46	0.05	0.15	0.37	0.19	0.16	-0.04	-0.02	0.45
47	-0.12	0.19	0.20	0.12	0.49	-0.08	0.26	0.22
48	0.10	0.15	-0.03	0.07	0.53	0.19	0.08	0.17
Connection to External Communities								

49	0.09	0.34	0.17	0.18	0.14	0.27	0.03	0.22
50	0.03	0.62	0.07	0.07	0.17	0.24	-0.01	0.08
51	0.01	0.76	0.05	-0.03	0.15	-0.06	0.10	0.21
52	-0.02	0.72	0.08	0.14	0.01	0.10	0.11	0.09
53	0.17	0.57	-0.02	0.30	0.14	0.18	0.09	-0.07
54	0.24	0.46	-0.02	0.15	0.20	0.23	0.02	0.14
55	0.05	0.42	0.08	-0.11	0.53	0.03	0.14	0.06
56	-0.12	0.53	0.09	0.24	0.31	0.12	0.00	0.02
57	0.03	0.16	0.06	0.09	0.64	-0.04	0.17	0.14
58	0.25	0.17	-0.05	0.13	0.65	-0.04	0.14	-0.06
59	0.01	0.37	0.00	0.46	0.38	0.10	0.05	-0.08
60	0.16	0.50	0.02	0.17	0.29	0.12	-0.17	0.11
Performance Accountability								
61	0.07	0.17	0.25	0.39	-0.08	0.20	0.13	0.33
62	0.12	0.24	0.12	0.58	-0.10	0.10	0.14	0.14
63	0.11	0.15	0.12	0.60	0.12	0.00	0.04	0.26
64	0.18	0.06	-0.01	0.39	0.08	0.21	0.16	0.29
65	-0.02	-0.03	0.36	0.43	0.02	0.25	0.20	0.16
66	0.01	0.02	0.49	0.33	0.04	0.13	-0.05	0.34
67	0.01	0.06	0.15	0.47	0.22	0.31	0.00	0.11
68	0.12	-0.03	0.06	0.32	0.14	0.32	0.18	0.31
69	0.06	0.09	0.19	0.31	0.31	0.38	-0.12	0.03
70	0.00	0.17	0.35	0.37	-0.08	0.27	-0.01	0.28
71	-0.12	0.13	0.25	0.35	-0.04	0.47	0.17	0.05
72	0.16	0.09	0.31	0.42	0.13	0.10	-0.02	0.22

Table 4.19 Factor Matrix - Form C

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
High Standards								
1	0.18	0.45	0.41	0.12	-0.04	0.00	0.08	0.15
2	0.37	0.45	0.36	-0.05	0.05	0.01	0.03	0.05
3	0.25	0.37	0.33	0.04	0.08	0.08	0.12	0.17
4	0.10	0.15	0.11	-0.03	0.15	-0.02	0.01	0.68
5	0.31	0.25	0.23	0.05	-0.01	0.04	0.09	0.37
6	0.32	0.19	0.17	0.13	-0.09	0.11	0.09	0.40
7	0.09	0.21	-0.07	0.75	0.10	-0.09	-0.05	0.19
8	0.18	0.40	0.27	0.28	-0.07	0.07	-0.06	0.12

9	0.16	0.41	0.29	0.14	0.02	0.11	0.00	0.23
10	0.17	0.43	0.19	0.09	-0.04	0.06	0.03	0.29
11	0.10	0.10	0.47	0.06	0.03	0.31	-0.07	0.21
12	-0.12	0.32	0.62	0.13	-0.05	0.19	0.04	0.13
Rigorous Curriculum								
13	0.06	0.15	0.14	0.65	-0.01	0.12	0.08	0.08
14	0.14	0.36	0.36	0.14	-0.06	0.16	0.21	0.01
15	0.00	0.15	0.09	0.57	-0.10	0.20	0.11	0.20
16	-0.05	0.47	0.28	0.10	0.04	0.08	0.31	0.16
17	0.15	0.60	-0.04	0.05	0.19	0.19	0.02	0.14
18	0.02	0.54	-0.05	0.15	0.11	0.24	0.03	0.16
19	0.22	0.40	0.15	0.36	-0.03	0.16	0.12	0.00
20	0.11	0.29	0.01	0.69	-0.04	0.07	0.04	0.05
21	0.03	0.31	0.14	0.02	-0.02	0.52	0.20	0.02
22	0.22	0.15	-0.13	0.11	-0.04	0.25	0.34	0.34
23	-0.13	0.30	0.46	0.10	0.15	0.28	-0.01	0.15
24	0.20	0.12	0.11	0.06	0.15	0.63	-0.01	0.02
Quality Instruction								
25	0.16	0.10	0.06	0.09	0.02	0.12	0.21	0.55
26	0.14	0.58	-0.01	0.02	0.16	0.21	0.10	0.17
27	0.00	0.40	0.21	0.18	0.10	0.20	0.16	0.13
28	-0.12	0.01	-0.11	0.22	0.13	0.35	0.24	0.40
29	-0.07	0.39	-0.01	0.06	0.25	-0.10	0.50	0.26
30	-0.01	0.38	0.10	0.11	0.13	-0.13	0.51	0.27
31	0.08	0.19	0.01	-0.06	0.56	-0.06	0.44	-0.08
32	0.12	0.24	0.21	0.38	0.24	0.03	0.19	-0.22
33	0.14	0.39	0.20	0.00	-0.06	0.45	0.14	0.05
34	0.29	0.06	0.02	0.02	-0.01	0.59	0.07	0.23
35	-0.11	0.24	0.43	-0.05	0.10	0.52	0.05	0.10
36	0.10	0.10	0.10	0.09	0.12	0.71	0.00	-0.03
Culture of Learning								
37	0.44	0.32	-0.02	-0.09	0.03	0.32	0.30	-0.01
38	0.47	0.27	0.04	0.04	0.11	0.18	0.18	0.03
39	0.42	0.34	0.17	0.03	0.02	0.05	0.14	0.21
40	0.50	-0.02	0.27	0.04	0.06	-0.19	0.10	0.40
41	0.55	0.20	-0.08	0.00	0.01	0.19	0.00	0.37
42	0.57	0.13	0.10	0.06	0.14	0.04	0.07	0.25
43	0.46	-0.01	0.05	0.41	0.01	0.16	0.02	0.08

44	0.43	0.05	-0.08	0.38	0.23	0.14	0.08	0.02
45	0.51	0.07	0.05	0.26	0.24	-0.01	0.10	0.07
46	0.52	0.16	0.14	0.18	0.04	0.13	0.06	0.16
47	0.53	0.03	0.02	0.13	0.14	0.12	0.14	0.26
48	0.50	0.09	0.02	0.23	0.06	0.25	0.13	0.11
Connection to External Communities								
49	0.14	0.26	-0.07	-0.06	0.65	0.07	0.13	0.04
50	0.07	0.11	0.13	0.03	0.76	0.09	-0.06	0.11
51	-0.06	0.13	0.15	0.12	0.74	0.02	-0.01	0.11
52	0.45	0.01	-0.05	0.07	0.53	-0.02	-0.01	0.16
53	0.13	0.06	-0.08	0.23	0.42	0.10	0.32	-0.07
54	0.27	-0.08	-0.02	0.32	0.38	0.12	0.15	0.13
55	0.01	-0.01	0.22	0.49	0.31	0.03	0.08	0.10
56	0.08	-0.11	0.22	0.55	0.21	-0.02	0.15	0.05
57	0.30	-0.08	0.15	0.31	0.35	-0.02	0.05	0.25
58	0.05	0.06	0.40	0.00	0.41	0.21	0.08	0.01
59	0.07	-0.05	0.03	-0.09	0.43	0.25	0.51	0.02
60	-0.03	-0.25	0.02	0.07	0.46	0.31	0.25	0.24
Performance Accountability								
61	0.16	0.08	0.64	-0.03	-0.01	0.00	0.23	0.06
62	0.41	0.11	0.16	-0.05	0.14	0.23	0.25	0.11
63	0.25	-0.06	0.48	0.17	0.17	0.00	0.12	0.20
64	0.14	0.25	0.42	0.04	0.06	0.18	0.20	0.21
65	0.09	-0.04	0.40	0.17	-0.09	0.03	0.58	0.13
66	0.18	-0.03	0.38	0.00	-0.04	0.14	0.56	0.17
67	0.11	-0.09	0.15	0.24	0.19	0.31	0.25	0.11
68	0.32	0.17	0.24	0.11	0.16	0.18	0.14	0.09
69	0.07	-0.12	0.49	0.23	0.33	0.04	0.26	-0.07
70	0.08	0.15	0.45	0.18	0.05	0.22	0.24	0.06
71	0.04	0.03	0.27	0.11	0.10	0.44	0.23	0.02
72	0.00	0.04	0.54	0.00	0.11	0.12	0.38	0.04

For Form C, Factor 1 might be labeled Culture of Learning and Professional Behavior. All of the factor loadings for the twelve Culture of Learning and Professional Behavior items on Factor 1 were .40 or larger. For the remainder of the 72 items, only two had factor loadings on Factor 1 of

.40 or greater: Planning for Performance Accountability and Implementing Connections to External Communities.

Factor 2 might be labeled Combination of High Standards for Student Learning and Rigorous Curriculum. Here, the clarity of support for the label is not quite as great as was Culture of Learning and Professional Behavior for Factor 1. Nevertheless, of the 24 items measuring High Standards and Rigorous Curriculum, 16 had factor loadings of .25 or greater on Factor 2 and all of the factor loadings were .10 or greater. There were, however, five of the twelve factor loadings for Quality Instruction also with .25 or greater factor loadings on Factor 2.

Factor 3 might be labeled Performance Accountability. Ten of the twelve items measuring Performance Accountability have factor loadings of .20 or greater on Factor 3 and all of them had factor loadings of .15 or greater. There were, however, nine other items that had .30 or greater factor loadings on Factor 3. Four of them were for Monitoring, three of them were for Planning, and five were for High Standards for Student Learning.

Factor 4 might be labeled Advocating. Of the twelve items measuring Advocating, eleven of them had factor loadings of .10 or greater for factor 4 and ten of them had factor loadings greater than .20. There were a couple other items, however, that had large factor loadings on Factor 4, one for Planning Rigorous Curriculum, and one for Implementing Rigorous Curriculum.

Factor 5 is quite clearly Connections to External Communities. All twelve items measuring External Communities had factor loadings of .20 or higher on Factor 5. Just two of the other 60 items had high loadings on Factor 5, one in Advocating Quality Instruction and one in Communicating Performance Accountability.

Factor 6 might be labeled Communicating and Monitoring Rigorous Curriculum and Quality Instruction. All eight of these items had factor loadings at .25 or higher on Factor 6.

Factor 7 might be labeled Supporting Quality Instruction and Performance Accountability. All four of the items measuring Supporting Quality Instruction and Performance Accountability had factor loadings of .50 or higher on Factor 7.

There was no clear interpretation of Factor 8.

Just as there were factors for Connections to External Communities and Performance Accountability for Form C, there were also factors for those two core components in the eight-factor solution for Form A. Factor 2 for the Form A solution might be labeled Connections to External Communities. Of the twelve items measuring Connections to External Communities, ten of them had factor loadings of .30 or greater and all twelve had factor loadings of .15 or greater. Similarly, Factor 4 might be labeled Performance Accountability. Of the twelve items measuring Performance Accountability, all twelve had factor loadings of .30 or higher on Factor 4 for Form A. Factor 1 might be labeled Support for High Standards, Rigorous Curriculum, Quality Instruction, and Culture of Learning and Professional Behavior (Factor 7 was labeled Supporting for Form C but that factor was not quite as strong, indicating a factor for Supporting only Quality Instruction and Performance Accountability). Of the eight items measuring Supporting High Standards, Rigorous Curriculum, Quality Instruction, and Culture of Learning and Professional Behavior, seven had loadings of .4 or higher on Factor 1. There were, however, some items measuring Planning and Implementing that had strong, positive factor loadings on Factor 1 as well, especially for Quality Instruction and Culture of Learning and Professional Behavior.

Factor 5 for Form A might be labeled Culture of Learning and Professional Behavior and Connections to External Communities, in combination with Advocating, Communicating, and Monitoring. Of the twelve items measuring the combination of those core components and key processes, all but one had factor loadings on Factor 5 of .29 or higher.

Factor 3 might be labeled Monitoring for High Standards, Quality Instruction, and Performance Accountability. Of the six items measuring that intersection of key process and core components, all had factor loadings of .25 or higher. For Factor 3, however, some of the items measuring Communicating also had reasonably high factor loadings on Factor 3, as seen for Rigorous Curriculum, Quality Instruction, and Culture of Learning.

For Form A, Factor 6 might be labeled High Standards for Student Learning. Of the twelve items, all but one had factor loadings of .20 or higher on Factor 6. Factor 6 also had a few other items with high loadings, but they were scattered across core components and key processes.

Factor 7 might be labeled Advocating for High Standards and Advocating and Communicating Rigorous Curriculum. Of the six items measuring those intersections of core components and key processes, all had factor loadings at .29 and higher. Other items with high factor loadings on Factor 7 were Planning for Rigorous Curriculum and Monitoring for Rigorous Curriculum.

Like for Form C, Factor 8 was not readily interpretable.

Thus, the exploratory factor analysis provided some support for the conceptual framework of the VAL-ED. Further, the results of the factor analysis for Form A replicated to a considerable extent the findings from the factor analysis for Form C. There was considerable support for the core components of Performance Accountability and Connections to External Communities and some support for the core component of Culture of Learning and Professional Behavior. There was also support for the key process of Supporting and the key process of Advocating.

Confirmatory Factor Analysis

To further test the fit of the data to the conceptual framework, a confirmatory factor analysis was completed on the 72 items for Form A and the 72 items for Form C where the data were

aggregated across the three respondent groups. Again, aggregate data were calculated by first forming the average across teachers within a school and then weighting the teacher average equally with the supervisor and the principal. The data set for confirmatory factor analysis was thus 72 items by 214 schools. Four schools of the 218 schools with complete data across respondent groups had inconsistent forms (e.g. the teachers filled out Form A and supervisor Form C). These confirmatory factor analyses are parallel to the confirmatory factor analysis done on the pilot data.

The hierarchical factor analytic model had four levels. The first level was for the 72 individual items, which were endogenous to the latent factors for the 36 cells representing six core components crossed with six key processes at the second level. At the third level were latent factors for the six core components or key processes. At the fourth level was a single latent trait representing overall principal leadership (i.e. the total score). Because each item contributed to both a core component and a key process, the factor analytic model was split into two separate analyses, one featuring core components and the other featuring key processes. To gauge agreement between the two models, factor scores for the overall leadership score were produced for both models and the correlation between them was estimated. Each CFA models was fit using PROC CALIS as implemented in SAS 9.1.3.

