



Investigating the Validity and Reliability of the Vanderbilt Assessment of Leadership in Education

Author(s): Andrew C. Porter, Morgan S. Polikoff, Ellen B. Goldring, Joseph Murphy, Stephen N. Elliott, Henry May

Source: *The Elementary School Journal*, Vol. 111, No. 2, The Conceptualization, Measurement, and Effects of School Leadership (December 2010), pp. 282-313

Published by: [The University of Chicago Press](http://www.press.uchicago.edu)

Stable URL: <http://www.jstor.org/stable/10.1086/656301>

Accessed: 23/06/2011 11:17

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=ucpress>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



The University of Chicago Press is collaborating with JSTOR to digitize, preserve and extend access to *The Elementary School Journal*.

INVESTIGATING THE VALIDITY AND RELIABILITY OF THE VANDERBILT ASSESSMENT OF LEADERSHIP IN EDUCATION

ABSTRACT

The Vanderbilt Assessment of Leadership in Education (VAL-ED) is a multirater assessment of principals' learning-centered leadership. The instrument was developed based on the Standards for Educational and Psychological Testing. In this article, we report on the validity and reliability evidence for the VAL-ED accumulated in a national field trial. Using data from more than 270 schools across a wide range of settings, we found that the instrument was reliable for measuring principals' learning-centered leadership. Furthermore, there was mostly positive evidence that the VAL-ED distinguished among subscales of principal performance. In conjunction with findings from the VAL-ED development phase, these results support the conclusion that the VAL-ED can be used by K–12 schools to assess learning-centered leadership.

Andrew C. Porter

UNIVERSITY OF
PENNSYLVANIA

Morgan S. Polikoff

UNIVERSITY OF
SOUTHERN CALIFORNIA

Ellen B. Goldring

Joseph Murphy

Stephen N. Elliott

VANDERBILT UNIVERSITY

Henry May

UNIVERSITY OF
PENNSYLVANIA

LEADERSHIP is a central ingredient in school success defined in terms of value added to student achievement (Leithwood, Louis, Anderson, & Wahlstrom, 2004; Marzano, Waters, & McNulty, 2005; Murphy & Hallinger, 1985). There have been a number of recent efforts to improve the quality of principal leadership, including the establishment of the Interstate Leadership Licensure Consortium (ISLLC) standards; growth in principal professional development, coaching, and mentoring; and training program accreditation (the National Council for Accreditation of Teacher Education process). However, there has been relatively little focus in research and practice on the assessment of principal leadership effec-

tiveness. When there has been movement toward leadership assessment, it has been in the form of assessment for certification (e.g., Education Testing Service's preservice assessments for principals), or it has been with instruments with few, if any, documented psychometric properties and little research support (Goldring, Cravens, et al., 2009).

Against this backdrop, a team of researchers from Vanderbilt University and the University of Pennsylvania began development and validation of the Vanderbilt Assessment of Leadership in Education (VAL-ED), a multirater, evidence-based assessment of principals' instructional leadership. The 3-year process began with development of a test framework (Porter, Goldring, Murphy, Elliott, & Cravens, 2006). Next, a series of qualitative and quantitative development studies were conducted (Porter et al., 2010). Finally, teachers, principals, and supervisors were engaged in a national field trial of the VAL-ED. More than 270 schools—elementary, middle, and high; suburban, urban, and rural; and from all four regions of the country—participated.

The purposes of the national field trial were several, but the primary purpose was to provide data with which to evaluate the VAL-ED against commonly accepted standards of test development (American Educational Research Association [AERA], American Psychological Association [APA], & National Council of Measurement in Education [NCME], 1999). In what follows, we summarize the results of the field trial, with a focus on validity, reliability, scales, norms, and form comparability. First, we briefly review the conceptual and theoretical foundation for the VAL-ED and sketch the development process. Then, we describe the national field trial and our findings regarding the psychometric properties of the VAL-ED. Specifically, we address the following four questions:

1. What evidence is there for the reliability of the score interpretations on the VAL-ED?
2. What evidence is there for the validity of the VAL-ED?
3. To what extent do the VAL-ED's items exhibit bias?
4. What evidence is there to support the VAL-ED's norms, performance standards, and the parallelness of its two forms?

Together with previous analyses, the results described here suggest that the VAL-ED can be used as a valid indicator of principals' learning-centered leadership behaviors in urban, suburban, and rural public elementary, middle, and high schools in all regions of the country.

Background

Before describing the VAL-ED, it is important to understand the instrument in context with the other principal leadership assessments currently in use. Our previous analysis of the field (Goldring, Cravens, et al., 2009) found 86 publicly available assessments currently in use by states or districts to evaluate principal leadership, 44 of which were analyzed for content and usage, and 42 of which were analyzed for content only. Of these, nine were from states or the District of Columbia, and 77 were from districts. Analyses revealed wide variation in the instruments' content and quality. For instance, the proportion of the instrument focusing on instruction var-

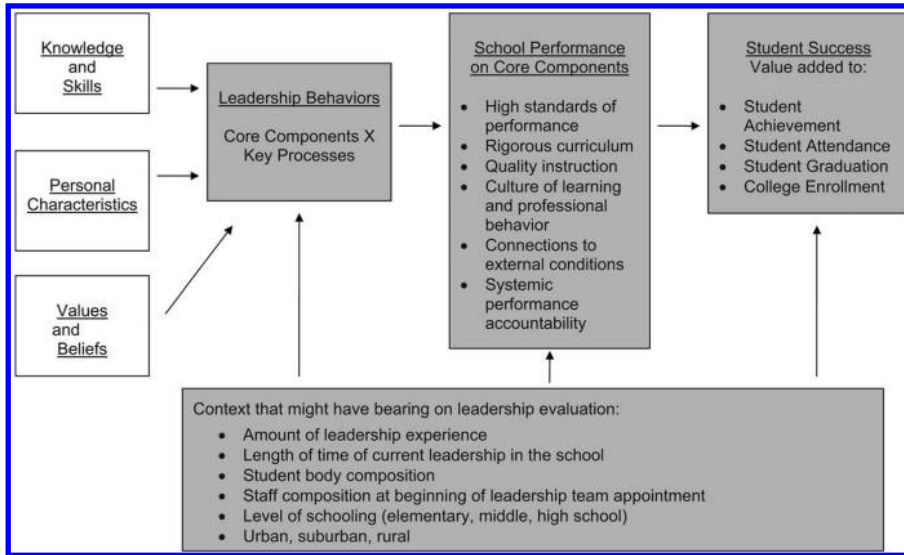


Figure 1. VAL-ED conceptual model.

ied from 23% to 85% across instruments. Furthermore, the instruments were poorly aligned to the ISLLC standards. Finally, just two of the instruments provided any psychometric evidence.

The VAL-ED Framework

The conceptual model that guided the development of the VAL-ED is shown in Figure 1 (for more details on the conceptual framework, see Goldring, Porter, Murphy, Elliott, & Cravens, 2009). Principals' backgrounds, along with their contexts, shape their leadership behaviors. Along with context, these behaviors affect their schools' performance in terms of, for instance, the quality of the school's instruction and the nature of its relationship with the external community. Finally, context and the school's performance on the core components affect student success, leading to value added to student academic and social learning.

The VAL-ED is focused on the second box from the left—the principal's learning-centered leadership behaviors. While the VAL-ED could have focused on the principal's attitudes and beliefs, research indicates that the principal's behaviors affect school processes and ultimately student learning (Leithwood et al., 2004; Marzano et al., 2005). Ultimately, a larger assessment system is envisioned, with the VAL-ED in the second box, but also direct measures of core components, student outcomes, and key processes and a rigorous accounting of contextual effects.

Focusing on the second box in Figure 1, we have the VAL-ED's conceptual framework (Fig. 2). This six-by-six matrix was used to guide the development of the VAL-ED and its items. While a full presentation of the research support for the conceptual framework is beyond the scope of this article, a summary of the framework is presented here. The framework contains core components—the features of schools that make them effective. These are High Standards for Student Learning, Rigorous Curriculum, Quality Instruction, Culture of Learning and Professional Behavior, Connections to External Communities, and Performance Accountability. Intersecting with the core components are key processes—the ways in which leaders

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

Figure 2. VAL-ED conceptual framework.

enact the core components in their schools. These are Planning, Implementing, Supporting, Advocating, Communicating, and Monitoring. The 36 cells at the intersection of core components and key processes define the learning-centered leadership behaviors that affect the quality of education provided by the school.

High Standards for Student Learning refers to the extent to which there are individual and team goals for student academic and social learning, reflecting the importance of establishing the school’s purpose (Hallinger & Heck, 2002) and having high expectations for students (Betts & Grogger, 2003; Purkey & Smith, 1983). Rigorous Curriculum is the extent to which ambitious academic content is provided to all students in core academic subjects, underlying the notion supported by standards-based reform that coverage of challenging academic content leads to improvements in student learning (Brophy & Good, 1986; Newmann & Wehlage, 1995). Also important is the delivery of the curriculum through Quality Instruction, effective instructional practices that maximize student academic and social learning. While the specific features of instruction that maximize student learning are not agreed upon, there is consistent evidence of the importance of such strategies as developing students’ depth of understanding (Newmann & Wehlage, 1995) and working with students’ previous knowledge and beliefs (National Research Council, 1999). The fourth core component, Culture of Learning and Professional Behavior, is the extent to which there are integrated communities of professional practice in the service of student academic and social learning, as well as a healthy school environment in which student learning is the central focus, emphasizing the importance of professional community (Louis, Marks, & Kruse, 1996) and a prosocial school environment (Rutter, Maughan, Mortimore, & Ouston, 1979). Successful schools also have strong Connections to External Communities, or linkages to families and/or other people and institutions in the community that advance academic and social learning (Henderson & Mapp, 2002). The final core component is Performance Accountability, that is, individual and collective responsibility among leadership, faculty, and students for achieving the rigorous student academic and social learning goals. There is evidence supporting the role of both internal (Bryk & Schneider, 2002) and external accountability (Carnoy & Loeb, 2004; Hanushek & Raymond, 2005) in influencing school success in terms of student learning.

High Standards for Student Learning		Sources of Evidence Check Key Sources of Evidence					Effectiveness Rating Mark One Circle to Indicate How Effective or Check DK						
		Reports from Others	Personal Observations	School Documents	School Projects or Activities	Other Sources	No Evidence	Ineffective	Minimally Effective	Satisfactorily Effective	Highly Effective	Outstandingly Effective	Don't Know
How effective is the principal at ensuring the school ...													
Planning	1. plans rigorous growth targets in learning for all students.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
	2. plans targets of faculty performance that emphasize improvement in student learning.	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 3. Sample VAL-ED items.

The key processes define the ways in which school leaders bring about the core components and support the notion that school improvement is a continual process (Fullan, 1982). The definitions of the key processes are:

- Planning: articulating shared direction and coherent policies, practices, and procedures for realizing high standards of student performance
- Implementing: putting into practice the activities necessary to realize high standards for student performance
- Supporting: creating enabling conditions; securing and using the financial, political, technological, and human resources necessary to promote academic and social learning
- Advocating: promoting the diverse needs of students within and beyond the school
- Communicating: developing, utilizing, and maintaining systems of exchange among members of the school and with its external communities
- Monitoring: systematically collecting and analyzing data to make judgments that guide decisions and actions for continuous improvement

Together, the six key processes outline the “how” of principal instructional leadership, while the core components identify the “what” that define the targets of effective leadership (Goldring, Porter, et al., 2009).

The VAL-ED Instrument

The VAL-ED is a multirater assessment of principals’ learning-centered leadership behaviors. The principal, his/her supervisor, and all the teachers in the school complete it (sometimes referred to as a 360 assessment). For each of the 72 items on the VAL-ED (see Fig. 3 for sample items), the respondent first checks the evidence on which an effectiveness rating is made: “reports from others,” “personal observations,” “school documents,” “school projects or activities,” “other sources,” or “no evidence.” Next, the respondent rates the principal’s effectiveness on a scale of 1 = ineffective, 2 = minimally effective, 3 = satisfactorily effective, 4 = highly effective, 5 = outstandingly effective. Teachers and supervisors can also select “don’t know” for any item; principals are not allowed to select “don’t know” because they should

know if they engaged in a particular behavior. There are additional rules for ratings: principals selecting “no evidence” must mark “ineffective” for the effectiveness rating, and teachers and supervisors selecting “no evidence” may mark either “ineffective” or “don’t know” for their effectiveness rating. Effectiveness ratings are only to be given on items for which there is some evidence. Invalid responses are treated as missing data and the items are ignored in the calculation of scale and total score means (i.e., the items are removed from the numerator and the denominator). Two parallel forms of the test were created with cell-by-cell randomized equivalent sets of items, Forms A and C.

Calculating scale and total score means for individual raters requires taking the mean item response for each item answered validly. Next, the scale or total score means for teachers are averaged (e.g., averaging the Planning scores for each teacher respondent to obtain a teacher rating for Planning). Finally, the scale and total score means for the three respondent groups are averaged; each group is weighted equally. All scores are on the 1–5 effectiveness scale. Core component and key process subscores are each based on responses to the same 72 items—each item contributes to both a core component and a key process. Results are reported in terms of mean item response for the total score and for each of the 12 subscales. Separate scores are provided for the aggregate, as well as for the three respondent groups. While the extant literature suggests differential effects of leadership types on student outcomes (Robinson, Lloyd, & Rowe, 2008), the current VAL-ED reporting system weights all items equally in the calculation of total scores. The literature is not sufficient to support differentiated weighting, and weighting would surely vary depending upon the operationalization of an external criterion. In addition to raw scores, percentile ranks and performance levels (e.g., proficient, distinguished) are reported for total score and for each subscale.

Development and Early Validation

The field test described in this article followed multiple years of development and testing of the instrument. The primary purpose of the early work was to build what promised to be a valid and reliable instrument for use in the national field test. Full details on the studies and the development process are available elsewhere (Porter et al., 2010). The following is a brief summary of the instrument development and initial testing phase of the project.

The development of the VAL-ED consisted of seven steps. The first two steps were the creation of the VAL-ED framework, described above, and the creation of the VAL-ED instrument. These took place over 1 year, as team members used the literature on effective principal instructional leadership and employed an iterative item-writing and selection process designed to build content validity. The third step was an item-sorting study with nine principals designed to improve item fit to the conceptual framework. Next, a series of cognitive interviews (Desimone & LeFloch, 2004) was conducted to solve lingering content problems and clarify instructions. Seventeen respondents participated, including teachers, principals, and supervisors from five urban districts. Then, a qualitative item bias study was conducted based on the test fairness guidelines published by the Educational Testing Service (2000), with nine respondents, including teachers, administrators, and behavior specialists. Finally, two small-scale pilot studies were conducted, one in nine schools in one urban

district, and one in 11 schools in one Midwestern state. Based on the findings of the various studies, several important changes to the instrument were made. These included the shortening of the instrument from 108 items to 72, the changing of the rating scale to the current version, and the alteration of four items to remove bias concerns.

The iterative test development phase was based on principles of best practice in test construction. The purpose was to construct a final version of the instrument for validation in the national field trial. By the 11-school pilot, the instrument was well prepared for such a large-scale evaluation.

Method and Data

In the spring of 2008, we undertook a nationally representative field trial of the VAL-ED. The primary purposes of the field trial were several: (a) to provide data with which to examine the validity and reliability of the VAL-ED, (b) to provide data for use in a standard-setting activity to establish the VAL-ED's performance standards, (c) to provide data to establish initial norms, and (d) to confirm that the two forms of the VAL-ED were parallel.

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 U.S. Census), of which 100 were to be elementary, 100 middle, and 100 high schools. There were to be 150 urban, 100 suburban, and 50 rural schools. In addition, the sample design called for the 150 urban schools to include 50 drawn from Wallace Foundation 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. Wallace districts played a special role in the sample design of the study because of the funding provided by the Wallace Foundation and the foundation's emphasis on effective principal leadership. Once a district had agreed to participate, schools in the district were selected by 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 by simple random sample.

Ninety-nine districts were contacted and 60 agreed to participate for a 61% response rate. By region, 18 of 23 participated from the Northeast (78%), 16 of 23 from the South (70%), 14 of 28 from the Midwest (50%), and 12 of 25 from the West (48%). By urbanicity, 27 of 51 suburban districts participated (53%), 20 of 32 rural districts (63%), and 13 of 16 urban districts (81%). Across the 60 participating districts, 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. Based on the 245 schools from which at least some teacher data were returned, teacher response was variable, with the median teacher response rate across partici-

pating schools 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.

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, matched 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 prevalent than targeted.

The obtained sample of schools is not representative of schools in the nation for several reasons. First, elementary, middle, and high schools were designed to be equally represented and this goal was nearly achieved, yet 75% of the schools in the nation are elementary. Second, the design overrepresented urban schools because urban schools represent some of the greatest challenges in education. We explore in our analyses the effects of these design variables. Nevertheless, given the above constructs, districts and schools were randomly stratified by geographic region and response rates are reported. In our analyses of the VAL-ED, we do not attempt to estimate national parameters.