Again, results from the confirmatory factor analysis reveal that both the core components and the key processes models fit the data very well, having goodness of fit indices (for both the GFI and the Adjusted GFI) of .99 for both core components and key processes analyses for Form A and .98 for both core components and key processes for Form C. Root mean square error was .02 for form A and .01 for form C. After adjusting for model complexity, the parsimonious goodness of fit indices were also very high and equal to .93 for core components both Form A and C and for key processes Form A and .92 for key processes Form C.

Item loadings for the core component solutions ranged from a low of .63 to a high of .98 for Form C and from a low of .64 to a high of .95 for Form A when investigating core components. When investigating key processes, the item factor loadings ranged from a low of .64 to a high of .95 for Form A and from a low of .62 to a high of .95 for Form C.

The level 2 factor loadings were also large in both sets of analyses and for both forms. When investigating core components, the factor loadings for key processes ranged from a low of .85 to a high of 1.0 for Form A and a low of .78 to a high of 1.0 for Form C. Similarly, when featuring key processes, the level 2 factor loadings for core components ranged from a low of .68 to a high of 1.0 for Form A and a low of .72 to a high of 1.0 for Form C.

Finally, the level 1 factor loadings were all large regardless of form and whether core components or key processes were being investigated. For core components, the lowest factor loading for Form A was .86 for External Communities and the highest was 1.0 for High Standards. For Form C, the lowest was .83 for External Communities and the highest was .99 for Quality Instruction. When investigating key processes for Form A, all of the factor loadings were .99 or higher. For Form C, all of the factor loadings were .95 or higher. The agreement between the CFA models of Core Components and Key Processes was consistently high, with a .99 correlation between overall leadership factor scores from the two models for both forms.

Despite the high co-linearity among core components and key processes in the data set and despite the complex nature of the data due to core components being crossed with key processes, the confirmatory factor analysis results are supportive of our conceptual framework.

Mean Differences in the Conceptual Framework

Thus far, we have investigated support for our conceptual framework through generalizability theory and the reliability of differences among the subscales, through confirmatory

factor analysis and through exploratory factor analysis. Yet another way to investigate the validity of the conceptual framework as realized in the national field trial is by asking the extent to which the means for core components differ one from another, the means for key processes differ one from another, and the means for the 36 cells differ one from another. In many data sets, comparing means is not a way to explore the structure of the data because variables are in different metrics. Here, the data are all in the five-point mean item response scale, making mean comparisons useful.

A two-dimensional analysis of variance (core components by key processes) was completed on the data aggregated across respondent groups for Form A; the results are shown in Table 4.20. The aggregation took place first at the item level. School by school, the mean item response for teachers was calculated. Then school by school, the mean of the teacher mean, the principal's and the supervisor's effectiveness ratings item by item was calculated. Thus, the respondent groups were weighted equally just as is done in reporting principal effectiveness as measured by the VAL-ED in the instrument's reporting forms. The mean square for core components was largest, 21.17, followed by the mean square for key processes, 6.19, followed by the interaction of core components by key processes, 2.91. All three were statistically significant with p-values less than .001.

Table 4.20. *Analysis of Variance for Core Components and Key Processes, Form A*

Source	*SS*	*df*	*MS*	*F*
Core Components (C)	105.83	5	21.17	81.13***
Key Processes (K)	30.97	5	6.19	24.04***
C x K	72.79	25	2.91	11.30***
Error	1846.13	7164	.26	

Note. N = 7200. *** $p < .001$

The mean for High Standards was 3.64, for Rigorous Curriculum, 3.58, for Quality Instruction, 3.74, for Culture of Learning, 3.74, for External Communities, 3.40, and Performance Accountability, 3.53. In terms of effect sizes, the differences range from .88 standard deviations (Quality Instruction or Culture of Learning compared with External Communities) to 0 (Quality Instruction compared with Culture of Learning), with a median effect size of .42 standard deviations. Thus, Quality of Instruction and Culture of Learning got nearly identical principal effectiveness mean ratings, while Connections to External Communities got the lowest rating. Tukey post-hoc pairwise contrasts were used to compare each core component to each other core component to see where the differences are statistically significant in mean principal effectiveness. Overall, there were only two instances where core components were not significantly different from one another. Rigorous Curriculum and Performance Accountability were not significantly different from each other; neither were Quality Instruction and Culture of Learning and Professional Behavior. All other pairwise comparisons showed significant differences.. Thus, the means for core components are significantly different among themselves with just a few exceptions, adding support for the core component dimension of the conceptual framework.

As for key processes, the mean for Planning was 3.58, Implementing, 3.58, Supporting, 3.75, Advocating, 3.56, Communicating, 3.61, and Monitoring, 3.56. These differences represent a maximum effect size of .53 standard deviations (Advocating compared with Supporting), a minimum effect size of 0 (Monitoring compared with Advocating or Planning compared with Implementing), and a median effect size of .08 standard deviations. Thus, across the entire sample and all three respondent groups, the key process of Supporting was judged to be most effectively accomplished by principals and the key process of Advocating was least effectively accomplished. Again using Tukey post-hoc pairwise comparisons and a .05 probability of type 1 error, the key

process of Supporting was significantly different from all other key processes. The other key processes were not significantly different one from another in their means on principal effectiveness.

Because the core components-by-key processes interaction was significant, Tukey post-hoc pairwise comparisons were used to contrast all of the pairs of cells, one from another. Again, at the statistical level of .05, even at the cell level where each cell is represented by only two of the 72 items on the VAL-ED, many cells were significantly different from other cells. In the 36 cells, 630 pairwise comparisons are possible. Of those, 276, or 44%, were significant at the .05 level. Had there been no real differences among the cell means, we would have expected only 5% significant by choice, only 32, not 276 as found. Of those 276 significant results, 111 involved contrasting a cell involving Supporting to another cell. Also, cells involving Connections to External Communities tended to have a high number of significant differences from other cells. There were 113 such statistically significant differences, or 43% of the 276 statistically significant differences found. There was evidence of the significant interaction as well. For example, many of the contrasts at the cell level between High Standards for Student Learning and Performance Accountability were not significantly different one from another. Nevertheless, Monitoring for High Standards was significantly different from five of the six key processes with Performance Accountability, the one exception being Supporting Performance Accountability. Another example is contrasts between Implementing for Quality Instruction and each of the six key processes. Yet another example is Performance Accountability. The one exception again that wasn't significant was for Supporting Performance Accountability. Yet another example is for the contrast of Monitoring Rigorous Curriculum paired with each of the key processes for Culture of Learning and Professional Behavior. The one contrast that wasn't significant was between Monitoring for Rigorous Curriculum and Monitoring for Culture of Learning.

Cell means ranged from a mean item response on the effectiveness scale of 3.99 for Supporting Quality Instruction to a low of 3.25 for Advocating Connections to External Communities. Table 4.21 presents cell means on the aggregate variable across respondent groups.

Table 4.21 Cell Mean Scores on the Aggregate Variable for Form A

Key Process	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Core Components						
High Standards	3.59	3.65	3.73	3.51	3.60	3.78
Rigorous Curriculum	3.47	3.43	3.89	3.63	3.58	3.48
Quality Instruction	3.68	3.78	3.99	3.68	3.61	3.71
Culture of Learning	3.83	3.87	3.86	3.75	3.70	3.43
Connections to External Communities	3.42	3.29	3.40	3.25	3.63	3.39
Performance Accountability	3.49	3.44	3.60	3.52	3.56	3.55

The results for Form C, shown in Table 4.22 were similar to the results for Form A. Again, mean effects for core components and key processes were significant as was the interaction. The mean square for core components was the largest, 20.32, followed by the mean square for key processes, 4.35, and the mean square for the interaction the smallest, 2.52. Again, the Tukey pairwise comparisons found that most of the core components were significant one from another. The only exceptions were that High Standards was not significantly different from Culture of Learning and Professional Behavior and Connections to External Communities was not significantly different from Performance Accountability. For key processes, again, as for Form A, Supporting was significantly different from most other key processes, the one exception being that it was not significantly different from Communicating. In addition, Communicating was significantly different from Planning, Implementing, Advocating, and Monitoring. The similarity of findings between Form A and C adds yet further support for the conclusion that the two forms are parallel.

Table 4.22. *Analysis of Variance for Core Components and Key Processes, Form C*				
Source	*SS*	*df*	*MS*	*F*
Core Components (C)	101.62	5	20.32	97.37***
Key Processes (K)	21.74	5	4.35	20.87***
C x K	63.01	25	2.52	12.08***
Error	1705.61	8172	.21	
Note. N = 7200. *** $p < .001$				

For Form C, cell means ranged from a high of 3.9 for Implementing Culture of Learning to a low of 3.36 for Monitoring Connections to External Communities. Of the 630 pairwise comparisons among the 36 cells, 296 were significant, or 47%, many more than could be expected by chance. Again, a great many of those significantly different cell means were accounted for by contrasting cells involving Supporting with other cells. Similarly, a great many of the differences were accounted for by contrasting Connections to External Communities and also Performance Accountability cells with other cells. Again, there was evidence of the significant interaction. For example, Planning for Rigorous Curriculum was significantly different from each of the key processes in connection with High Standards for Student Learning.

Similar analyses of variance can and were run by form and respondent group. For each respondent group in both forms, the core components' mean effect was statistically significant, the key process mean effect was statistically significant and the core components by key process interaction was statistically significant. These mean differences add further support to the validity of our conceptual framework and our VAL-ED instrument's ability to assess that conceptual framework.

Correlations among Response Groups

For 218 schools in the national field trial, data are available from each of the three respondent groups: principals, supervisors, and teachers. Just as we investigated the correlations among subscales within respondent groups, it is possible to investigate the correlations among respondent groups within subscale and total score. Average, minimum, and maximum correlations are presented in Table 4.23. For total score, the correlation between principal and supervisor was .13 and between principal and teacher .27. The correlation between supervisor and teacher was .18. These correlations are modest, indicating that teachers and principals have 7% of their variance on total score in common while teachers and supervisors have only 3% and principals and supervisors only 2% of their variance in common.

Table 4.23. *Total Score and Subscale Correlations*

		Core Components			Key Processes		
	Total Score	Average	Minimum	Maximum	Average	Minimum	Maximum
Teachers & Principals	.27	.26	.24	.29	.27	.23	.30
Supervisors & Principals	.13	.13	.02	.20	.13	.05	.18
Teachers & Supervisors	.18	.19	.12	.25	.17	.11	.21

It is also possible to investigate the extent to which core components are assessed the same by the three respondent groups. For core components, the average correlation across the six core components is again highest between teacher and principal, .26, next highest between teacher and supervisor, .19, and lowest between principal and supervisor, .13. For key processes, the pattern is once again the same. Principal and teacher are on average correlated the highest, .27, followed by supervisors and teachers, .17, and .13 for principals and supervisors. Looking below the averages,

no interesting patterns emerge either for core components or key processes. The correlation between principal and teacher respondents was uniformly relatively high, ranging from a low of .24 to a high of .29 for core components and from a low of .23 to a high of .30 for key processes. Similarly, the correlations between principal and supervisor data are uniformly low for both core components and key processes and the correlations between supervisors and teachers were uniformly in between the other two data sets, with the possible exception of High Standards, where the correlation between teachers and supervisors was .25.

These results for correlations among respondent groups indicate that the 360 degree approach to principal assessment is useful. The results across the three respondent groups are not redundant; each adds new information. Still, it would have been troublesome if the correlations among respondent groups were zero or even negative, which they were not.

The modest between respondent group correlations, ranging from .30 to .13 are typical of what is reported in the literature. Unfortunately, none of the relevant literature is on assessment of principals. Harris and Schaubroeck (1988) in a review of studies in education psychology and measurement psychology found an average correlation between self and supervisor of .35. Atwater, Ostroff, Yammarino, and Fleenor (1998) report a correlation of .25 between self and supervisor, .25 between self and subordinate, and .33 between supervisor and subordinate. Murphy and Deshon (2000, p.822) speculate about the less than perfect correlations, saying, "There are at least four reasons, none of which can be sensibly interpreted as random measurement error, why raters might be expected to disagree: (a) systematic differences in what is observed, (b) systematic differences in access to information other than observations of performance, (c) systematic differences in expertise in interpreting what is observed, and (d) systematic differences in evaluating what is observed."

Parallel Forms

As has been described elsewhere, there are two forms to the VAL-ED, A and C. Each form was created by randomly selecting two items from the pool of items written for each of the 36 cells in the six-by-six conceptual framework. In that sense, the 72 items in Form A are, stratified by cell, randomly equivalent to the 72 items in Form C. In the national field trial, within strata, forms were randomly assigned to districts as they were recruited. Unfortunately, this did not result in equal numbers of schools using Form A versus Form C. Neither did it result in equal numbers of schools who returned data for Form A and Form C. Part of the explanation is that the districts differ dramatically in size and, perhaps by chance, some larger districts got assigned Form C. In any event, for principals, there are data for 106 Form A schools and 130 Form C schools. For supervisors, there are data for 124 Form A schools and 130 Form C schools and for teachers, there are data for 113 Form A schools and 132 Form C schools. Tables 4.24, 4.25, and 4.26 compare Form A to Form C on means and standard deviations first for principal data, then for supervisor data, and last for teacher data. As can be seen across the three tables for the entry on total score, the means for Form A are very similar to the means for Form C. For principal data, they are different by four hundredths of a point, for supervisor data, three hundredths of a point, and for teacher data, three hundredths of a point. The standard deviations are similar as well. For principal data, they are different by two hundredths of a point and for teacher data, one hundredth of a point. However, for supervisor data, they are different by .12.