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. While there are paper and online versions of the VAL-ED, only the paper version was used in the field trial.

Results

Throughout the field testing phase of our project, our research was guided by the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999). Thus, we focus on the categories for reliability, validity, bias, scales, norms, performance standards, and score comparability that are described in the standards.

Reliability and Errors of Measurement

Internal consistency. We estimated the internal consistency reliability of the VAL-ED for a total score and for each of the six core component and six key process subscales, separately for Form A and Form C. Table 1 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, the internal consistency reliability ranged from .98 to .99. Internal consistencies were slightly lower for principals (.87 to .93) than for supervisors or teachers (.93 to .97), and the reliabilities for key processes (.87 to .90) based on principal data were slightly lower than the reliabilities for core components (.89 to .93). The conclusion is that there is strong internal consistency reliability for both total score and each of the 12 subscales, as based on national field trial data.

Standard error of measurement. From Table 1, the lowest internal consistency reliability coefficient for total score is .98. When results are aggregated across respondent groups, the reliability of the aggregate total score is certainly at least .98 reliable.

Table 1. Internal Consistency Reliability for National Field Trial

Variable	Principal		Supervisor		Teachers	
	Form A	Form C	Form A	Form C	Form A	Form C
Total score	.98	.98	.99	.99	.99	.99
Core components:						
High Standards	.89	.91	.96	.95	.96	.95
Rigorous Curriculum	.90	.92	.96	.95	.96	.95
Quality Instruction	.90	.89	.95	.94	.95	.94
Culture of Learning	.90	.93	.96	.95	.96	.96
External Community	.92	.91	.96	.93	.97	.96
Performance Accountability	.91	.92	.97	.95	.97	.97
Key processes:						
Planning	.88	.90	.96	.93	.96	.95
Implementing	.88	.89	.96	.93	.96	.95
Supporting	.90	.88	.96	.94	.95	.95
Advocating	.87	.89	.95	.93	.95	.95
Communicating	.87	.91	.95	.94	.95	.95
Monitoring	.89	.91	.95	.94	.96	.95

The standard deviation of the aggregate total score is .35. Thus, the standard error of measurement for the aggregate total score is at least as small as .05 on the five-point effectiveness scale. 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 on the five-point effectiveness scale. Standard errors of measurement for disaggregated data in the subscales vary. Using the reliability of .90 and the standard deviation of .55, a standard error of measurement is .17. Most of the subscales are .90 reliable or greater and most of the standard deviations are .55 or lower. The standard deviations for supervisor data are slightly larger, ranging from .60 to .80. Thus, all standard errors of measurement across subscales, forms, and respondents are .20 or less.

Validity

Evidence based on test content. The standards describe evidence based on test content as “the relationship between a test’s content and the construct it is intended to measure” (AERA, APA, & NCME, 1999, p. 11). This form of validity also includes the clarity of directions, wording, and rules for administration. The instrument development phase of the study provided substantial evidence related to content and domain validity. For instance, a sorting study indicated that naive respondents were able to accurately sort VAL-ED items into their appropriate cells in the 36-cell framework. A series of cognitive interviews indicated that respondents understood the VAL-ED items and instrument instructions as intended. However, the field test provided additional evidence of content validity, which is summarized here.

Response errors. One potential source of invalidity is respondents not understanding the items and thus answering them incorrectly. 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 check a source of evidence, (c) failing to check 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 giving an

effectiveness rating, (g) checking “no evidence” and not completing the effectiveness rating scale including not indicating “don’t know,” and (h) not checking “no evidence” or another source of evidence and checking “don’t know.” All of the eight types of errors were infrequent, with the possible exception of type *h* for teachers, which occurred on 5.54% of items. For both principals and supervisors, less than 5% of items (roughly three to four items on the VAL-ED) had an error of some kind. For teachers, 11.07%, or about eight items per teacher, had some form of error, with half of these errors of the type *h* variety, which would not affect the principal’s effectiveness rating even without an error.

Face validity. We also asked respondents nine survey questions at the end of the VAL-ED. The questions were on a 1–4 scale, with 1 indicating strongly disagree and 4 indicating strongly agree. Relevant to content validity, we asked respondents if they believed “the vast majority of items focus on important leadership behaviors.” The mean response to this item was 3.18 for principals, 3.05 for teachers, and 3.30 for supervisors, with just 5% of principals, 12% of teachers, and 2% of supervisors disagreeing or strongly disagreeing. We also asked if “this assessment is appropriate for use at the elementary, middle, and high school levels,” with mean responses of 3.15 for principals, 3.05 for teachers, and 3.27 for supervisors and just 10% of principals, 14% of teachers, and 6% of supervisors disagreeing or strongly disagreeing. Finally, we asked if respondents “understood the vast majority of items,” with mean responses of 3.20 for principals, 2.94 for teachers, and 3.27 for supervisors and 5% of principals, 18% of teachers, and 1% of supervisors disagreeing or strongly disagreeing. Thus, all three respondent groups understood the VAL-ED, believed it was appropriate for use in all kinds of schools, and agreed that the VAL-ED items were focused on important leadership behaviors, a measure of their perception of the instrument’s content validity. These results support that the development phase had led to the creation of a content-valid instrument for measuring principal learning-centered leadership.

Evidence Based on Internal Structure

The development of the VAL-ED was based on the 36-cell conceptual framework that emerged from the literature on effective principal learning-centered leadership. We asked to what extent the national field trial data supported the conceptual framework.

Scale and total score intercorrelations. Because there are so many correlations to consider (65 for each of the three respondent groups), the patterns are summarized here. For principals, the average intercorrelation of core components was .73, with a minimum of .65 and a maximum of .85. For key processes, the average was .85, the minimum .80, and the maximum .88. Results were similar for supervisors and teachers, except that average correlations were higher for both respondent groups than for principals. Average core component intercorrelations were .93 for teachers and .86 for supervisors; average key process intercorrelations were .95 for teachers and .92 for supervisors. Thus, examination of the intercorrelations reveals that the subscales and total score are highly correlated, especially for key processes and for teachers.

Principal effectiveness on the VAL-ED is reported not only on the total score but also by core component and by key process. The set of core components and the set of key processes are redundant because 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 finding the 36-factor structure unlikely. Neither is it likely to find a 12-factor structure, one for each of the six core components and six key processes, because the two sets of six are totally redundant. While the analyses in the following sections provide information about factor structure, the most important questions are whether the core components and key processes can be reliably distinguished, one from another. Many assessments, including student achievement assessments, report subscale results without reporting on the reliability of differences among subscales. This is probably because the reliability of differences is notoriously low due to typical high correlations among the subscales. The reliabilities reported below should be interpreted within this context.

Traditional reliability of differences. One way to examine the reliability of the difference between subscales is to use the classical approach (Stanley, 1967).

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

Tables 2, 3, and 4 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. Three out of the 180 reliabilities were negative. 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).

In terms of differentiating core components from one another, the reliability of the difference calculations indicates that Culture of Learning and Professional Behavior, Connections to External Communities, and Performance Accountability are reliably 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 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. Comparatively, these are strong reliabilities of differences.

The results are weaker 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 somewhat stronger. The large majority of reliabilities are between .40 and .60. There

Table 2. Reliability of the Differences among Scales, Principal Data

Form A Core Components						
	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	.43	1				
Quality Instruction	.46	.40	1			
Culture of Learning	.60	.69	.54	1		
External Community	.72	.73	.71	.74	1	
Performance Accountability	.63	.62	.60	.72	.74	1

Form A Key Processes						
	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	.14	1				
Supporting	.17	.17	1			
Advocating	.25	.25	.22	1		
Communicating	.38	.26	.36	.24	1	
Monitoring	.34	.08	.35	.32	.04	1

Form C Core Components						
	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	.36	1				
Quality Instruction	.57	.49	1			
Culture of Learning	.64	.69	.59	1		
External Community	.75	.73	.64	.72	1	
Performance Accountability	.72	.72	.66	.67	.58	1

Form C Key Processes						
	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	.05	1				
Supporting	.06	-.10	1			
Advocating	.35	.12	.17	1		
Communicating	.44	.39	.24	.43	1	
Monitoring	.51	.49	.43	.55	.32	1

are a few particularly strong comparisons, such as the comparison between Advocating or Supporting and Monitoring. Nevertheless, there are no comparisons with reliabilities greater than .70 for any respondent group for any key process. The overall results from the traditional reliability of the difference analysis are mostly positive for core components, weaker for key processes for teachers, and somewhat weak for key processes for principals and supervisors. If, over time, key processes

Table 3. Reliability of the Differences among Scales, Teacher Data

	Form A Core Components					
	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	.60	1				
Quality Instruction	.64	.58	1			
Culture of Learning	.70	.70	.58	1		
External Community	.84	.84	.82	.79	1	
Performance Accountability	.74	.73	.69	.70	.82	1

	Form A Key Processes					
	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	.26	1				
Supporting	.48	.42	1			
Advocating	.48	.50	.55	1		
Communicating	.45	.47	.52	.45	1	
Monitoring	.54	.58	.60	.60	.52	1

	Form C Core Components					
	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	.58	1				
Quality Instruction	.66	.50	1			
Culture of Learning	.69	.69	.64	1		
External Community	.80	.79	.75	.76	1	
Performance Accountability	.79	.76	.72	.77	.81	1

	Form C Key Processes					
	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	.20	1				
Supporting	.40	.25	1			
Advocating	.49	.48	.57	1		
Communicating	.46	.45	.46	.58	1	
Monitoring	.54	.55	.59	.66	.47	1

become more distinctive one from another, the reliabilities of their differences might improve. For the national field trial data, they are less distinct one from another than are the core components.

Table 4. Reliability of the Differences among Scales, Supervisor Data

	Form A Core Components					
	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	.29	1				
Quality Instruction	.16	.38	1			
Culture of Learning	.59	.63	.41	1		
External Community	.75	.78	.77	.72	1	
Performance Accountability	.49	.62	.60	.73	.78	1

	Form A Key Processes					
	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	.26	1				
Supporting	.17	.07	1			
Advocating	.37	.38	.36	1		
Communicating	.14	.00	-.40	.10	1	
Monitoring	.07	.31	.03	.41	.02	1

	Form C Core Components					
	High Standards	Rigorous Curriculum	Quality Instruction	Culture of Learning	External Community	Performance Accountability
High Standards	1					
Rigorous Curriculum	.42	1				
Quality Instruction	.48	.37	1			
Culture of Learning	.70	.69	.59	1		
External Community	.80	.78	.70	.75	1	
Performance Accountability	.68	.71	.64	.73	.79	1

	Form C Key Processes					
	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	1					
Implementing	-.13	1				
Supporting	.34	.19	1			
Advocating	.50	.50	.62	1		
Communicating	.36	.26	.33	.57	1	
Monitoring	.40	.31	.64	.64	.39	1

Exploratory factor analysis. We examined the internal structure of the VAL-ED using both exploratory and confirmatory techniques. The exploratory analysis was designed to examine whether item responses tended to cluster in ways that indicated

the presence or absence of core components or key processes. To conduct the exploratory factor analysis, we used a promax approach to oblique rotations on the 72 items for Form A and the 72 items for Form C where the data were aggregated across the three respondent groups. 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 for exploratory factor analysis were thus 72 items by 214 schools. Four 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).

Form A had eight eigenvalues greater than or equal to 1; for Form C, there were nine. Nevertheless, we tried six-, eight-, and twelve-factor solutions for both forms. 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. The six-factor solution had difficulty obtaining simple structure (i.e., a factor matrix with one large factor loading per item with the rest of the loadings for that item near zero). The twelve-factor solution was driven by the hypothesis that perhaps there would be a factor for each core component and each key process. For the twelve-factor solutions, the last three factors were noninterpretable. The eight-factor solution was driven by the rule of thumb that there are as many factors as there are eigenvalues greater than or equal to 1. We chose an eight-factor solution for each form. Table 5 contains the factor matrix for Form C, the more easily interpreted of the two. Each factor loading equal to or greater than .40 is flagged in the table. While it would be ideal to compare the factor loadings with the text of the individual items, it is not possible to display all the items here due to copyrighting. However, the highest-loading item for each factor is presented as an illustration. As is always the case in exploratory factor analysis, there is some art in naming factors.

For Form C, Factor 1 might be labeled Culture of Learning and Professional Behavior. All of the factor loadings for the 12 Culture of Learning and Professional Behavior items (items 37–48) on Factor 1 were .40 or larger, as compared to only two of the other 60 items with factor loadings that large. The largest factor loading for Factor 1 is item 42, which reads, “How effective is the principal at ensuring the school provides a positive environment in which student learning is the central focus?” This item has a factor loading of .57 for Factor 1 and no greater than .25 for any other factor.

Factor 2 is a combination of High Standards for Student Learning, Rigorous Curriculum, and Quality Instruction. Of the 36 items measuring these core components (items 1–36), nine had factor loadings .40 or greater, 21 had factor loadings of .25 or greater, and all but two had loadings of .10 or greater. There were just five other items with loadings above .25 on this factor. For this factor, the highest loading is for item 17, which reads, “How effective is the principal at ensuring the school supports participation in professional development that deepens teachers’ understanding of a rigorous curriculum?” Item 17 has a factor loading of .60 for Factor 2 and .19 or less for all other factors.

Factor 3 represents Performance Accountability. Ten of the 12 items measuring Performance Accountability (items 61–72) had factor loadings of .20 or greater on Factor 3 and all had factor loadings of .15 or greater. There are some moderate loadings on this factor under High Standards for Student Learning (items 1, 11, 12) and Monitoring Rigorous Curriculum (item 23). The highest loading is for item 61, which

reads, “How effective is the principal at ensuring the school plans data collection to hold students accountable for academic and social learning?” Item 61 has a factor loading of .64 for Factor 3 and .16 or lower for all other factors.

Factor 4 represents Advocating. Of the 12 items measuring Advocating (items 7–8, 19–20, 31–32, 43–44, 55–56, and 67–68), 11 had factor loadings of .10 or greater for Factor 4 and 10 had factor loadings greater than .20. There were two items from Rigorous Curriculum (items 13 and 15) that had large factor loadings on Factor 4. The highest factor loading for Factor 4 is item 7, which asks, “How effective is the principal at ensuring the school advocates for students with special needs when making decisions about high standards for student learning?” This item has a factor loading of .75 for Factor 4 and .21 or lower for all other factors.

Factor 5 represents Connections to External Communities. Items measuring external communities (items 49–60) comprise eight of the nine items with factor loadings of .40 or higher on Factor 5—the only exception is item 31, which is under Advocating Quality Instruction. The highest loading for Factor 5 is for item 50, which reads, “How effective is the principal at ensuring the school plans activities to engage families in student learning?” This item has a factor loading of .76 for Factor 5 and .13 or lower for all other factors.

Factor 6 might be labeled Communicating and Monitoring Rigorous Curriculum and Quality Instruction. All eight of these items (items 21–24 and 33–36) had factor loadings at .25 or higher on Factor 6. However, there are eight other items with factor loadings of at least .25 on that factor, with five of those eight falling under Monitoring and some other core component. The highest factor loading for Factor 6 was item 36, “How effective is the principal at ensuring the school observes each teacher’s instructional practices routinely to provide feedback?” Item 36 has a factor loading of .71 for Factor 6 and .12 or lower for the other factors.

Factor 7 might be labeled Supporting Quality Instruction and Performance Accountability. All four of these items (items 29–30 and 65–66) had factor loadings of .50 or higher on Factor 7, though two other items did as well. The highest loading item is item 65, with a loading of .58; the item reads, “How effective is the principal at ensuring the school provides procedures that hold students accountable for their learning?” Item 65 also has a moderate factor loading of .40 on Factor 3. 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 (not shown due to space limitations). Factor 2 for the Form A solution represented Connections to External Communities. Of the 12 items measuring Connections to External Communities, 10 had factor loadings of .30 or greater and all 12 had factor loadings of .15 or greater. Factor 4 on Form A might be labeled Performance Accountability. Of the 12 items measuring Performance Accountability, all 12 had factor loadings of .30 or higher on Factor 4. Also similar to a Form C factor was Form A Factor 1, which might be labeled Support for High Standards, Rigorous Curriculum, Quality Instruction, and Culture of Learning and Professional Behavior. Of the eight items in those cells (items 5–6, 17–18, 29–30, and 41–42), seven had loadings of .4 or higher on Factor 1. Similarly, Factor 7 on Form C was labeled Supporting, but that factor for Supporting included only Quality Instruction and Performance Accountability.