Table 4.24 Comparing Forms: Principal Data

	Means		Standard Deviation	
	A (N=106)	C (N=130)	A	C
Total Score	3.50	3.54	0.490	0.510
Core Components				
High Standards	3.59	3.71	0.502	0.533
Rigorous Curriculum	3.47	3.52	0.538	0.609
Quality Instruction	3.70	3.63	0.568	0.564
Culture of Learning	3.72	3.74	0.565	0.570
External Community	3.12	3.34	0.633	0.595

Performance Accountability	3.39	3.31	0.552	0.595
Key Processes				
Planning	3.51	3.50	0.518	0.533
Implementing	3.50	3.55	0.499	0.524
Supporting	3.71	3.66	0.544	0.517
Advocating	3.42	3.48	0.513	0.534
Communicating	3.47	3.59	0.542	0.559
Monitoring	3.40	3.47	0.537	0.610

Table 4.25 Comparing Forms: Supervisor Data

	Means		Standard Deviation	
	A (N=124)	C (N=130)	A	C
Total Score	3.66	3.69	0.709	0.590
Core Components				
High Standards	3.65	3.75	0.735	0.628
Rigorous Curriculum	3.63	3.66	0.752	0.659
Quality Instruction	3.74	3.72	0.765	0.622
Culture of Learning	3.78	3.83	0.722	0.634
External Community	3.51	3.59	0.740	0.639
Performance Accountability	3.60	3.55	0.779	0.664
Key Processes				
Planning	3.64	3.68	0.729	0.617
Implementing	3.63	3.65	0.743	0.645
Supporting	3.78	3.78	0.709	0.596
Advocating	3.61	3.62	0.754	0.566
Communicating	3.67	3.76	0.717	0.614
Monitoring	3.64	3.67	0.741	0.684

Table 4.26 Comparing Forms: Teacher Data

	Means		Standard Deviation	
	A (N=113)	C (N=132)	A	C
Total Score	3.61	3.58	0.428	0.415
Core Components				
High Standards	3.61	3.66	0.449	0.412
Rigorous Curriculum	3.60	3.51	0.412	0.423
Quality Instruction	3.71	3.58	0.429	0.422
Culture of Learning	3.69	3.67	0.441	0.462
External Community	3.46	3.54	0.444	0.414
Performance Accountability	3.53	3.45	0.462	0.428
Key Processes				
Planning	3.56	3.54	0.447	0.404
Implementing	3.59	3.58	0.456	0.417
Supporting	3.73	3.63	0.428	0.438
Advocating	3.57	3.54	0.402	0.384
Communicating	3.62	3.64	0.436	0.440
Monitoring	3.59	3.52	0.443	0.455

Each of the tables also compares means and standard deviations on the two forms for each subscale. Generally, the differences are small, though the differences noted between the standard deviation for Form A and Form C for supervisor data are seen across the data set, with Form C in all cases being less variable than Form A, sometimes by as much as .15. More generally, however, there are some differences by subscale. The differences are as large as .22 for the means in the principal data and teacher data, though most differences are .10 or less. The largest difference of .22 for Connections to External Communities on the principal forms represents a standardized effect size of approximately .36. Whether this is evidence of some lack of parallelness of the two forms or of non-randomly equivalent samples is impossible to say.

The distribution of item means for the aggregate data across respondent groups for Form A versus Form C is found on Figure 4.1. The distributions of item means look similar, although not identical. Form A had the easiest item and the hardest item. Still, the distributions are parallel, one to the other, and again provide support for Forms A and Forms C operating as parallel forms.

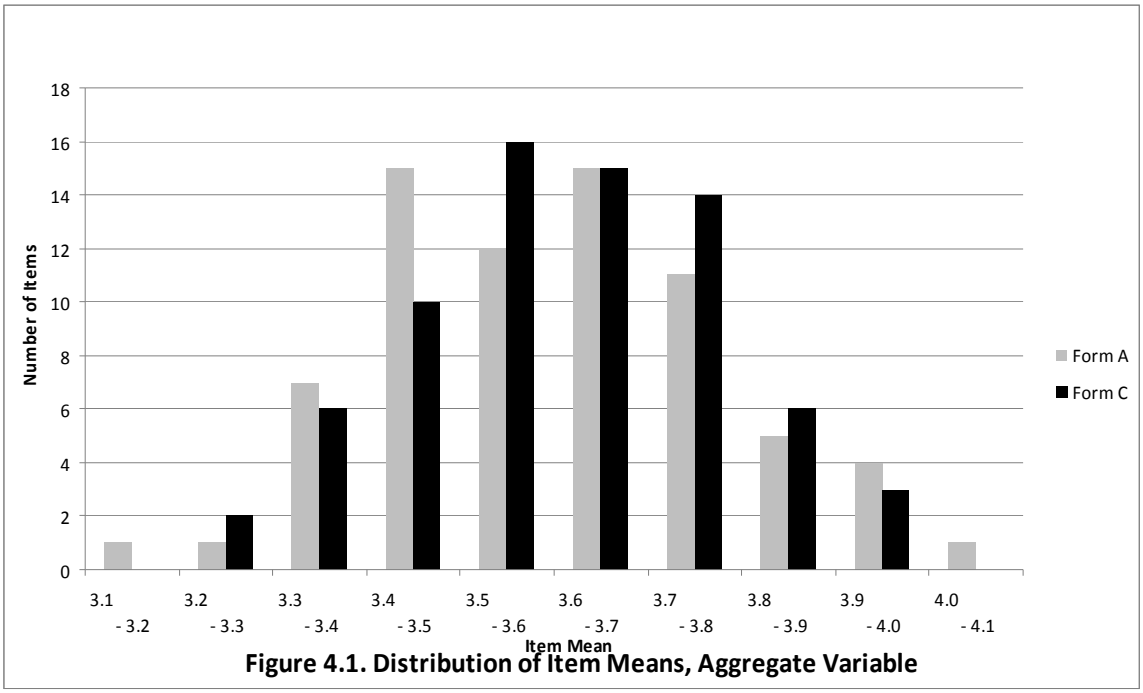


Figure 4.1. Distribution of Item Means, Aggregate Variable

Uses of Evidence and "Don't Know"

The VAL-ED instrument asks each respondent to think about the item describing a principal's behavior and before rating the principal on effectiveness as to that item, first identify what sources of evidence they have for making their effectiveness rating. Alternatives from which they can select as many as appropriate are: a) reports from others, b) personal observations, c) school documents, d) school projects or activities, e) other sources, and f) no evidence. When "No evidence" was checked by a supervisor or teacher, the effectiveness rating had to be "ineffective" or the respondent indicated "Don't know." Principals were not given the option of "Don't know." If they checked no evidence then they were required to rate themselves ineffective on that item. For the online version, the respondent is forced to conform to these rules. For the paper and pencil form an item is dropped when a respondent does not conform to these rules. A principal's effectiveness then is the mean item response for the reduced set of items. Similarly, when no effectiveness rating is given and instead "Don't know" is checked the item does not count toward the principal's effectiveness rating.

Table 4.27reports as an average across the 72 items (and disregarding form) the percent of times respondents indicated each type of source of evidence or no evidence. Since there were no restrictions on the number of sources of evidence that could be checked for any individual item, the percentages do not add to 100% across types of evidence within a respondent group. No Evidence was checked 2.6% of the time by principals, 7.5% of the time by supervisors, and 10.3% of the time by teachers. For teachers, principals, and supervisors, the most common source of evidence was personal observation. In contrast to teachers, however, supervisors and principals indicated school documents as a source of evidence nearly as frequently as personal observations, while teachers were less than half as likely to select school documents as they were personal observations. All of

the sources of evidence options were selected frequently by each respondent group, though Other Sources of evidence were used less frequently by teachers, 10% of the time, than by supervisors, 20% of the time, or principals, 26% of the time.

Table 4.27 Sources of Evidence Used by Respondent Group

	Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence
Principals	33.27%	61.91%	56.84%	36.19%	26.31%	2.60%
Supervisors	36.78%	58.01%	53.26%	27.18%	19.96%	7.52%
Teachers	24.24%	65.04%	29.27%	20.54%	10.04%	10.25%

The hypothesis is that by having respondents first select one or more sources of evidence before indicating a principal's effectiveness on an item, the respondent will think more deeply about the item and the principals' effectiveness and the resulting data will be better than the data would if just an effectiveness rating scale were used. Unfortunately, there is no way to test this hypothesis using our field test data. During the development of the instrument, however, both in cognitive interviews and in pilot studies, participants indicated that they liked that the instrument required them to reflect on sources of evidence and felt that it was useful.

An analysis of types of sources of evidence checked by core component and by key process revealed no differences for principals. For supervisors, however, they were much more likely to check "no evidence" for rating effectiveness having to do with Connections to External Communities (16%) and somewhat more likely to indicate "No evidence" for Monitoring (11%), Advocating (10%), and Performance Accountability (10%). There were not, however, substantial differences among core components or key processes in the frequency with which specific types of evidence were checked, including "Other Sources." For teachers, there were substantial differences on the frequency with which "no evidence" was checked. Again, Connections to External Communities was most often checked for "no evidence" (18% of the time), but Monitoring,

Advocating, Planning, and Performance Accountability were all checked at least 10% "No evidence," and Rigorous Curriculum, Quality Instruction, Implementing, and Communicating were close behind with more than 8%. As for supervisors, however, there were no notable differences in type of evidence checked by core component or key process other than their checking no evidence at all.

Table 4.28 reports the percent of respondents checking "don't know" averaged across items for total score and each of the core components and each of the key processes. The highest percentage of "Don't knows" were for External Communities and Performance Accountability for both supervisors and teachers. Twenty-nine percent of teachers checked "don't know" for the items asking about Connections to External Communities and 21% checked "don't know" for items on Performance Accountability. The statistics for supervisors were 19% and 12%, respectively. Similarly, Monitoring had a relatively high percent of "don't knows," with 23% for teachers and 13% for supervisors. In fact, of the twelve subscales, the percent of "Don't knows" averaged across items for teachers was above 10% for all but High Standards, Culture of Learning, and Supporting. For supervisors, the percent of "Don't knows" was above 10% for External Communities, Performance Accountability, Advocating, and Monitoring.

Table 4.28 Percent Don't Know by Respondent and Subscale

	Supervisors Percent Don't Know	Teacher Percent Don't Know
Total Score	9.31%	15.05%
Core Components		
High Standards	4.07%	7.99%
Rigorous Curriculum	7.64%	13.32%
Quality Instruction	7.64%	10.86%
Culture of Learning	5.94%	8.49%
External Community	18.96%	28.97%
Performance Accountability	11.58%	20.64%
Key Processes		
Planning	7.28%	14.41%
Implementing	9.48%	12.82%
Supporting	5.91%	8.37%

Advocating	11.58%	19.34%
Communicating	8.30%	12.16%
Monitoring	13.29%	23.16%

We investigated the prevalence of each of eight different types of errors that respondents could make in completing the instrument: a) omitting the item, b) failing to complete a source of evidence, c) failing to complete an effectiveness rating, d) checking a source of evidence but also checking "No Evidence" and giving a rating, e) checking a source of evidence and indicating "Don't Know" for the effectiveness rating, f) checking no evidence and then giving an effectiveness rating, g) checking "No Evidence" and then not completing the effectiveness rating scale including not indicating "Don't Know," and h) not completing the "Sources of evidence" portion of the item and checking "Don't Know." Table 4.29 provides the prevalence of each of the 8 kinds of errors for each of the 3 respondent groups. As can be seen in the table, all of the 8 types of errors were infrequent and in no case did they happen in 2% or more of the cases. The most common error was to fail to check a source of evidence and still provide an effectiveness rating. This happened on 1.94% of the items for principals, .73% of the items for supervisors, and 1.57% of the items for teachers.

Table 4.29 Prevalence of Response Errors by Respondent

	Complete Item Omission	Omission of Evidence	Omission of Rating	Evidence + No Evidence + Rating	Evidence + DK	No Evidence + Rating	No Evidence + DK Not Checked	No Evidence Not Checked + DK Checked
Principal	0.29%	0.38%	0.18%	0.29%	0.00%	1.94%	0.35%	0.00%
Supervisor	0.27%	0.33%	0.14%	0.30%	0.10%	0.73%	0.22%	2.49%
Teacher	0.76%	1.18%	0.28%	0.25%	1.15%	1.57%	0.34%	5.54%

The results of uses of evidence and "Don't know" seem reasonable. Principals were much more likely to check sources of evidence than were supervisors and teachers. Teachers were more

likely to check "Don't know" than were supervisors. Further, the error analysis indicates a low prevalence of errors in how respondents completed the instrument with no trouble spots either for the separate core components or the separate key processes.

Evidence of the VAL-ED's Feasibility

In the national field trial, once a respondent had completed the VAL-ED, the respondent was asked to respond to nine items on a scale of 1=strongly disagree to 4=strongly agree. Table 4.30 provides the results by respondent group and aggregated across respondent groups. Three of the questions had to do with perceptions of the validity of the items. The means for all three respondent groups were above "agree" and below "strongly agree" on the response scale when asked, "I do not believe the items are biased against any race or gender of a principal being assessed." When asked if "I believe the vast majority of items focus on important leadership behaviors," both principals and supervisors were well above "agree" on the scale and teachers were slightly above "agree." Results were very similar when respondents were asked, "This assessment is appropriate for use at the elementary, middle, and high school levels."

Table 4.30 Responses to VAL-ED Feasibility Questions by Respondent Group, National Field Test

	Principal		Teacher		Supervisor		Overall	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
I found this response form easy to use.	2.85	0.63	2.65	0.29	2.95	0.58	2.81	0.33
I understood the vast majority of the items	3.20	0.53	2.94	0.21	3.27	0.45	3.13	0.26
I believe the vast majority of the items focus on important leadership behaviors.	3.18	0.50	3.05	0.20	3.30	0.55	3.17	0.27
I do not believe the items are biased against any race or gender of a principal being assessed.	3.44	0.50	3.33	0.16	3.44	0.50	3.41	0.25
This assessment is appropriate for use at the elementary, middle, and high school levels.	3.15	0.62	3.05	0.20	3.27	0.61	3.16	0.32

I would prefer a web-based format for this assessment over the paper-and-pencil version I just completed.	3.09	0.86	2.83	0.32	3.22	0.80	3.04	0.45
Teachers should have input into the assessment of their principal's leadership.	3.36	0.62	3.45	0.16	3.18	0.61	3.33	0.29
I would support the use of this assessment instrument to hold principals accountable in my district.	2.72	0.80	2.85	0.29	2.74	0.71	2.78	0.38
The amount of time required to complete this instrument is reasonable.	2.99	0.63	2.76	0.24	2.71	0.60	2.82	0.34
1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree								
Significant Differences:								
Question 1, Teachers significantly lower than Principals and Supervisors								
Question 2, Teachers significantly lower than Principals and Supervisors								
Question 3, Teachers lower than Principals, Principals and Teachers lower than supervisors								
Question 4, Teachers significantly lower than Principals and Supervisors								
Question 5, Principals and teachers significantly lower than supervisors								
Question 6, Teachers significantly lower than Principals and Supervisors								
Question 7, Supervisors significantly lower than Principals and Teachers								
Question 9, Teachers and supervisors significantly lower than Principals								

All respondents were well above the "agree" point on the response scale when asked, "Teachers should have input into the assessment of their principal's leadership." Principals were nearly as enthusiastic about having teachers involved in their assessment as were teachers themselves.

When asked, "I would support the use of this assessment instrument to hold principals accountable in my district," results were slightly above the neutral point and leaning toward agreement with principals having a mean of 2.72, supervisors 2.74, and teachers 2.85. This item implies using the instrument for summative evaluation purposes by using the word "accountability." Respondents might have been even more positively inclined toward the use of the VAL-ED for formative purposes to help the principal identify strengths and weaknesses that could be targeted for future improvement.

When asked whether they would prefer a web-based format to the paper and pencil version that they had just completed, all three respondent groups tended to agree, though supervisors were the most enthusiastic, followed by principals, and teachers the least.

Answers to the feasibility questions were also compared across levels of schooling (Table 4.31). For each question, the mean response was not significantly different regardless of whether the respondent came from elementary school, middle school, or high school. The one exception was on whether the assessment was appropriate for use at elementary, middle, and high schools. Here, middle school respondents were significantly more likely to agree, though the difference was roughly .1 points on the four-point agreement scale, and all three groups still more than agreed with the statement.