Other factors on Form A were somewhat different from Form C factors. Factor 5

Table 5. Exploratory Factor Analysis Factor Matrix for Form C

Item	Factor							
	1	2	3	4	5	6	7	8
1	.18	.45 ^a	.41 ^a	.12	-.04	.00	.08	.15
2	.37	.45 ^a	.36	-.05	.05	.01	.03	.05
3	.25	.37	.33	.04	.08	.08	.12	.17
4	.10	.15	.11	-.03	.15	-.02	.01	.68 ^a
5	.31	.25	.23	.05	-.01	.04	.09	.37
6	.32	.19	.17	.13	-.09	.11	.09	.40 ^a
7	.09	.21	-.07	.75 ^a	.10	-.09	-.05	.19
8	.18	.40 ^a	.27	.28	-.07	.07	-.06	.12
9	.16	.41 ^a	.29	.14	.02	.11	.00	.23
10	.17	.43 ^a	.19	.09	-.04	.06	.03	.29
11	.10	.10	.47 ^a	.06	.03	.31	-.07	.21
12	-.12	.32	.62 ^a	.13	-.05	.19	.04	.13
13	.06	.15	.14	.65 ^a	-.01	.12	.08	.08
14	.14	.36	.36	.14	-.06	.16	.21	.01
15	.00	.15	.09	.57 ^a	-.10	.20	.11	.20
16	-.05	.47 ^a	.28	.10	.04	.08	.31	.16
17	.15	.60 ^a	-.04	.05	.19	.19	.02	.14
18	.02	.54 ^a	-.05	.15	.11	.24	.03	.16
19	.22	.40 ^a	.15	.36	-.03	.16	.12	.00
20	.11	.29	.01	.69 ^a	-.04	.07	.04	.05
21	.03	.31	.14	.02	-.02	.52 ^a	.20	.02
22	.22	.15	-.13	.11	-.04	.25	.34	.34
23	-.13	.30	.46 ^a	.10	.15	.28	-.01	.15
24	.20	.12	.11	.06	.15	.63 ^a	-.01	.02
25	.16	.10	.06	.09	.02	.12	.21	.55 ^a
26	.14	.58 ^a	-.01	.02	.16	.21	.10	.17
27	.00	.40 ^a	.21	.18	.10	.20	.16	.13
28	-.12	.01	-.11	.22	.13	.35	.24	.40 ^a
29	-.07	.39	-.01	.06	.25	-.10	.50 ^a	.26
30	-.01	.38	.10	.11	.13	-.13	.51 ^a	.27
31	.08	.19	.01	-.06	.56 ^a	-.06	.44 ^a	-.08
32	.12	.24	.21	.38	.24	.03	.19	-.22
33	.14	.39	.20	.00	-.06	.45 ^a	.14	.05
34	.29	.06	.02	.02	-.01	.59 ^a	.07	.23
35	-.11	.24	.43 ^a	-.05	.10	.52 ^a	.05	.10
36	.10	.10	.10	.09	.12	.71 ^a	.00	-.03
37	.44 ^a	.32	-.02	-.09	.03	.32	.30	-.01
38	.47 ^a	.27	.04	.04	.11	.18	.18	.03
39	.42 ^a	.34	.17	.03	.02	.05	.14	.21
40	.50 ^a	-.02	.27	.04	.06	-.19	.10	.40 ^a
41	.55 ^a	.20	-.08	.00	.01	.19	.00	.37
42	.57 ^a	.13	.10	.06	.14	.04	.07	.25
43	.46 ^a	-.01	.05	.41 ^a	.01	.16	.02	.08
44	.43 ^a	.05	-.08	.38	.23	.14	.08	.02
45	.51 ^a	.07	.05	.26	.24	-.01	.10	.07
46	.52 ^a	.16	.14	.18	.04	.13	.06	.16
47	.53 ^a	.03	.02	.13	.14	.12	.14	.26
48	.50 ^a	.09	.02	.23	.06	.25	.13	.11
49	.14	.26	-.07	-.06	.65 ^a	.07	.13	.04
50	.07	.11	.13	.03	.76 ^a	.09	-.06	.11
51	-.06	.13	.15	.12	.74 ^a	.02	-.01	.11
52	.45 ^a	.01	-.05	.07	.53 ^a	-.02	-.01	.16
53	.13	.06	-.08	.23	.42 ^a	.10	.32	-.07
54	.27	-.08	-.02	.32	.38	.12	.15	.13
55	.01	-.01	.22	.49 ^a	.31	.03	.08	.10

(Continued)

Table 5. Exploratory Factor Analysis Factor Matrix for Form C (Continued)

Item	Factor							
	1	2	3	4	5	6	7	8
56	.08	-.11	.22	.55 ^a	.21	-.02	.15	.05
57	.30	-.08	.15	.31	.35	-.02	.05	.25
58	.05	.06	.40 ^a	.00	.41 ^a	.21	.08	.01
59	.07	-.05	.03	-.09	.43 ^a	.25	.51 ^a	.02
60	-.03	-.25	.02	.07	.46 ^a	.31	.25	.24
61	.16	.08	.64 ^a	-.03	-.01	.00	.23	.06
62	.41 ^a	.11	.16	-.05	.14	.23	.25	.11
63	.25	-.06	.48 ^a	.17	.17	.00	.12	.20
64	.14	.25	.42 ^a	.04	.06	.18	.20	.21
65	.09	-.04	.40 ^a	.17	-.09	.03	.58 ^a	.13
66	.18	-.03	.38	.00	-.04	.14	.56 ^a	.17
67	.11	-.09	.15	.24	.19	.31	.25	.11
68	.32	.17	.24	.11	.16	.18	.14	.09
69	.07	-.12	.49 ^a	.23	.33	.04	.26	-.07
70	.08	.15	.45 ^a	.18	.05	.22	.24	.06
71	.04	.03	.27	.11	.10	.44 ^a	.23	.02
72	.00	.04	.54 ^a	.00	.11	.12	.38	.04

^aIndicates the entry is .40 or larger.

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 12 items measuring the combination of those core components and key processes (items 43–48 and 55–60), 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 (items 11–12, 35–36, and 71–72), all had factor loadings of .25 or higher. Factor 6 is for High Standards for Student Learning. Of the 12 items (items 1–12), all but one had factor loadings of .20 or higher on Factor 6. 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 (items 7–8 and 19–22), all had factor loadings at .29 and higher. As with Form C, Factor 8 on Form A 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. For all three solutions for each form, there was strong evidence in support of a factor for Connections to External Community and a factor for Performance Accountability. There was some support for the core component of Culture of Learning and Professional Behavior and the key processes of Supporting and Advocating. Most often, there was evidence of simple structure insofar as most of the items with large factor loadings on one factor had small factor loadings on other factors. In general, however, the factors that emerged in the factor analysis did not perfectly conform to the conceptual framework in the sense that, in all cases, there were some items from conceptually distinct cells in the framework that had factor loadings as large as the more conceptually related items we identified in the factors.

Confirmatory factor analysis (CFA). To further test the fit of the data to the

Form A 360-Score Factor Loadings for Core Components Confirmatory Factor Analysis																	
Item	Level 1 Factor Loading	Level 1 Factor	Level 2 Factor Loading (Core Components)	Level 2 Factor (Core Components)	Level 3 Factor Loading	Level 3 Factor	Level 3 Factor Loading (Core Components)	Level 3 Factor (Core Components)	Level 2 Factor Loading	Level 2 Factor (Core Components)	Level 1 Factor Loading	Level 1 Factor	Item				
Item 1	.87	HighStandards_Planning	.96	High Standards	1.00	.93	Culture of Learning		.87	CultureofLearning_Planning	.84	Item 27					
Item 2	.88	HighStandards_Implementing	1.00						.90	CultureofLearning_Implementing	.83	Item 28					
Item 3	.78	HighStandards_Supporting	.95						.98	CultureofLearning_Supporting	.89	Item 43					
Item 4	.89	HighStandards_Advocating	.79						.96	CultureofLearning_Advocating	.74	Item 41					
Item 5	.80	HighStandards_Communicating	.98						1.00	CultureofLearning_Communicating	.74	Item 42					
Item 6	.84	HighStandards_Monitoring	.92						.92	CultureofLearning_Monitoring	.75	Item 44					
Item 7	.81	RigorousCurriculum_Planning	.99						Rigorous Curriculum	.97	.86	External Communities		.99	ExtCommunities_Planning	.93	Item 45
Item 8	.86	RigorousCurriculum_Implementing	.90											.85	ExtCommunities_Implementing	.83	Item 50
Item 9	.90	RigorousCurriculum_Supporting	.94											1.00	ExtCommunities_Supporting	.86	Item 51
Item 10	.76	RigorousCurriculum_Advocating	.94											.96	ExtCommunities_Advocating	.71	Item 52
Item 11	.91	RigorousCurriculum_Communicating	.92	.75	ExtCommunities_Communicating	.83	Item 53										
Item 12	.90	RigorousCurriculum_Monitoring	.96	.94	ExtCommunities_Monitoring	.83	Item 54										
Item 13	.82	QualityInstruction_Planning	1.00	Quality Instruction	.98	.94	Performance Accountability							.94	Accountability_Planning	.95	Item 55
Item 14	.87	QualityInstruction_Implementing	.96											.97	Accountability_Implementing	.84	Item 56
Item 15	.88	QualityInstruction_Supporting	.85											.92	Accountability_Supporting	.90	Item 57
Item 16	.86	QualityInstruction_Advocating	1.00											.98	Accountability_Advocating	.83	Item 58
Item 17	.79	QualityInstruction_Communicating	.95						1.00	Accountability_Communicating	.76	Item 59					
Item 18	.85	QualityInstruction_Monitoring	.89						1.00	Accountability_Monitoring	.85	Item 60					
Item 19	.80																
Item 20	.90																
Item 21	.82																
Item 22	.73																
Item 23	.86																
Item 24	.89																
Item 25	.68																
Item 26	.76																
Item 27	.90																
Item 28	.73																
Item 29	.89																
Item 30	.89																
Item 31	.80																
Item 32	.64																
Item 33	.71																
Item 34	.94																
Item 35	.86																
Item 36	.91																

Figure 4. Confirmatory factor analysis for Form A.

conceptual framework, a confirmatory factor analysis was completed using the same aggregated data as in the exploratory analysis. 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., 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 for core components and the other for 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 model was fit using PROC CALIS as implemented in SAS 9.1.3. The results of the core components analysis for Form A are displayed in Figure 4 as an illustration. The rest of the results are summarized here.

Results from the confirmatory factor analysis reveal that both the core components and the key processes models fit the data well, having goodness-of-fit indices (GFI) for both the GFI and the adjusted GFI of .99 for both analyses for Form A and .98 for both analyses 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 high—.92 or greater for each analysis on both forms.

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, as seen in Figure 4. 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 for both forms. For the core components analysis, 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 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 analysis. For core components, the lowest factor loading for Form A was .86 for Connections to External Communities and the highest was 1.0 for High Standards. For Form C, the lowest

was .83 for Connections to External Communities and the highest was .99 for Quality Instruction. All of the factor loadings were .99 or higher for key processes for Form A. 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.

Mean differences in the conceptual framework. Since results are all in the mean item response metric, yet another way to investigate the validity of the conceptual framework is by asking the extent to which the means for core components differ reliably one from another, the means for key processes differ one from another, and the means for the 36 cells differ one from another. The six core components by six key processes conceptual framework hypothesizes that there are 12 distinct subscales and 36 distinct domains of leadership behavior. It could be that currently, principal behaviors do not differ across these conceptual distinctions. Still, if differences are found, it is yet one more indication that the items measure these conceptual distinctions. A two-dimensional analysis of variance (core components by key processes) was completed on the data aggregated across respondent groups for the two forms. The aggregation took place first at the item level as described in previous analyses.

For Forms A and C, the mean square for core components was largest, followed by the mean square for key processes, followed by the interaction of core components by key processes. For each form, all three were statistically significant at the .001 level. For Form A, the means were as follows: High Standards, 3.64; Rigorous Curriculum, 3.58; Quality Instruction, 3.74; Culture of Learning, 3.74; Connections to 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. Overall, there were only two instances where core components were not significantly different from one another using a Tukey adjustment for multiple comparisons. Rigorous Curriculum and Performance Accountability were not significantly different from each other; nor were Quality Instruction and Culture of Learning and Professional Behavior. 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.

For key processes, the Form A means were as follows: Planning, 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. 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. 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 perceived effectiveness.

Because the core components-by-key processes interaction was significant, Tukey post-hoc pairwise comparisons were used to contrast each pair of cells. Even at the cell level, where each cell consists of only two items, many cells were significantly

different from others at the .05 level. Of the 630 possible pairwise comparisons for Form A, 276 (44%) were significant. Of these, 111 involved contrasting a cell for Supporting to another cell, and 113 involved contrasting a cell for Connections to External Communities to another cell. There was evidence of the significant interaction as well. For example, many of the cell contrasts between High Standards for Student Learning and Performance Accountability were not significant. Nevertheless, Monitoring for High Standards was significantly different from five of the six cells for Performance Accountability, the one exception being Supporting Performance Accountability. Cell means ranged from a mean item response of 3.99 for Supporting Quality Instruction to a low of 3.25 for Advocating Connections to External Communities.

The results for Form C were similar to the results for Form A. Again, mean effects for core components and key processes were significant, as was the interaction. Tukey pairwise comparisons found that most of the core components were significantly different one from another. The only exceptions were that High Standards was not significantly different from Culture of Learning, and Connections to External Communities was not significantly different from Performance Accountability. For key processes, Supporting was significantly different from all other key processes except Communicating. In addition, Communicating was significantly different from Planning, Implementing, Advocating, and Monitoring.

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 cells, 296 were significant. Again, many of the significantly different cell means were accounted for by contrasting cells for Supporting with other cells. Similarly, many of the differences were accounted for by contrasting Connections to External Communities or Performance Accountability cells with other cells. Again, there was evidence of 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 were run by form and respondent group. For each respondent group in both forms, the core components' main effect, the key process main effect, and the interaction effect were statistically significant.

Evidence Based on Relations to Other Variables

The data set generally did not support predictive studies of the validity of the VAL-ED. We used regression analyses to see if design variables and other variables were significant predictors of VAL-ED scores. These analyses were done on the aggregate sample as well as the sample for each response group (see Table 6). Two findings offer modest validity evidence. First, principals' experience in the school was statistically significant at the .05 level. Principals with more years of experience in the school were rated more highly when data were aggregated across groups. One might hypothesize that more experienced principals are on average more learning centered. Second, principals in high schools had .17 points lower effectiveness ratings than principals in elementary schools, a difference of .50 standard deviations. Generally, high school principals are seen as less learning centered than elementary school principals.

Respondent group correlations. The literature on multirater assessments indicates that respondent group correlations should be positive, but there is no definite

Table 6. Design Factors Regression Results for Total Score Aggregated Sample

Design Factors	Coefficient	Standard Error
Number of teachers	.00005	.001
Wallace	.07	.09
Midwest	-.08	.07
West	-.21**	.07
South	-.05	.07
Suburban	.08	.09
Rural	.02	.10
Form C	.03	.05
Middle	-.07	.06
High	-.17*	.07
Teacher response rate	.002	.001
Years principal	.008*	.004
Supervisor feasibility:		
I found this response form easy to use	.15**	.05
The amount of time required to complete this instrument is reasonable	-.11*	.05

Note.—For feasibility questions, only significant results are reported.

* $p < .05$.

** $p < .01$.

*** $p < .001$.

standard for the magnitude of the correlation coefficient (Borman, 1997). Studies in psychology and management generally find respondent group correlations in the .25 to .35 range (Atwater, Ostroff, Yammarino, & Fleenor, 1998; Harris & Schaubroeck, 1988). Nonperfect correlations are expected, and they can be attributed to measurement error, along with four more substantive potential explanations: “(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” (Murphy & Deshon, 2000, p. 822). Of course, if they were perfectly correlated, there would be no value in using the 360 strategy of multiple respondent groups.

To calculate the respondent group correlations, we ignored form and used all 218 schools with data from all three respondent groups. We first calculated the teacher mean for the school by averaging across all the teachers in each school. Next, we correlated the scale and total scores, with the school as the unit of analysis. Results indicate that the multirater approach to principal assessment is useful. Correlations among the respondent groups on total score were .13 for supervisors and principals, .18 for teachers and supervisors, and .27 for teachers and principals. Across all core components and key processes, the correlations among any pair of respondent groups were positive and between .02 and .30. These correlations are in line with typical respondent group correlations on multirater assessments.

Bias

In addition to examining the validity and reliability of the VAL-ED, item bias was examined using a differential item functioning (DIF) framework. The purpose was

Table 7. Text of Items Exhibiting DIF

Form	Item	Item Text
C	7	... advocates for students with special needs when making decisions about high standards for student learning.
C	26	... plans opportunities for teachers to improve their instruction through professional development.
C	38	... plans for a culture of shared responsibility for the social and academic learning of students.
C	40	... builds a school environment that is safe and orderly for all students.
C	49	... plans for the use of external community resources to promote academic and social learning goals.

to investigate the possible presence of construct-irrelevant variance in VAL-ED item responses (for a full reporting of DIF results, see Polikoff et al., 2009).