Table 4.31 Responses to VAL-ED Feasibility Questions by School Level, National Field Test								
	Elementary		Middle		High		Overall	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
I found this response form easy to use.	2.80	0.34	2.82	0.32	2.82	0.32	2.81	0.33
I understood the vast majority of the items	3.10	0.26	3.17	0.24	3.14	0.29	3.13	0.26
I believe the vast majority of the items focus on important leadership behaviors.	3.13	0.28	3.20	0.26	3.21	0.26	3.17	0.27
I do not believe the items are biased against any race or gender of a principal being assessed.	3.37	0.26	3.45	0.23	3.41	0.25	3.41	0.25
This assessment is appropriate for use at the elementary, middle, and high school levels.	3.11	0.29	3.24	0.30	3.13	0.35	3.16	0.32
I would prefer a web-based format for this assessment over the paper-and-pencil version I just completed.	3.00	0.45	3.08	0.46	3.05	0.42	3.04	0.45
Teachers should have input into the assessment of their principal's leadership.	3.32	0.30	3.35	0.31	3.33	0.25	3.33	0.29
I would support the use of this assessment instrument to hold principals accountable in my district.	2.78	0.38	2.80	0.37	2.75	0.38	2.78	0.38
The amount of time required to complete this instrument is reasonable.	2.77	0.37	2.85	0.31	2.86	0.33	2.82	0.34
1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree								
Significant differences:								
Question 5, Elementary significantly lower than Middle								

Similarly, the responses to the nine items were compared and contrasted across urban, suburban, and rural (Table 4.32). Here, rural teachers were slightly less likely to agree that they understood the vast majority of items though their mean was still above 3.0 on the agreement scale. Also, urban and suburban respondents were somewhat less likely than rural respondents to agree that the amount of time required to complete the assessment was appropriate.

Table 4.32 Responses to VAL-ED Feasibility Questions by School Location, National Field Test								
	Urban		Suburban		Rural		Overall	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
I found this response form easy to use.	2.76	0.35	2.88	0.32	2.79	0.27	2.81	0.33
I understood the vast majority of the items	3.19	0.26	3.13	0.26	3.05	0.25	3.13	0.26
I believe the vast majority of the items focus on important leadership behaviors.	3.15	0.28	3.20	0.26	3.17	0.27	3.17	0.27
I do not believe the items are biased against any race or gender of a principal being assessed.	3.38	0.25	3.42	0.24	3.42	0.26	3.41	0.25
This assessment is appropriate for use at the elementary, middle, and high school levels.	3.16	0.32	3.18	0.32	3.12	0.31	3.16	0.32
I would prefer a web-based format for this assessment over the paper-and-pencil version I just completed.	3.03	0.45	3.05	0.47	3.04	0.41	3.04	0.45
Teachers should have input into the assessment of their principal's leadership.	3.30	0.31	3.36	0.26	3.33	0.30	3.33	0.29
I would support the use of this assessment instrument to hold principals accountable in my district.	2.75	0.39	2.81	0.39	2.77	0.32	2.78	0.38
The amount of time required to complete this instrument is reasonable.	2.76	0.37	2.81	0.34	2.95	0.25	2.82	0.34
1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree								
Significant differences:								
Question 2, Rural significantly lower than Urban								
Question 9, Urban and suburban significantly lower than rural								

Performance Standards

The results of the principal assessment are reported in terms of performance levels: distinguished, proficient, basic, and below basic as well as in terms of percentile ranks. Seven educators were recruited to participate in the evaluation of draft Performance Level Descriptors (PLDs) to be reported with the criterion-referenced results from the VAL-ED. The initial draft of the PLDs was:`

- Below Basic – Leadership behaviors of core components and key processes of insufficient quality that over time are unlikely to bring the school to produce acceptable value-added to student achievement and social learning.
- Basic – Leadership behaviors of core components and key processes of a quality that over time are likely to bring the school to produce acceptable value-added to student achievement and social learning for some sub-groups of students but not all.
- Proficient – Leadership behaviors of core components and key processes of sufficient quality that over time are likely to bring the school to produce acceptable value-added to student achievement and social learning for all students.
- Distinguished – Leadership behaviors of core components and key processes at levels of excellence that over time are virtually certain to bring the school to a point that produces strong value-added to student achievement and social learning for all students.

The seven educators included two supervisors of principals, three principals (one each from elementary, middle, and high schools), and two teachers (one each from elementary and high schools). The participants were from five different states. Participants were mailed packets including unlabeled PLDs and a set of instructions describing the task and the questions to answer. The packet also included a paper copy of the VAL-ED instrument with the framework and definitions of core components and key processes. The questions were:

1) Does each description clearly and adequately describe a different level of proficiency for school leaders?
2) Is the wording clear?
3) Do the descriptions distinguish and differentiate amongst the levels of performance?
4) Do you have other reactions to the PLDs?
5) If, in your opinion, changes are needed, please change or edit the wording of each of the PLDs to better describe the outcomes of effective and ineffective leadership behaviors at the four levels of proficiency.

Respondents were also asked to sort the unlabeled PLDs into the proper order from high to low effectiveness. All seven educators responded to the task by answering the questions and returning their responses to study personnel.

All respondents were able to correctly sort the unlabeled PLDs into the correct order, indicating that the PLDs clearly describe levels of effectiveness that are interpretable by practitioners. Respondents also agreed that the PLDs distinguished and differentiated amongst levels of performance. However, respondents did not believe that the PLDs were as clearly worded as they should have been.

One set of comments from respondents indicated excessive wordiness. One respondent called the PLDs "too complicated," a second said they were "too wordy but nevertheless clear," and a third said they were "too long." Another respondent suggested that the PLDs were too long because they were only one sentence. Finally, one respondent pinned the excess length on the beginning of the sentence, saying "the first part of the PLDs is a little lengthy." Overall, respondents were in agreement that length and/or wordiness was a concern.

A second set of comments concerned wording and ambiguity. Several respondents were concerned about the similarities among the PLDs, with one respondent saying they were "worded in such a similar way that could easily lead to misinterpretation," and another relating that he/she "had to read them several times looking for key vocabulary to differentiate between descriptions." Other respondents suggested the PLDs were "rather ambiguous," "not easily understood," or "too complicated," suggesting the "statements could be written in simpler terms."

Finally, two respondents were concerned about specific wording within the PLDs. One respondent was unclear on the interpretation of the word "virtually," suggesting this word be

removed from the PLDs. A second was unclear as to the distinction between measuring effectiveness and sufficient effectiveness.

Overall, results from the PLD evaluation task suggested that the study team needed to further refine the PLDs. Respondents clearly indicated a level of confusion with the PLDs that should be addressed in order to help respondents make valid decisions using VAL-ED results. In response to the concerns, the final version of the PLDs was created by study team members based on the three sets of modified PLDs provided by respondents. The final PLDs take into account the feedback from the seven respondents.

The final proficiency level descriptors are as follows:

- Below Basic – A leader at the below basic level of proficiency exhibits leadership behaviors of core components and key processes at levels of effectiveness that over time are unlikely to influence teachers to bring the school to a point that results in acceptable value-added to student achievement and social learning for students.

- Basic – A leader at the basic level of proficiency exhibits leadership behaviors of core components and key processes at levels of effectiveness that over time are likely to influence teachers to bring the school to a point that results in acceptable value-added to student achievement and social learning for some sub-groups of students, but not all.

- Proficient – A proficient leader exhibits leadership behaviors of core components and key processes at levels of effectiveness that over time are likely to influence teachers to bring the school to a point that results in acceptable value-added to student achievement and social learning for all students.

- Distinguished – A distinguished leader exhibits leadership behaviors of core components and key processes at levels of effectiveness that over time are virtually certain to

influence teachers to bring the school to a point that results in strong value-added to student achievement and social learning for all students.

With the PLDs in final form, a Bookmark approach to standard setting was undertaken (for more detail on the standard setting, see Porter et al., 2008). The Bookmark requires an item-ordered booklet. To order the items in difficulty, an aggregate variable was created. The aggregate variable was defined as the arithmetic mean of an item's mean item response across the principal, the supervisor, and the mean for the teachers in the principal's school. The principal, the supervisor, and the mean of the teachers were thus equally weighted in creating the aggregate variable. The variable could range from 1 to 5, representing the levels on the effectiveness rating scale. The data for the item-ordered booklet come from the national field trial.

The item-ordered booklet consisted of the 72 items from Form A. The decision to use Form A was based on the need to make the task for panelists manageable, the fact that Forms A and C were constructed to be parallel, and the findings from the national field trial that the two forms have nearly identical means and standard deviations.

Item means on the aggregate variable ranged from a low of 3.18 for the most difficult item to 4.04 for the easiest item. The distribution of schools on the aggregate variable ranged from 2.57 for the lowest-rated principal to 4.51 for the highest-rated principal. The range for schools was larger than the range for the item means, as might be expected.

A panel of 22 experts was recruited from across the nation, consisting of ten principals, four teachers, four supervisors of principals, two researchers of school leadership, and two education policymakers. The panel was convened in August of 2008. At the end of the standard setting event, panelists had placed three cuts on the effectiveness rating scale continuum from 1.0 to 5.0. The cut to distinguish proficient from basic was set at 3.60, the cut between distinguished and proficient at

3.77, and the cut between basic and below basic at 3.42. These cuts resulted in 30% of principals in the national field trial below basic, 50% below proficient, and 70% below distinguished.

Panelists were positive about the process and generally satisfied with the cuts set. Nevertheless, 24% expressed some concern about where the cuts were set between basic and below basic and between distinguished and proficient. In response to panelists' concerns, a post-standard setting communication with panelists asked them whether they wished to a) keep the cuts where they were set at the end of the standard-setting event, b) move the cuts to the median cut for the least demanding table (there were five tables that operated independently in the standard setting task) for the distinction between basic and below basic and the most demanding table for the distinction between proficient and distinguished, or c)move the cut for basic and below basic to the least demanding table and the cut for the distinction between proficient and distinguished to 4.0 on the impact scale. All 22 panelists responded. Twenty-one favored moving the basic to below basic cut to be less demanding and the proficient to distinguished cut to be more demanding. Of those twenty-one, all favored moving the basic to below basic cut to 3.29, yielding 17% of principals below basic, and 18 of the 21 favored moving the cut between distinguished and proficient to 4.00, yielding 14.2% of the principals distinguished. The final decision, then, is to set the cuts consistent with the panelists' preferences. The cut between basic and below basic is 3.29, between basic and proficient is 3.60, and between proficient and distinguished is 4.00.

The proficiency level cut scores are used in reporting principal performance, first and foremost on the total score aggregated across the three respondent groups. Because the proficiency cuts are made on the mean item response scale, even though they were made based on judgments for total score aggregated across the three respondent groups, they can be used to distinguish proficiency levels by respondent group and on each of the twelve subscales. To use proficiency

levels with subscales, we must make the assumption that the judgments on the total score apply equally to the VAL-ED subscales. This assumption allows us to report, for instance, if a principal's performance as reported by teachers on rigorous curriculum is distinguished, proficient, basic, or below basic? Further, the proficiency cuts can be used to distinguish performance at the cell level. For example, is a principal's behavior as rated across the three respondent groups distinguished, proficient, basic, or below basic for planning for high standards for student learning? As seen in Appendix B, the principal report form indicates for each of the 36 cells in the six-by-six conceptual framework for which cells performance was rated proficient or better, for which cells the behavior was rated as basic and for which cells the behavior was rated below basic. These three levels of performance are indicated by green, yellow, and red colors respectively.

Norms

Using the results from the National Field Trial, initial norms for the VAL-ED were set for total score and each of the twelve sub-scores, once for the data aggregated across the three respondent groups, once for the principal data, once for the supervisor data, and once for the teacher data. Collectively then, there are 52 sets of norms. Table 4.33 provides the norms for total score on the aggregated data across respondent groups. As can be seen in Table 4.33, total score mean item response ranged from a lower 2.573 to a high of 4.506 with the median of 3.604. Using this norms table, for example, a principal with an aggregated total score mean item response of 4.00 would have the percentile rank of 86.

Table 4.33 Total Score Aggregated Data Across Respondent Groups

Total Score	Percentile Rank
2.573	0.4
2.745	0.9
2.770	1.3
2.847	1.8
2.893	2.2
2.911	2.7

2.926	3.1
2.936	3.6
2.954	4.1
2.978	4.5
2.993	5.0
3.044	5.4
3.046	5.9
3.065	6.3
3.073	6.8
3.077	7.2
3.080	7.7
3.095	8.2
3.096	8.6
3.141	9.1
3.155	9.5
3.156	10.0
3.165	10.4
3.183	10.9
3.187	11.4
3.195	11.8
3.198	12.3
3.211	12.7
3.212	13.2
3.216	13.6
3.235	14.1
3.269	14.5
3.273	15.0
3.274	15.5
3.275	15.9
3.276	16.4
3.277	16.8
3.293	17.3
3.310	17.7
3.315	18.2
3.330	18.7
3.333	19.1
3.338	19.6
3.341	20.0
3.354	20.5
3.361	20.9
3.365	21.4
3.368	21.8
3.368	22.3
3.376	22.8
3.380	23.2
3.381	23.7
3.383	24.1
3.391	24.6

3.402	25.0
3.404	25.5
3.404	26.0
3.405	26.4
3.415	26.9
3.419	27.3
3.421	27.8
3.425	28.2
3.433	28.7
3.438	29.1
3.444	29.6
3.447	30.1
3.458	30.5
3.463	31.0
3.465	31.4
3.465	31.9
3.467	32.3
3.469	32.8
3.473	33.3
3.474	33.7
3.474	34.2
3.476	34.6
3.490	35.1
3.490	35.5
3.495	36.0
3.509	36.4
3.509	36.9
3.511	37.4
3.515	37.8
3.515	38.3
3.519	38.7
3.522	39.2
3.526	39.6
3.531	40.1
3.535	40.5
3.537	41.0
3.540	41.5
3.540	41.9
3.542	42.4
3.545	42.8
3.547	43.3
3.551	43.7
3.553	44.2
3.554	44.7
3.555	45.1
3.560	45.6
3.572	46.0
3.574	46.5

3.583	46.9
3.584	47.4
3.587	47.8
3.592	48.3
3.593	48.8
3.596	49.2
3.603	49.7
3.604	50.1
3.605	50.6
3.606	51.0
3.611	51.5
3.612	52.0
3.614	52.4
3.618	52.9
3.623	53.3
3.626	53.8
3.627	54.2
3.629	54.7
3.631	55.1
3.639	55.6
3.642	56.1
3.649	56.5
3.651	57.0
3.664	57.4
3.665	57.9
3.670	58.3
3.671	58.8
3.674	59.3
3.682	59.7
3.689	60.2
3.691	60.6
3.705	61.1
3.707	61.5
3.710	62.0
3.717	62.4
3.720	62.9
3.723	63.4
3.724	63.8
3.724	64.3
3.728	64.7
3.734	65.2
3.738	65.6
3.738	66.1
3.742	66.6
3.756	67.0
3.757	67.5
3.758	67.9
3.761	68.4

3.763	68.8
3.775	69.3
3.782	69.7
3.784	70.2
3.799	70.7
3.803	71.1
3.806	71.6
3.809	72.0
3.811	72.5
3.812	72.9
3.817	73.4
3.817	73.8
3.827	74.3
3.833	74.8
3.844	75.2
3.866	75.7
3.867	76.1
3.867	76.6
3.868	77.0
3.871	77.5
3.880	78.0
3.881	78.4
3.885	78.9
3.889	79.3
3.894	79.8
3.909	80.2
3.920	80.7
3.923	81.1
3.925	81.6
3.926	82.1
3.947	82.5
3.962	83.0
3.967	83.4
3.968	83.9
3.969	84.3
3.989	84.8
3.994	85.3
4.002	85.7
4.017	86.2
4.019	86.6
4.032	87.1
4.037	87.5
4.047	88.0
4.052	88.4
4.067	88.9
4.076	89.4
4.086	89.8
4.100	90.3

4.112	90.7
4.124	91.2
4.138	91.6
4.154	92.1
4.158	92.6
4.181	93.0
4.183	93.5
4.184	93.9
4.199	94.4
4.214	94.8
4.243	95.3
4.254	95.7
4.293	96.2
4.312	96.7
4.345	97.1
4.360	97.6
4.372	98.0
4.395	98.5
4.489	98.9
4.506	99.4

Summary and Conclusions

The VAL-ED is an assessment of principal instructional leadership in K-12 schools. The project was stimulated and funded by the Wallace Foundation and completed during a three-year period, 2005 – 2008.

The VAL-ED assessment consists of two parallel forms (A and C) and is available in both a paper and an online version. The instrument is a 360-degree assessment; for each school, the principal completes a self-evaluation, the teachers in the school evaluate the principal, and the supervisor of the principal evaluates the principal, all using the same 72-item instrument, which requires 20 – 30 minutes to complete.

The conceptual framework that drives the development, reporting, and interpretation of the VAL-ED consists of six core components by six key processes. The core components are features of effective schools. The key processes are leadership behaviors that principals can employ to lead their schools to a point in time when they are strong on each of the core components. Thus, for each

core component, there are six key processes of leadership behaviors. The six-by-six matrix identifies 36 cells with two items in each cell. Forms A and C were created by randomly selecting two items from a set of items written for each of the 36 cells.

Results from the VAL-ED are reported in terms of mean item effectiveness on a five point effectiveness rating scale. Percentile ranks are based on a national field trial and performance standards were set by a 22-member panel of experts. There is a total aggregate score, which is a function of the responses to all 72 items across the supervisor, the principal, and the teachers, where supervisor, principal, and teachers are weighted equally. Results are also reported separately for each respondent group and for each core component and each key process.

The first and most important argument for the content validity of the VAL-ED is that the items were written against the six-by-six conceptual framework. The conceptual framework is based on the literature on school leadership effects on student achievement. The VAL-ED assessment was developed to: a) work well in a variety of settings and circumstances; b) be construct valid; c) be reliable; d) be feasible for widespread use; e) provide accurate and useful reporting of the results; f) be unbiased; g) yield a diagnostic profile for summative and formative purposes; h) be able to measure progress over time in the development of leadership; and i) predict important outcomes.

The development of the VAL-ED was embedded in a research paradigm. First, items were written to fit each of the 36 cells in the core components by key processes conceptual framework. As many items were written as leadership behaviors could be identified. For some cells more than 10 behaviors were identified and items written. Principals were asked to sort items into cells as a second check on the content validity of the items; sorting accuracy was good but some items were dropped and some re-written for clarity of target cell. Two rounds of cognitive interviews were

conducted in three districts each. In each case, principals, teachers, and principals' supervisors from elementary, middle, and high schools participated in the cognitive interviews. Based on the results, the instrument was revised. When the instrument was judged to be ready, a nine-school pilot test was conducted in one district involving elementary, middle, and high schools. Based on that pilot, the instrument was seen to have good internal consistency reliability, good construct validity, good face validity, but the 108-item instrument was too long and the effectiveness scale was not being used across its full range.

The instrument was revised to be shorter and have different benchmarks for the effectiveness scale. Cognitive interviews were conducted on the online instrument (the paper and pencil and online instruments are virtually identical). The positive results led to piloting the instrument once again, this time in eleven schools, again across elementary, middle, and high school. The results for this second pilot were encouraging. The reliability remained high. More of the range of the response scale was used. Completion time was seen as less of a problem and confirmatory factor analysis on the teacher data supported the conceptual framework against which the items were written.

A fairness review of the VAL-ED instructions and items was conducted to identify and remove aspects of test items or directions that might hinder respondents from completing the instrument. The fairness review was based on the fairness guidelines published and used by ETS. A panel of nine individuals completed the fairness review. The results indicated no fairness concerns in instructions or introductory content. Three items on Form A and one items on Form C were identified as requiring some edits to improve their fairness characteristics and these edits were made.

With Form A and Form C of the VAL-ED assessment of school leadership in final form, a national field trial was undertaken to establish the psychometric properties of the assessment, to establish percentile ranks for reporting, to build an item-ordered booklet for the Bookmark performance standard setting, to further investigate the perceived feasibility of the instrument, and to investigate design factors, including level of schooling, locale, and the parallelness of Forms A and C. The target was set at 300 schools: 100 elementary, 100 middle, and 100 high schools. The obtained sample had principal data on 235 schools, supervisor data on 253, and teacher data on 245. For 218 schools, there were data from all three respondent groups.

Based on the national field trial, supervisors were seen as slightly more positive on principal effectiveness than were principals with teachers in between. Controlling for other factors, high school principals were seen as slightly less effective than elementary or middle school principals, but the difference was small. Suburban school principals were seen as more effective than rural school principals and no significant differences were found between Forms A and C.

Confirmatory factor analysis, exploratory factor analysis with oblique factor rotations, and investigations of mean differences among core components, key processes, and their interactions were conducted to investigate the construct validity of the instrument. Confirmatory factor analysis supported both core components and key processes. Exploratory factor analysis identified factors for Performance Accountability, Connections to External Community, and Culture of Learning and Professional Behavior, as well as the key processes of Supporting and Advocating. There were significant differences between the means of the core components, with the exception that Rigorous Curriculum and Performance Accountability were not significantly different one from the other, nor were Quality of Instruction and Culture of Learning and Professional Behavior. For key processes, Supporting was significantly different from all other key processes, but the other key processes were

not significantly different one from another. Of the 630 pairs of contrasts among the 36 cells, for Form A 44% were significant and for Form C, 47% were significant. Finally, two analyses of the reliability of the difference between subscales were conducted. In the traditional analysis, the reliability of the difference of core components from the total score was generally good, and the reliability of the difference of key processes from the total score was generally poor. When comparing across core components and key processes, it was found that the reliability of the difference between core component subscales was quite strong, with mixed evidence as to the reliability of the difference between key process subscales. A second analysis used generalizability theory to investigate the reliability of contrasts between core components and between key processes. The reliabilities of the differences were surprisingly good given the notoriously low reliability of different scores. The results were similar to the results from the exploratory factor analysis. The reliabilities contrasting Culture of Learning, External Communities, Performance Accountability, and Rigorous Curriculum were all strong. For key processes, the reliabilities contrasting Supporting and Advocating were both strong.

See Tables 4.34 and 4.35 for a summary of the empirical support for the conceptual framework. Table 4.34 provides the evidence for the six core components and Table 4.35 provides evidence for the six key processes. In each case, there were four sources of evidence: a) the extent to which effectiveness ratings differed in their mean value among the core components (or key processes); b) the reliability of contrasting a core component (or key process) from the overall total score with the criterion being reliabilities of .50 or higher; c) using generalizability theory of finding that the core component (or key process) has significant unique variance; and d) exploratory factor analysis using oblique notation indicates a clear factor on both forms (a double plus) or a clear factor on one form but not the other (indicated with a single plus). Results of distinctions among the

six core components are presented in the form of one of three entries: positive evidence on both forms (++), positive evidence on one form but not the other (+), and no positive evidence (0) for each pair wise comparison of one core component to another. The first entry in a cell is for significant mean difference, the second for classical reliability of the differences, the third based on G-theory, and the fourth based on exploratory factor analysis. Most of the entries are ++ and every pair wise comparison has at least two ++'s. The conclusion is strong empirical support for the construct of the instrument. As is seen in Table 4.34, the empirical support for the core components is overwhelmingly positive. In Table 4.35, it is seen that the evidence in support of key processes is less strong than for core components. Of the six key processes, the most distinct were Support, Advocate, and Communicate.

Table 4.34 Evidence on the Ability to Differentiate VAL-ED Factors: Core Components

ANOVA Alpha diff. G-theory EFA	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	Connection Ext. Comm.	Perform. Accy.
Rigorous Curriculum	++ + ++ +					
Quality Instruction	++ + ++ ++	++ + ++ +				
Culture of Learning	+ ++ ++ ++	++ ++ ++ +	+ + ++ +			
Connection Ext. Comm	++ ++ ++ ++	++ ++ ++ ++	++ ++ ++ ++	++ ++ ++ ++		
Perform. Accy.	++ + ++ ++	+ ++ ++ ++	++ ++ ++ ++	++ ++ ++ ++	+ ++ ++ ++	

ANOVA:

++ indicates $p < .05$ on both forms
+ indicates p > .05 on one form, p < .05 on other form
0 indicates $p > .05$

Reliability of Difference:

++ indicates alpha > .50 for all respondents/forms
+ indicates alpha >.50 for at least one form/respondent
0 indicates alpha < .50 for all forms/respondents

G-theory:

++ indicates CC/KP has unique variance for all 6 form/respondent groups
+ indicates both CCs/KPs have at least one form/respondent with no unique variance

EFA:

++ indicates a clear factor on both forms
+ indicates a clear factor on one form
0 indicates no clear factors

Table 4.35 Evidence on the Ability to Differentiate VAL-ED Factors: Key Processes						
ANOVA Alpha diff. G-theory EFA	Planning	Implementing	Support	Advocate	Communic.	Monitor
Implementing	0 0 + 0					
Support	++ 0 ++ +	++ 0 ++ +				
Advocate	0 0 ++ +	0 + ++ +	++ + ++ +			
Communic.	+ 0 + 0	+ 0 + 0	+ + ++ +	+ + ++ +		
Monitor	0 + + 0	0 + + 0	++ + ++ +	0 + ++ +	+ + + 0	

ANOVA:

++ indicates $p < .05$

+ indicates p > .05 on one form, p < .05 on other form
0 indicates *p*>.05

Reliability of Difference:
++ indicates alpha > .50 for all respondents/forms
+ indicates alpha >.50 for at least one form/respondent
0 indicates alpha < .50 for all forms/respondents

G-theory:
++ indicates CC/KP has unique variance for all 6 form/respondent groups
+ indicates both CCs/KPs have at least one form/respondent with no unique variance

EFA:
++ indicates a clear factor on both forms
+ indicates a clear factor on one form
0 indicates no clear factors

The relationship between responses from principals, supervisors, and teachers was investigated. All three respondent groups were positively intercorrelated with the highest intercorrelation between principals and teachers (approximately .25). The correlation between supervisors and teachers was lower, and the lowest correlation was between principals and supervisors. Clearly, the information from one respondent group was not redundant with the information from the other respondent groups, a finding that supports the 360 degree approach to assessment.

Investigations of the use of sources of evidence, the "Don't know" option on the effectiveness scale, and a variety of possible errors that could be made in filling out the instrument revealed no problems.

Nine questions were asked of respondents after they completed the assessment. All three respondent groups indicated that they found the instrument to be easy to use and the items to be a) focusing on important leadership behaviors, b) understandable, c) not biased, d) appropriate for elementary, middle, and high school levels. Similarly, all respondent groups agreed that teachers

should have input into the assessment of principal leadership, but they neither agreed nor disagreed that the VAL-ED should be used to hold principals accountable in their district. Perhaps had the item asked about formative as well as summative evaluations, the responses would have been more enthusiastic about use of the instrument.

A Bookmark method was used to set performance standards for distinguished, proficient, basic, and below basic. A national panel of 22 experts participated. Ultimately, the standards were set yielding 17% of the national field trial principals below basic, 50% below proficient, and 86% below distinguished.

With the completion of the national field trial, the VAL-ED has been documented to have excellent reliability, strong validity, initial national norms for reporting percentile ranks, and performance standards to identify distinguished, proficient, basic, and below basic principals. The norms and the proficiency levels apply to both Form A and C, which can be used interchangeably, and for a paper and pencil as well as an online version of the assessment.

Work on establishing the psychometric properties of the VAL-ED is continuing with support from the US Department of Education's Institute for Education Sciences. During the period 2008 to 2012, studies will be completed on item bias, using differential item functioning (DIF) procedures, known group studies to see if the VAL-ED can distinguish between principals who are identified as in the top 20% of their professional peer group versus the lowest 20% of their professional peer group, stability of performance over time, a study of the consequences of using the VAL-ED to see how the results are used by supervisors and principals, and finally, a study to investigate on a longitudinal sample, the extent to which performance on the VAL-ED predicts a school's value added to student achievement.

References

Adams, J. E., & Kirst, M. W. (1999). New demands and concepts for educational accountability: Striving for results in an era of accountability. In J. Murphy & K. Louis (Eds.), *Handbook of research on educational administration, 2nd edition* (pp. 463-489). San Francisco: Jossey-Bass.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing.* Washington, DC: AERA, APA, & NCME.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: AERA, APA, & NCME.

Atwater, L. E., Ostroff, C., Yammarino, F. J., & Fleenor, J. W. (1998). Self-other agreement: Does it really matter? *Personnel Psychology, 51*(3), 577-598.

Betts, J. R., & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings. *Economics of Education Review, 22*, 343-352.

Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., & Sudman, S. (1991). *Measurement Errors in Surveys*. New York: J. Wiley.

Boyer, E.L. (1983). *High school: A report on secondary education in America*. New York: Harper & Row.

Brookover, W. B., & Lezotte, L. W. (1977). *Changes in school characteristics coincident with changes in student achievement.* East Lansing: College of Urban Development, Michigan State University.

Bryk, A., Camburn, E., & Louis, K. S. (1999). Professional community in Chicago elementary schools. Facilitating factors and organizational consequences. *Educational Administration Quarterly, 35*, 751-781.

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

Bryk, A. S., & Driscoll, M. E. (1988). *The high school as community: Contextual influences and consequences for students and teachers.* Madison, WI: University of Wisconsin-Madison, National Center on Effective Secondary Schools.

Bryk, A. S., & Schneider, B. (2002). *Trust in schools: A core resource for improvement.* New York: Russell Sage.

Burns, J. M. (1978). *Leadership.* New York: Harper & Row.

Butty, J., LaPoint, V., Thomas, V., & Thompson, D. (2001). The changing face of after school programs: Advocating talent development for urban middle and high school students. *NASSP Bulletin, 58*(262), 22-34.