Item response data from the VAL-ED national field trial were subjected to DIF analyses using PARSCALE, version 4.1 (Muraki & Bock, 1996). Samejima's (1969) graded response item response theory model with a logit link function was used to model the cumulative probability of responses to the 72 VAL-ED survey items employing a five-point Likert scale. This model leverages the ordinal nature of the item responses by extending the traditional two-parameter logit IRT model to include a set of four category parameters that separate the five response categories. In the analyses, each item had distinct intercept (i.e., difficulty) and slope (i.e., discrimination) parameters, but the category parameters were constrained to be the same for all items, implying that while it may have been easier or harder to earn high ratings on various items, and while the relative contribution of items to the total score varied, the spacing between response categories was assumed to be constant from one item to the next.

The DIF analyses assessed item bias by allowing the difficulty and discrimination parameters for each item to vary across specific subgroups. In the analyses presented here, we tested for differences between subgroups defined by grade level (i.e., elementary, middle, high), urbanicity (i.e., rural, suburban, urban), and region (i.e., Midwest, Northeast, South, West). While it would be useful to examine DIF by principal characteristics, such as race and gender, we did not collect information on these variables in the field trial. DIF was identified using a chi-square test of the standardized difference between the item parameters of two groups. Because of the large number of comparisons, only p -values below .01 were considered for DIF. The reference categories for these analyses were elementary, urban, and Northeast.

The results, shown in Table 7, revealed few items exhibiting DIF for any subgroup comparisons. Four items exhibited intercept DIF for urbanicity on Form C, two for the urban-suburban comparison and two for the urban-rural comparison. Item mean differences between groups were between .22 and .46 standard deviations. One item on Form C exhibited slope DIF for the urban-rural comparison. No items on Form A or for region or school level exhibited significant slope or intercept DIF, indicating that the instrument can be used in elementary, middle, and high schools without concerns about item bias.

The results of the DIF analysis largely support the findings of an earlier qualitative item bias study based on ETS guidelines, where four items were identified as having potential bias concerns. Those four items were corrected based on the feedback from

Table 8. Scale and Total Score Means and Standard Deviations by Respondent Group and Form

Form	Principal				Supervisor				Teacher			
	Mean		SD		Mean		SD		Mean		SD	
	A	C	A	C	A	C	A	C	A	C	A	C
Total score	3.50	3.54	.49	.51	3.66	3.69	.71	.59	3.61	3.58	.43	.42
Core components:												
High Standards	3.59	3.71	.50	.53	3.65	3.75	.74	.63	3.61	3.66	.45	.41
Rigorous Curriculum	3.47	3.52	.54	.61	3.63	3.66	.75	.66	3.60	3.51	.41	.42
Quality Instruction	3.70	3.63	.57	.56	3.74	3.72	.77	.62	3.71	3.58	.43	.42
Culture of Learning	3.72	3.74	.57	.57	3.78	3.83	.72	.63	3.69	3.67	.44	.46
External Community	3.12	3.34	.63	.60	3.51	3.59	.74	.64	3.46	3.54	.44	.41
Performance Accountability	3.39	3.31	.55	.60	3.60	3.55	.78	.66	3.53	3.45	.46	.43
Key processes:												
Planning	3.51	3.50	.52	.53	3.64	3.68	.73	.62	3.56	3.54	.45	.40
Implementing	3.50	3.55	.50	.52	3.63	3.65	.74	.65	3.59	3.58	.46	.42
Supporting	3.71	3.66	.54	.52	3.78	3.78	.71	.60	3.73	3.63	.43	.44
Advocating	3.42	3.48	.51	.53	3.61	3.62	.75	.57	3.57	3.54	.40	.38
Communicating	3.47	3.59	.54	.56	3.67	3.76	.72	.61	3.62	3.64	.44	.44
Monitoring	3.40	3.47	.54	.61	3.64	3.67	.74	.68	3.59	3.52	.44	.46
N	106	130			124	130			113	132		

the item bias panel (Porter et al., 2010). After these two analyses, there is good evidence that few if any items exhibit serious item bias. All items will continue to be monitored for DIF using operational data.

Parallel Forms

Forms A and C were created by randomly selecting two items from the pool of items written for each of the 36 cells in the 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, nor did it result in equal numbers of schools returning data for the two forms. Part of the explanation is that the districts differ in size and, by chance, some larger districts were 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.

Table 8 compares Form A to Form C on means and standard deviations for each of the respondent groups. For total score, the means for Form A are 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 8 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 deviations 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, there are some differences by subscale. The differences are as large as .22 for

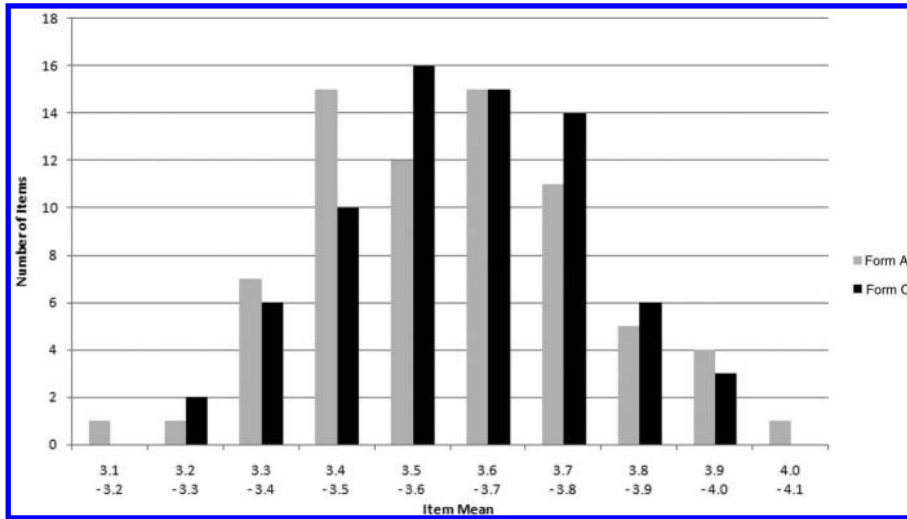


Figure 5. Distribution of aggregate item means by form.

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 nonrandomly 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 5. Recall that the standard error of measurement across scales was .20 or less. 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 Form A and Form C operating as parallel forms.

Norms and Performance Standards

In addition to reporting total score and scale scores by respondent group and in the aggregate, VAL-ED scores are reported against national norms and performance standards.

Norms. Using the results from the national field trial, initial norms for the VAL-ED were set for total score and each of the 12 subscores, 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, there are 52 (13×4) sets of norms to potentially be used on the VAL-ED score report. As an example, for aggregate total score, mean item response ranged from a low of 2.573 to a high of 4.506 with a median of 3.604. Using the norms table constructed for the aggregate total score, a principal with an aggregated total score mean item response of 4.00 would have the percentile rank of 86. The norm-referenced reporting of VAL-ED scores gives leaders a statement of their learning-centered leadership relative to other principals nationwide who have taken the VAL-ED. The norms are not perfect in the sense of being a national probability sample of schools for reasons discussed earlier (e.g., an underrepresentation of elementary schools).

Performance standards. The results of the VAL-ED can also be reported in terms of performance levels: distinguished, proficient, basic, and below basic. To set the performance standards for the VAL-ED, a panel of experts was assembled to complete a modified bookmark procedure (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 based on the field trial data, an aggregate variable was defined as the arithmetic mean of an item's mean item response across the three response groups, equally weighted. The variable could take values on the 1 to 5 effectiveness rating scale.

Standard-setting panelists based their judgments on a set of performance level descriptors (PLDs). The final PLDs used in the standard setting task 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 subgroups 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.

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 and the fact that the forms were constructed to be and were operating as parallel. 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 national panel of 22 experts was recruited, with 10 principals, four teachers, four supervisors of principals, two school leadership researchers, and two education policy makers. The panelists included four recent recipients and one finalist for Principal of the Year designation for their states, two American Association of School Administrators (AASA) Distinguished Principals, two principals in National Association of Secondary School Principals Breakthrough Schools, one Middle School National Teacher of the Year, one State Teacher of the Year, one Distinguished Distributive Leadership Teacher, one Outstanding Educator at the district level, and four district superintendents with an AASA designation of excellence. The panel consisted of five tables placing bookmarks independently. At the end of the 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 national field trial principals being 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 poststandard setting communication with panelists asked them whether they wished to (a) keep the cuts where they were set at the end of the panel, (b) move the cuts to the median cut for the least demanding table 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 effectiveness scale. All 22 panelists responded; 21 favored moving the basic to below basic cut to be less demanding and the proficient to distinguished cut to be more demanding. Of those 21, 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 12 subscales. To use proficiency levels with subscales, the assumption is made that the judgments on the total score apply equally to the VAL-ED subscales. This assumption allows reporting, for instance, if a principal's performance as reported by teachers on Rigorous Curriculum is distinguished, proficient, basic, or below basic.