Council of Chief State School Officers (CCSSO). (1996). *Interstate School Leaders Licensure Consortium Standards for School Leaders*. (Washington, DC: Council of Chief State School Officers

Conley, D. T. (1991). Lessons from laboratories in school restructuring and site-based decision making. *Oregon School Study Council Bulletin, 34*(7), 1-61.

Conley, D. T., & Goldman, P. (1994). Ten propositions for facilitative leadership. In J. Murphy & K. S. Louis (Eds.), *Reshaping the principalship: Insights from transformational reform efforts.* Thousand Oaks, CA: Corwin Press.

Cronbach L. J. (1970). How should we measure "change"—or should we? *Psychological Bulletin. 74(1),* 68-80.

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

Edmonds, R., & Frederiksen, J. R. (1978). *Search for effective schools: The identification and analysis of city schools that are instructionally effective for poor children.* Cambridge, MA: Harvard University, Center for Urban Studies.

Educational Testing Service (ETS). (2000). *ETS standards for quality and fairness.* Princeton, NJ: Educational Testing Service.

Elmore, R. F. (2005). Accountable leadership. *The Educational Forum, 69,* 134-142.

Feldt, L. S. (1995). Estimation of the Reliability of Differences Under Revised Reliabilities of Component Scores. *Journal of Educational Measurement. 32*(3) 295-301.

Eubanks, E. E., & Levine, D. U. (1983, June). A first look at effective schools projects in New York City and Milwaukee. *Phi Delta Kappan, 64*(10), 697-702.

Fullan, M., & Pomfret, A. (1977). Research on curriculum and instructional implementation. *Review of Educational Research, 47,* 335-397.

Garibaldi, A. M. (1993). *Improving urban schools in inner-city communities* (Occasional Paper No. 3). Cleveland, OH: Cleveland State University, Levine College of Urban Affairs, Urban Child Research Center.

Ginsberg, R., & Berry, B. (1990). The folklore of principal evaluation. *Journal of Personnel Evaluation in Education, 3*, 205-230.

Ginsberg, R., & Thompson, T. (1992). Dilemmas and Solutions Regarding Principal Evaluation *Peabody Journal of Education 68*(1), 58-74.

Glasman, N. S., & Heck, R. H. (1992). The changing leadership role of the principal: implications for principal assessment. *Peabody Journal of Education, 68*(1), 5-24.

Goldring, E., Cravens, X., Murphy, J., Porter, A., Elliott, S., & Carson, B. (In Press). The evaluation of principals: What and how do states and urban districts assess leadership? *Elementary School Journal (Accepted*)

Goldring, E., Cravens, X., Porter, A., Murphy, J., & Elliott, S.N. (2007). *The State of Educational Leadership Evaluation: What do States and Districts Assess?*

Goldring, E., & Hausman, C. (2001). Civic capacity and school principals: The missing link in community development. In R. Crowson & B. Boyd (Eds.), *Community development and school reform.* Greenwich, CT: JAI Press.

Goldring, E., Porter, A.C., Murphy, J., Elliott, S.N., & Cravens, X. (2007, March). Assessing learning-centered leadership: Connections to research, professional standards, and current practice. New York, N.Y.: Wallace Foundation.

Goldring , E., Porter, A., Murphy, J., Elliot, S., & Cravens, X. (in press). Assessing Learning-Centered Leadership: Connections to Research, Standards and Practice. Leadership and Policy in Schools.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hallinger, P., & Heck, R. (2002). What do you call people with visions? The role of vision, mission, and goals in school improvement. In K. Leithwood, P. Hallinger, G. Furman, J. MacBeath, B. Mulford, & K. Riley (Eds.), *The second international handbook of educational leadership and administration*. Dordrecht, The Netherlands: Kluwer.

Hallinger, P., & Heck, R. (1996). Reassessing the principal's role in school effectiveness: A review of empirical research, 1980-1985. *Educational Administration Quarterly*, 32(1), 5-44.

Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *Elementary School Journal, 86*, 217-247.

Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*(1). 43-62.

Heck, R. (1992). Principals' instructional leadership and school performance: Implications for policy development. *Educational Evaluation and Policy Analysis, 14*(1), 21-34.

Heck, R. H., & Hallinger, P. (1999). Next generation methods in the study of leadership and school improvement. In J. Murphy, & K. S. Louis (Eds.). *Handbook of Research on Educational Administration* (2nd ed., pp. 141-162). San Francisco: Jossey-Bass.

Henderson, A. T., & Mapp, K. L. (2002). *A new wave of evidence: The impact of school, family, and community connections on student achievement.* Austin, TX: Southwest Educational Development Laboratory.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport: American Council on Education & Praeger Publishers.

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

Lawson, H. A. (1999). Two new mental models for schools and their implications for principals' roles, responsibilities, and preparation. *Bulletin, National Association of Secondary School Principals, 83*(611), 8-27.

Lee, V. E., Smith, J. B., & Croninger, R. G. (1995). *Another look at high school restructuring. More evidence that it improves student achievement and more insights into why.* Madison, WI: Center on Organization and Restructuring of Schools.

Leithwood, K., Louis, K.S., Anderson, S., & Wahlstrom, K. (2004) *How leadership influences student learning.* New York, NY: The Wallace Foundation.

Leithwood, K. (1994). Leadership for school restructuring. *Educational Administration Quarterly, 30*, 498-518.

Leithwood, K., & Jantzi, D. (1990). *Transformational leadership: How principals can help reform school cultures*. Paper presented at the annual meeting of the Canadian Association for Curriculum Studies, Victoria, B.C.

Leithwood, K., & Jantzi, D. (2000). The effects of transformational leadership on organizational conditions and student engagement. *Journal of Educational Administration, 38*(2), 112-129.

Leithwood, K., & Montgomery, D. J. (1982). The role of the elementary school principal in program improvement. *Review of Educational Research, 52*(3), 309-339.

Lieberman, A., Falk, B., & Alexander, L. (1994). *A culture in the making: Leadership in learner-centered schools.* New York: Columbia University Teachers College, National Center for Restructuring Education, Schools, and Teaching.

Little, J. W. (1982, Fall). Norm of collegiality and experimentation: Work-place conditions of school success. *American Educational Research Journal, 19*(3), 325-340.

Loucks, S. F., Bauchner, J.E., Crandal, D., Schmidt, W., and Eisman, J. (1982). *Portraits of the Changes, the Players, and the Contexts. A Study of Dissemination Efforts Supporting School Improvement. People, Policies, and Practices: Examining the Chain of School Improvement*. Andover, MA: The Network, Inc.

Louis, K. S., Marks, H., & Kruse, S. (1996). Teachers' professional community in restructuring schools. *American Educational Research Journal, 33*(4), 757-798.

Louis, K. S., & Miles, M. B. (1990). *Improving the urban high school: What works and why.* New York: Teachers College Press.

Manasse, A. L. (1985). Improving conditions for principal effectiveness: Policy implications for research. *The Elementary School Journal, 85*, 439-463.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan Publishing Co.

Murphy, J. (2005). Unpacking the foundations of ISLLC Standards and addressing concerns in the academic community. *Educational Administration Quarterly*, 41, 154-191.

Murphy, J., Elliott, S. N., Goldring, E., & Porter, A. C. (2006). *Learning-Centered Leadership: A Conceptual Foundation*. Nashville, TN: Learning Sciences Institute, Vanderbilt University and The Wallace Foundation.

Murphy, J., & Hallinger, P. (1985, January). Effective high schools: What are the common characteristics? *NASSP Bulletin, 69*(477), 18-22.

Murphy, J., Elliott, S. N., Goldring, E., & Porter A. (2007, April). Leadership for learning: A research-based model and taxonomy of behaviors. School Leadership & Management, 27(2), 179-201.

Murphy, J., Elliott, S. N., Goldring, E., & Porter, A. C. (in press) Leaders for productive schools. International Encyclopedia of Education (3rd ed.). Oxford, Elsevier.

Murphy, J.F., Goldring, E.B., Cravens, X.C., Elliott, S.N., Porter, A.C. (2007, August). The Vanderbilt Assessment of Leadership in Education: Measuring Learning-Centered Leadership. Journal of East China Normal University.

Murphy, J. & Meyers, C. V. (2008). *Turning around failing schools: Leadership lessons from the organizational sciences.* Thousand Oaks, CA: Corwin.

Murphy, K. R., & Deshon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*. *53*(4), 873-900.

National Research Council. (1999). *How people learn: Bridging research and practice*. Washington, DC: National Academy Press.

Newmann, F. M. (1997). How secondary schools contribute to academic success. In K. Borman & B. Schneider (Eds.), *Youth experiences and development: Social influences and educational challenges*. Berkeley, CA: McCutchan.

Newmann, F., & Wehlage, G. (1995). *Successful school restructuring. A report to the public and educators by the Center on Organization and Restructuring of Schools.* Alexandria, VA and Reston, VA: Association for Supervision and Curriculum Development, and the National Association for Secondary School Principals.

Ogden, E. H., & Germinario, V. (1995). *The nation's best schools: Blueprints for excellence*. Lancaster, PA: Technomic.

Porter, A.C., Goldring, E.B., Murphy, J., Elliott, S.N., & Cravens, X. (2006). A framework for the assessment of learning-centered leadership. New York, NY: Wallace Foundation.

Porter, A.C., Goldring, E.B., Elliott, S.N., Murphy, J., Polikoff, M., and Cravens, X. (2008). Setting Performance Standards for the VAL-ED Assessment of Principal Leadership, New York: NY: Wallace Foundation.

Portin, B. S., Feldman, S., & Knapp, M. S. (2006). *Purposes, uses, and practices of leadership assessment in education* Wallace Foundation.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behaviour and Research*, 32(4), 329–353.

Reeves, D. B. (2005). *Assessing educational leaders: Evaluating performance for improved individual and organizational results*. Thousand Oaks, CA: Corwin Press.

Rosenholtz, S. J. (1989). *Teachers' workplace: The social organization of schools.* New York: Longman.

Roueche, J. E., Baker, G. A., Mullin, P. L., & Boy, N. H. O. (1986). *Profiling excellence in America's schools*. Arlington, VA: American Association of School Administrators.

Sebring, P., & Bryk, A. (2000). School leadership and the bottom line in Chicago. *Phi Delta Kappan, 81*, 440-443.

Shaver, A. V., & Walls, R. T. (1998). Effect of Title I parent involvement on student reading and mathematics achievement. *Journal of Research and Development in Education, 31(2), 90-97.*

Sheppard, B. (1996). Exploring the transformational nature of instructional leadership. *Alberta Journal of Educational Research, 42*(4), 325-344.

Stanley, J. C. (1967). General and specific formulas for reliability of differences. *Journal of Educational Measurement*, 4, 249-252.

Teddlie, C., Stringfield, S., Wimpelberg, R., & Kirby, P. (1989). Contextual differences in models for effective schooling in the United States. In B. Creemers, T. Peters, & D. Reynolds (Eds.), *School effectiveness and school improvement: Selected proceedings from the Second Inferential Congress* (pp. 117-130). Amsterdam: Swets and Zeitlinger.

Thomas, D. W., Holdaway, E., & Ward, K. (2000). Policies and practices involved in the evaluation of school principals. *Journal of personnel evaluation in education, 14*(3), 215-240.

Waters, T., & Grubb, S. (2004). *The leadership we need: Using research to strengthen the use of standards for administrator preparation and licensure programs.* Aurora, CO: Mid-continent Research for Education and Learning.

Wellisch, J. B., MacQueen, A. H., Carriere, R. A., & Duck, G. A. (1978, July). School management and organization in successful schools. *Sociology of Education, 51*, 211-226.

Westat and Policy Studies Associates. (2001). *The longitudinal evaluation of change and performance in Title I schools.* Washington, DC: U.S. Department of Education, Office of the Deputy Secretary, Planning and Evaluation Service.

Appendix A. Sample VAL-ED Principal's Response Form

VANDERBILT ASSESSMENT *of* LEADERSHIP *in* EDUCATION™

Principal Response – Form A

Name: _____ **Date:** _____

School District: _____

School: _____ **Years as Principal of this school:** ☐☐

Directions: **The Vanderbilt Assessment of Leadership in Education (VAL-ED) measures the effectiveness of a principal's key leadership behaviors that influence teacher performance and student learning. You will be asked to make effectiveness ratings for each of 72 leadership behaviors based on evidence from the current school year.**

1. **Read each item describing a leadership behavior. In some cases, you may not have actually performed the behavior, but you have ensured that it was done by others in the school. Either way the behavior should be rated.**
2. **Check (✔) the key Sources of Evidence you use for the basis of your assessment. Note, at least one source of evidence must be checked for an item before you make an Effectiveness rating. If you check No Evidence, then Ineffective must be marked in the Effectiveness column.**
3. **If you check any sources of evidence other than No Evidence, always make an effectiveness rating. The number of Sources of Evidence checked is not indicative of the effectiveness rating.**
4. **Mark one Effectiveness Rating circle to indicate how effectively the behavior was performed.**

   Outstandingly effective means you (or your designee) has carried out a particular behavior (e.g., providing necessary support) with a very strong, positive effect on the targeted area of school activity (e.g., rigorous curriculum).

   Ineffective means you (or your designee) has either not done the particular behavior (e.g., not provided necessary support) or has carried out the behavior with very low quality that does not have a positive effect on the targeted area of school activity (e.g., rigorous curriculum).

Completion Tips:

- **Review the VAL-ED Conceptual Framework to see how the core components and key processes assessed provide a comprehensive picture of leadership behaviors.**
- **Definitions of key leadership behaviors are provided in the VAL-ED Glossary.**
- **Most respondents take 20 minutes to complete all items. You should try to complete the evaluation in one sitting.**

Review the completed example items below before starting the evaluation.

Leadership Behaviors	Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
	Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …											
1. plans for a culture of learning that serves all students.	☐	☒	☐	☒	☐	☐	⊗ 1	○ 2	○ 3	○ 4	○ 5
2. evaluates the rigor of the curriculum.	☐	☐	☒	☐	☐	☐	○ 1	○ 2	⊗ 3	○ 4	○ 5

- For Item #1, which states "How effective is the principal at ensuring the school plans for a culture of learning that serves all students," the respondent checked two sources of evidence for the basis of her evaluation of effectiveness and then checked one effectiveness category to indicate she perceived the principal as being *ineffective* regarding this leadership behavior.
- For Item #2, which states "How effective is the principal at ensuring the school evaluates the rigor of the curriculum," the respondent checked one source of evidence for the basis of her evaluation and then checked one effectiveness category to indicate she perceived the principal as being *satisfactorily effective* regarding this leadership behavior.

This table represents the Conceptual Framework of this assessment, where each cell represents the cross-section of one core component and one key process of principal leadership. Every item of the assessment represents a cross-section of one core component and one key process.

Core Components	Key Processes						
	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
High Standards for Student Learning						
Rigorous Curriculum (content)						
Quality Instruction (pedagogy)						
Culture of Learning & Professional Behavior						
Connections to External Communities						
Performance Accountability						

The following provides definitions for the core components and key processes of principal leadership.

Core Components of School Performance

High Standards for Student Learning—There are individual, team, and school goals for rigorous student academic and social learning.

Rigorous Curriculum (content)—There is ambitious academic content provided to all students in core academic subjects.

Quality Instruction (pedagogy)—There are effective instructional practices that maximize student academic and social learning.

Culture of Learning & Professional Behavior—There are integrated communities of professional practice in the service of student academic and social learning. There is a healthy school environment in which student learning is the central focus.

Connections to External Communities—There are linkages to family and/or other people and institutions in the community that advance academic and social learning.

Performance Accountability—Leadership holds itself and others responsible for realizing high standards of performance for student academic and social learning. There is individual and collective responsibility among the professional staff and students.

Key Processes of Leadership

Planning—Articulate shared direction and coherent policies, practices, and procedures for realizing high standards of student performance.

Implementing—Engage people, ideas, and resources to put into practice the activities necessary to realize high standards for student performance.

Supporting—Create enabling conditions; secure and use the financial, political, technological, and human resources necessary to promote academic and social learning.

Advocating—Promotes the diverse needs of students within and beyond the school.

Communicating—Develop, utilize, and maintain systems of exchange among members of the school and with its external communities.

Monitoring—Systematically collect and analyze data to make judgments that guide decisions and actions for continuous improvement.

High Standards for Student Learning		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Planning	1. plans rigorous growth targets in learning for all students.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	2. plans targets of faculty performance that emphasize improvement in student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Implementing	3. creates buy-in among faculty for actions required to promote high standards of learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	4. creates expectations that faculty maintain high standards for student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Supporting	5. encourages students to successfully achieve rigorous goals for student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	6. supports teachers in meeting school goals.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

High Standards for Student Learning		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Advocating	7. advocates for high standards for student learning when writing and implementing Individualized Education Plans (IEPs).	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	8. challenges low expectations for students with special needs.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Communicating	9. communicates rigorous goals for student learning to faculty.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	10. communicates with families and the community about goals for rigorous student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Monitoring	11. monitors student learning against high standards of achievement.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	12. monitors disaggregated test results.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Rigorous Curriculum		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Planning	13. develops a rigorous curriculum for all students.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	14. plans access to rigorous curricula for students with special needs.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Implementing	15. creates rigorous sequences of learning experiences/courses.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	16. implements a rigorous curriculum in all classes.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Supporting	17. secures the teaching materials necessary for a rigorous curriculum.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	18. supports teachers to teach a curriculum consistent with state and national content standards.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Rigorous Curriculum		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Advocating	19. advocates a rigorous curriculum that honors the diversity of students and their families.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	20. challenges faculty to teach a rigorous curriculum to students at risk of failure.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Communicating	21. discusses state curriculum frameworks.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	22. discusses the importance of addressing the same academic content in special and regular programs.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Monitoring	23. evaluates the extent to which all students complete a rigorous curricular program.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	24. evaluates the rigor of the curriculum.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Quality Instruction		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Planning	25. plans instructional services for students with special needs using assessment data.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	26. plans a schedule that enables quality instruction.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Implementing	27. coordinates efforts to improve instruction in all classes.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	28. recruits teachers with the expertise to deliver instruction that maximizes student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Supporting	29. supports collaboration among faculty to improve instruction that maximizes student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	30. supports teachers' opportunities to improve their instructional practices.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Quality Instruction		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Advocating	31. advocates for all students to regularly experience effective instruction.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	32. advocates opportunities for high quality instruction beyond the regular school day and school year.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Communicating	33. discusses instructional practices during faculty meetings.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	34. communicates with faculty about removing barriers that prevent students from experiencing quality instruction.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Monitoring	35. evaluates how instructional time is used.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	36. evaluates teachers' instructional practices.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Culture of Learning and Professional Behavior		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Planning	37. plans programs and policies that promote discipline and order.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	38. plans for a positive environment in which student learning is the central focus.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Implementing	39. implements a learning environment in which all students are known and cared for.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	40. builds a culture that honors academic achievement.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Supporting	41. allocates resources to build a culture focused on student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	42. supports collaborative teams to improve instruction.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Culture of Learning and Professional Behavior		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Advocating	43. advocates a culture of learning that respects diversity of students.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	44. advocates for students to be involved in the school community.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Communicating	45. communicates with parents about the aspects of a positive school culture.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	46. discusses standards of professional behavior with faculty.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Monitoring	47. monitors the participation of every student in social and academic activities.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	48. assesses the culture of the school from students' perspectives.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Connections to External Communities		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Planning	49. develops a plan for school/community relations that revolves around the academic mission.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	50. develops a plan for community outreach programs consistent with instructional goals.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Implementing	51. implements programs to help address community needs.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	52. builds business partnerships to support social and academic learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Supporting	53. secures additional resources through partnering with external agencies to enhance teaching and learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	54. allocates resources that build family and community partnerships to advance student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Connections to External Communities		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Advocating	55. promotes mechanisms for reaching families who are least comfortable at school.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	56. challenges teachers to work with community agencies to support students with low achievement.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Communicating	57. listens to feedback from the community.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	58. listens to the diverse opinions and needs of all families.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Monitoring	59. collects information to learn about resources and assets in the community.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	60. monitors the effectiveness of community-school connections.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Performance Accountability		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Mark One Circle to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Planning	61. develops a plan for individual and collective accountability among faculty for student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	62. develops a plan emphasizing accountability to stakeholders for student academic and social learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Implementing	63. uses faculty input to create methods to hold faculty accountable.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	64. implements social and academic accountability equitably for all students.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Supporting	65. allocates time to evaluate student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	66. allocates time to evaluate faculty for student learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Performance Accountability		Sources of Evidence (Check Key Sources of Evidence)						Effectiveness Rating (Circle One Number to Indicate How Effective)				
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective
How effective is the principal at ensuring the school …												
Advocating	67. challenges faculty who attribute student failure to others.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	68. advocates that all students are accountable for achieving high levels of performance in both academic and social learning.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Communicating	69. discusses progress toward meeting school goals with parents.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	70. communicates to faculty how accountability results will be used for school improvement.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
Monitoring	71. analyzes the influence of faculty evaluations on the rigor of the curriculum.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5
	72. monitors the accuracy and appropriateness of data used for student accountability.	☐	☐	☐	☐	☐	☐	○ 1	○ 2	○ 3	○ 4	○ 5

Questions for VAL-Ed Users

	Strongly Disagree	Disagree	Agree	Strongly Agree
1. I found this response form easy to use.	○	○	○	○
2. I understood the vast majority of the items	○	○	○	○
3. I believe the vast majority of the items focus on important leadership behaviors.	○	○	○	○
4. I do not believe the items are biased against any race or gender of a principal being assessed.	○	○	○	○
5. This assessment is appropriate for use at the elementary, middle, and high school levels.	○	○	○	○
6. I would prefer a web-based format for this assessment over the paper-and-pencil version I just completed.	○	○	○	○
7. Teachers should have input into the assessment of their principal's leadership.	○	○	○	○
8. I would support the use of this assessment instrument to hold principals accountable in my district.	○	○	○	○
9. The amount of time required to complete this instrument is reasonable.	○	○	○	○

Comments/Suggestions for Improvement:

Thank you for your participation.

Appendix B. Sample VAL-ED Multi-Rater Report for Principal

VANDERBILT ASSESSMENT of LEADERSHIP in EDUCATION

Principal Report

Survey ID:	1067162	**Date of Report:**	November 03, 2008
School District:	District ABC	**Date of Evaluation:**	N/A
School:	ABC School	**VAL-ED Form:**	A

Purpose of the Assessment

The Vanderbilt Assessment of Leadership in Education or VAL-ED is designed to provide a summary of effectiveness of a principal's learning-centered leadership behaviors during the current school year. A comprehensive picture has emerged and is reported with input from teachers, the principal's supervisor and his or her own self-report.

The VAL-ED focuses on leadership behaviors defined by six core components and six key processes known to influence student achievement:

Core Components

- High Standards for Student Learning
- Rigorous Curriculum
- Quality Instruction
- Culture of Learning & Professional Behavior
- Connections to External Communities
- Performance Accountability

Key Processes

- Planning
- Implementing
- Supporting
- Advocating
- Communicating
- Monitoring

Respondents to the VAL-ED were asked: How effective the principal is at ensuring the school carries out specific actions that affect core components of learning-centered leadership. The effectiveness ratings, based on evidence, range from 1 (ineffective) to 5 (outstandingly effective) for each of 72 leadership behaviors.

This VAL-ED report addresses the questions of:
(1) who responded?
(2) what evidence was used to evaluate the principal?
(3) what do the results say about the principal's current leadership behaviors?

The results are interpreted against both norm-referenced and standards-referenced criteria that highlight areas of strengths and possible areas for improvement. A leadership development plan can be developed based on these results.

The VAL-ED provides technically sound scores when used as designed, however, it is recommended that it be used along with other information when making important evaluative decisions.

For more information about the VAL-ED please visit our website:
http://www.thinklinkassessment.com/corporate/valed.html.

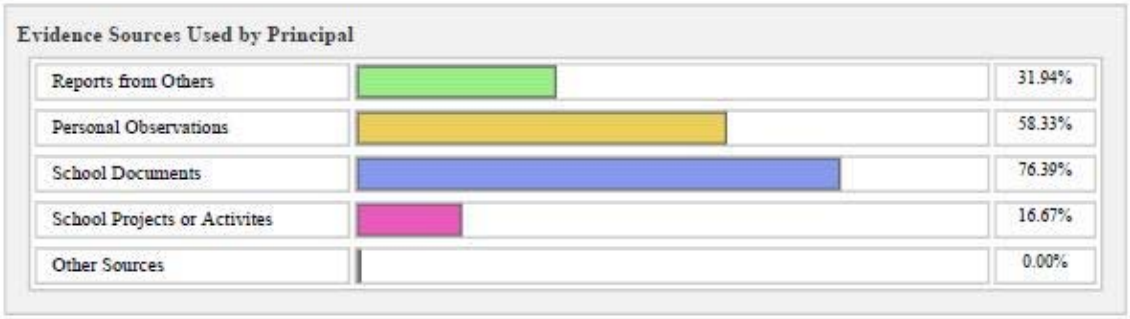
Who Responded and What Evidence Did They Use?

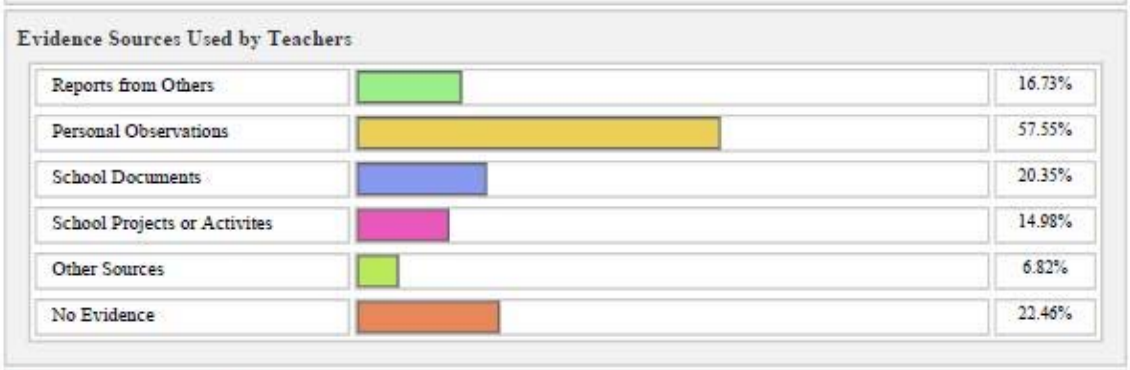
	Possible Respondents	Actual Respondents	Percent (%) Responding
Principal	1	1	100 %
Teachers	23	23	100 %
Supervisor	1	1	100 %

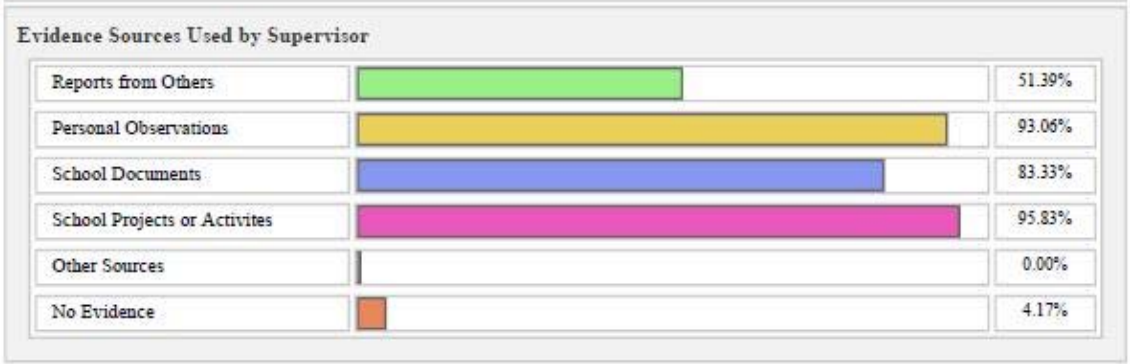
A response rate of greater than or equal to 75% is high, 50% to 74% is moderate, and below 50% is low. When response rates are low, resulting scores should be interpreted with caution.

Sources of Evidence

Ratings of a principal's behaviors should be based on evidence that is recent, relevant and representative. Evidence comes in many forms (e.g., observations of behavior, review of documents that record leadership actions and communications with people who have directly observed the principal's behavior). After reflecting on a sample of evidence, respondents effectiveness ratings of leadership behaviors are behaviorally-anchored and more accurate. The graphs below summarize each type of evidence used as a basis for their effectiveness ratings of the leadership behaviors. The bars display the sources of evidence for each item used by the principal, all teacher and supervisor respondents in the school.







What are the Results of the Assessment?

VAL-ED provides a total score across all respondents as well as separately by respondent group. The scores from the teachers are all based on the average across all teacher respondents. The total score, core component, and key process effectiveness ratings are interpreted against a national representative sample that included principals, supervisors, and teachers, providing a **percentile rank**. The results are also interpreted against a set of performance standards ranging from **Below Basic** to **Distinguished**. The scores associated with performance levels were determined by a national panel of principals, supervisors and teachers.

Below Basic	Basic	Proficient	Distinguished
A leader at the below basic level of proficiency exhibits learning-centered leadership behaviors at levels of effectiveness that are unlikely to influence teachers positively nor result in acceptable value-added to student achievement and social learning for students.	A leader at the basic level of proficiency exhibits learning-centered leadership behaviors at levels of effectiveness that are likely to influence teachers positively and that result in acceptable value-added to student achievement and social learning for some sub-groups of students, but not all.	A proficient leader exhibits learning-centered leadership behaviors at levels of effectiveness that are likely to influence teachers positively and result in acceptable value-added to student achievement and social learning for all students.	A distinguished leader exhibits learning-centered leadership behaviors at levels of effectiveness that are virtually certain to influence teachers positively and result in strong value-added to student achievement and social learning for all students.

Overview of Assessment Results

The Principal's Overall Total Effectiveness score based on the averaged ratings of all respondents is 3.54. Remember, this score is based on a 5-point effectiveness scale where 1=Ineffective; 2=Minimally Effective; 3=Satisfactorily Effective; 4=Highly Effective; 5=Outstandingly Effective. The Performance Level and national Percentile Rank for this score are documented in the table below.

Overall Effectiveness Score		
Mean Score	Performance Level	Percentile Rank
3.54	Basic	42.4
The standard error of measurement is .05		

Summary of Core Components Scores			
	Mean	Performance Level	Percentile Rank
High Standards for Student Learning	3.73	Proficient	54.7
Rigorous Curriculum	3.51	Basic	40.1
Quality Instruction	3.61	Proficient	40.1
Culture of Learning & Professional Behavior	3.72	Proficient	46.9
Connections to External Communities	3.22	Below Basic	27.3
Performance Accountability	3.43	Basic	44.2

Summary of Key Processes Scores			
	Mean	Performance Level	Percentile Rank
Planning	3.60	Proficient	55.1
Implementing	3.60	Basic	51.5
Supporting	3.54	Basic	27.3
Advocating	3.28	Below Basic	20
Communicating	3.58	Basic	43.7
Monitoring	3.66	Proficient	57.9

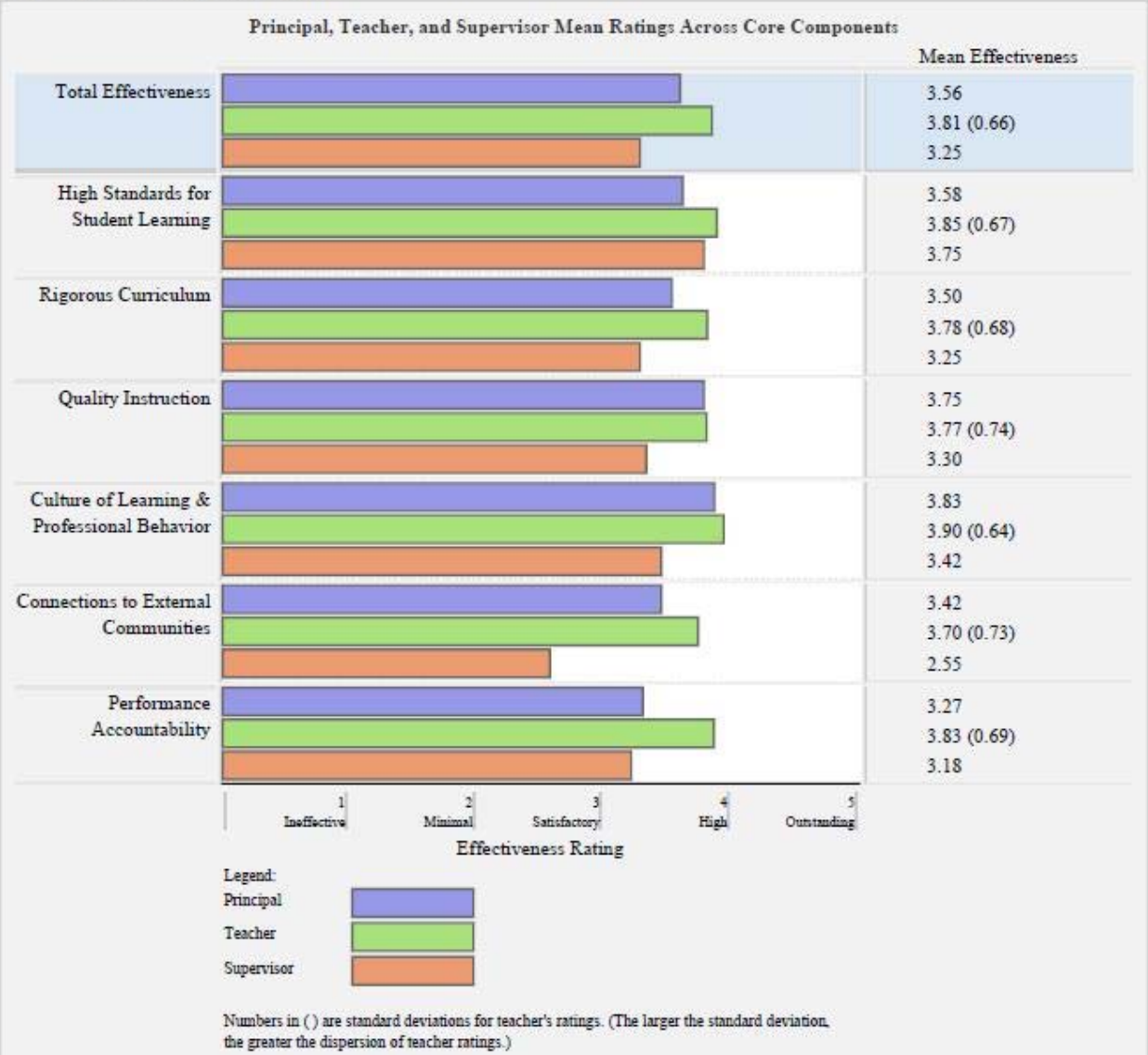
An examination of the principal's Core Components mean item ratings ranged from a low of 3.22 for Connections to External Communities to a high of 3.73 for High Standards for Student Learning. Similarly the principal's Key Processes mean item ratings indicates they ranged from a low of 3.28 for Advocating to a high of 3.66 for Monitoring.

Assessment Profile and Respondent Comparisons

The principal's relative strengths and areas for development can be determined by comparing scores for each of the 6 Core Components and 6 Key Processes across different respondent groups. The next two graphs present an integrated visual summary of the results. They show the **Mean Effectiveness** associated with each Core Component and Key Process.

First, examine the profiles as recorded by each of the three respondent groups. These scores can be interpreted by

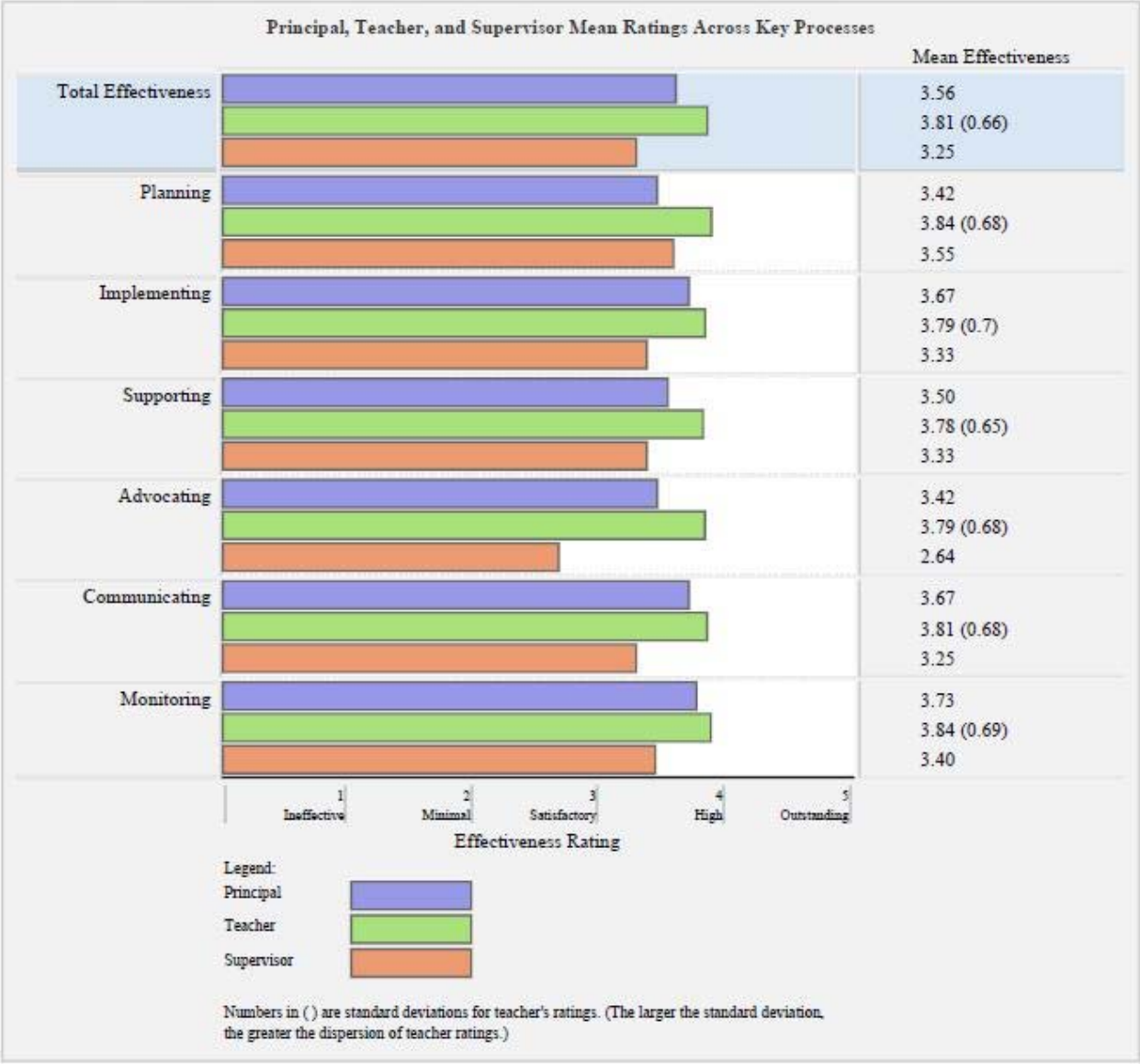
(a) Comparisons among Core Components and Key Processes
(b) Examination of scores among respondent groups
(c) Comparisons to the mean effectiveness scale
(d) Reference to national percentile ranks



For each of the six Core Components in the graph, examine the effectiveness ratings. The ratings on 12 items focus on a given Core Component. The higher the ratings, the more effective the leadership behaviors of the principal. When there are large differences between respondent groups, the focus should be on the results for each respondent group rather than the overall effectiveness score.

Assessment Profile and Respondent Comparisons (Cont'd.)

The ratings of the six Key Processes are based on 12 items that focus on a given Key Process. Again, the higher the score, the more effective the leadership behaviors of the principal. For more details about the technical aspects of the VAL-ED scores and tips on interpreting scores, visit the VAL-ED website http://www.thinklinkassessment.com/corporate/valed.html



Using Results to Plan for Profesional Growth

The matrix below provides an integrated summary of the principal's relative strengths and areas for growth based on the mean item scores for the intersection of Core Components by Key Processes across the three respondent groups.

- Cells that are green represent areas of behavior that are 'proficient' or 'distinguished'.
- Cells that are yellow represent areas of behavior that are 'basic'.
- Cells that are red represent areas of behavior that are 'below basic'.

The leadership behaviors listed in each cluster on the following pages are representative of the lowest rated core component by key process areas of behavior. If no behavior clusters are provided it indicates the principal's current learning-centered leadership behaviors are considered acceptable.

The behaviors on each page that are **boldface** type are those that were actually assessed in the evaluation. The other behaviors represent the entire pool of VAL-ED behaviors for each core component by key process. All of these behaviors are relevant targets for improvement.

For a list of all the leadership behaviors associated with each core component area, consult the VAL-ED Users' Guide.

Leadership Behaviors for Possible Improvement

Connections to External Communities X Monitoring

- Analyzes data about parental involvement.
- Uses data to make decisions about community engagement.
- **Monitors the effectiveness of community school connections.**
- Uses data on parent involvement in teacher evaluations. (Removed after 9-school pilot)
- Evaluates the effectiveness of its partnerships with the community in advancing academic and social learning.
- Collects information about the needs and interests of parents.
- **Collects information to learn about resources and assets in the community.**

Connections to External Communities X Supporting

- Supports teachers to work with community agencies on behalf of students.
- **Secures additional resources through partnering with external agencies to enhance teaching and learning.**
- Secures technology from the district and/or the community to enhance teaching and learning.
- Secures resources to support school-community relationships.
- **Allocates resources that build family and community partnerships to advance student learning.**
- Motivates teachers to be responsive to all families.

Performance Accountability X Advocating

- Advocates that leaders are accountable for meeting the needs of diverse students.
- **Advocates that all students are accountable for achieving high levels of performance in both academic and social learning.**
- Advocates that the faculty is accountable for meeting the needs of diverse students.
- Promotes an accountability system that represents the diverse views of families and the community.
- **Challenges faculty who attribute student failure to others.**
- Advocates for shared accountability by faculty for student academic and social learning.

Leadership Behaviors for Possible Improvement

Rigorous Curriculum X Advocating
• Challenges all students to complete a rigorous, academically focused program of study. • **Challenges faculty to teach a rigorous curriculum to students at risk of failure.** • Advocates that all programs for students with special needs deliver a rigorous curriculum. • **Advocates rigorous curriculum that honors the diversity of students and their families.** • Promotes the importance of a rigorous curriculum to students of all ability levels. • Advocates for families to learn about the curricular program.

Connections to External Communities X Implementing
• **Builds business partnerships to support social and academic learning.** • Implements programs to involve families in the educational mission. • **Implements programs to help address community needs.** • Builds a positive, open relationship with the community. • Coordinates access to social service agencies to support students. • Implements programs to help parents assist their children to be successful in school.

Connections to External Communities X Planning
• Plans family education programs consistent with instructional goals. • Plans for the use of external community resources to promote academic and social learning goals. • **Develops a plan for community outreach programs consistent with instructional goals.** • Plans activities with volunteers to advance social and academic goals. • Plans activities to engage families in student learning. • **Develops a plan for school/community relations that revolves around the academic mission.**

About the VAL-ED

The Vanderbilt Assessment of Leadership in Education (VAL-ED) is conceptually and theoretically grounded and its resulting scores are reliable and valid for purposes of evaluating learning-centered leadership.

The VAL-ED uses 360 degree feedback from teachers, principals, and supervisors.

Content focuses on learning-centered leadership behaviors that influence teachers and staff, and in turn are related to increases in student achievement.

Assessment is of leadership behaviors, not knowledge, dispositions, or personal characteristics of leadership.

The VAL-ED requires respondents to identify evidence on which they are basing their assessment of principal behaviors.

The psychometric properties of the VAL-ED are clearly documented. Information on norms, standards, and uses are available through a comprehensive technical manual.

"Leadership is a central ingredient -
often the keystone element in school and district success as defined in terms of student achievement."

- Joseph Murphy
Vanderbilt University

"Assessments that provide ongoing performance feedback to school leaders about their learning-centered leadership behaviors can substantially help school leaders develop effective leadership for school improvement."

- Ellen Goldring
Vanderbilt University

Visit

http://www.vanderbilt.edu/lsi/valed

for more information and periodic updates on research and related articles on the use of the VAL-ED.

VAL-ED Authors
Andrew Porter, Joseph Murphy,
Ellen Goldring, & Stephen N. Elliott