Summary and Conclusions

The Vanderbilt Assessment of Leadership in Education is a multirater assessment of principals' learning-centered leadership behaviors based on the research on effective school leadership. In this article, we described the validity, reliability, bias, parallel forms, norms, and performance standards for the VAL-ED based on results from the national field trial. The field trial sample consisted of more than 270 schools and more than 8,000 individual evaluations, with 218 schools having complete sets of responses. The sample included urban, suburban, and rural schools; elementary, middle, and high schools; and schools from all regions of the country. Results provided evidence on the validity of the conceptual framework and showed that VAL-ED scores are higher for principals who have been in the building longer. Scale and total score reliabilities were high, and the standard error of measurement was low across scales and respondents. The forms, as designed, tended to operate as parallel though not perfectly. Finally, the instrument's norms and performance standards were briefly described.

Core Components						
ANOVA Alpha diff. 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.	++ + ++	+ ++ ++	++ ++ ++	++ ++ ++	+ ++ ++	
Key Processes						
ANOVA Alpha diff. 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	

Figure 6. Evidence on the VAL-ED’s ability to distinguish core components and key processes. 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. EFA: ++ indicates a clear factor on both forms; + indicates a clear factor on one form; 0 indicates no clear factors.

Figure 6 provides a summary of the empirical support for the conceptual framework for core components and key processes. In each case, there were three 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 subscale with another subscale, the criterion being reliabilities of .50 or higher; and (c) the presence of clear factors in the exploratory factor analysis using oblique rotation. Results of distinctions among the core components and key processes 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 (o) for each pairwise comparison. The first entry in a cell is for significant mean difference, the second for classical reliability of the differences, and the third based on exploratory factor analysis. As is seen in Figure 6, the empirical support for the core components is overwhelmingly positive, while the evidence in support of key processes is weak for Planning, Implementing, and Monitoring and moderate for Supporting, Advocating, and Communicating.

The results also indicate that the VAL-ED can appropriately be used in elementary, middle, and high schools, and that the items do not operate differentially based on school characteristics. Differential item functioning analyses indicated no significant item bias on school region or level, and limited item bias (five items) on urbanicity. Furthermore, survey responses on the 1–4, strongly disagree to strongly agree scale indicate that principals (mean = 3.44), teachers (mean = 3.33), and supervisors (mean = 3.44) believe the items are not biased based on principal race or gender. As for school level, elementary school (mean = 3.11), middle school (mean = 3.24), and high school (mean = 3.13) respondents agree that the VAL-ED is appropriate for use at all levels of schooling.

A final conclusion about the VAL-ED based on these results is that the instrument better differentiates core components than key processes. There are several possible explanations for this difference. One potential explanation is that items are clustered on the VAL-ED instrument by core component and, within core component, key process. This format might exaggerate the differences among core components and minimize the differences among key processes. A second potential explanation is that it may be more difficult for respondents to differentiate the key processes than the core components because respondents think about the results of particular behaviors when rating items. This tendency arose in our earlier cognitive interviews, where respondents indicated a tendency to defer to results in some cases rather than differentiating the planning, for instance, from the implementation (Porter et al., 2010). A third possibility is that there actually are smaller differences in the effectiveness of leadership behaviors across key processes than across core components—perhaps principals tend to implement, support, and monitor programs and policies they have planned with approximately equal effectiveness, whereas they differ in their effectiveness across behavioral targets (e.g., curriculum vs. accountability). It is impossible to say which explanation is true, though it seems likely that all three explanations might contribute to the phenomenon.

The results described here have certain limitations. The most important is that the data used for the analyses are not from “operational” sources in the sense that schools and districts were recruited to participate in a research study and thus may or may not have been using the VAL-ED to formally evaluate their principals. It is impossible to say whether the results would have been the same if the data were operational. The addition of data from schools using the VAL-ED in future years will help us study this potential problem, as well as update the VAL-ED’s norms based on operational data. Nevertheless, the studies described here represent best practice in test construction and validation based on a large national sample of principals. A second potential limitation is with regard to the response rates. Results indicated that Form C was completed at a higher rate than Form A, and that the distribution of response rates was not uniform across strata. Despite this concern, the evidence gathered

suggests that the forms are operating in a parallel way. Nevertheless, the imbalance in response rates was troubling.

The validation of the VAL-ED, like that of any assessment instrument, is ongoing; future work will continue to shed light on the validity of the instrument. Among the future studies to be conducted are criterion validity studies to examine the relation of VAL-ED scores to other criteria of principal effectiveness. These criterion studies include an examination of the convergent and divergent validity of the VAL-ED against other instruments, and an examination of the relationship between VAL-ED scores and school effectiveness as measured by value added to student achievement. Further evidence with regard to the consequences of the VAL-ED will also be collected in a consequences study, where respondents of all types will be surveyed and interviewed to determine the results arising from use of the VAL-ED. As for reliability, a test-retest study will be conducted to examine the extent to which VAL-ED scores are stable over a short time. These studies will add important validity and reliability evidence.

The results described here indicate that the VAL-ED is a reliable instrument for validly measuring principals' learning-centered leadership behaviors. This is especially so when compared with the instruments currently in use, the large majority of which have little connection to the research on effective leadership or documented psychometric properties. If the VAL-ED is used for formative purposes, the findings suggest it can provide important information about relative areas of strength and weakness in a principal's learning-centered leadership. If used for summative purposes, the VAL-ED can help schools and districts make important decisions about staffing and contracts. Together with high-quality professional development and growth plans, the use of the VAL-ED may help improve the quality of principals' learning-centered leadership behaviors in U.S. schools.

Note

The authors gratefully acknowledge the generous support of the Wallace Foundation. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305C050041-05 to the University of Pennsylvania. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The Vanderbilt Assessment of Leadership in Education (VAL-ED) instrument is authored by Drs. Porter, Murphy, Goldring, and Elliott and copyrighted by Vanderbilt University, all of whom receive a royalty from its sales by Discovery Education Assessment. The VAL-ED authors and their research partners have made every effort to be objective and data-based in statements about the instrument, and value the independent peer review process of their research. With any publication, readers in the end must judge the facts and related materials for themselves.

References

- American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Organization.
- 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.

- Borman, W. C. (1997). 360 ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review*, 7(3), 299–315.
- Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 340–370). New York: Macmillan.
- Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York: Russell Sage.
- Carnoy, M., & Loeb, S. (2004). Does external accountability affect student outcomes? A cross-state analysis. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 189–219). New York: Teachers College Press.
- Desimone, L. M., & LeFloch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, 26(1), 1–22.
- Educational Testing Service. (2000). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Fullan, M. (1982). *The meaning of educational change*. New York: Teachers College Press.
- Goldring, E., Cravens, X. C., Murphy, J., Elliott, S. N., Carson, B., & Porter, A. C. (2009). The evaluation of principals: What and how do states and districts assess? *Elementary School Journal*, 110(1), 19–39.
- Goldring, E., Porter, A. C., Murphy, J., Elliott, S. N., & Cravens, X. C. (2009). Assessing learning-centered leadership: Connections to research, professional standards, and current practices. *Leadership and Policy in Schools*, 8(1), 1–36.
- Hallinger, P., & Heck, R. H. (2002). What do you call people with visions? The role of vision, missions, 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.
- Hanushek, E. A., & Raymond, M. A. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2), 297–327.
- Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41(1), 43–62.
- 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.
- Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning*. Minneapolis: University of Minnesota.
- Louis, K. S., Marks, H., & Kruse, S. (1996). Teachers' professional community in restructuring schools. *American Educational Research Journal*, 33(4), 757–798.
- Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works: From leadership to results*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Muraki, E., & Bock, R. D. (1996). *PARSCALE: IRT Based Test Scoring and Item Analysis for Graded Open-Ended Exercises and Performance Tasks*. Chicago: Scientific Software.
- Murphy, J., & Hallinger, P. (1985). Effective high schools: What are common characteristics? *NASSP Bulletin*, 69(477), 18–22.
- 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., & 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.
- Polikoff, M. S., May, H., Porter, A. C., Elliott, S. N., Goldring, E., & Murphy, J. F. (2009). An analysis of differential item functioning on the Vanderbilt Assessment of Leadership in Education. *Journal of School Leadership*, 19(6), 661–679.
- Porter, A. C., Goldring, E., Elliott, S. N., Murphy, J., Polikoff, M. S., & Cravens, X. C. (2008). *Setting performance standards VAL-ED assessment of principal leadership*. (ERIC Document No. ED505799)

- Porter, A. C., Goldring, E., Murphy, J., Elliott, S. N., & Cravens, X. C. (2006). *A framework for the assessment of learning-centered leadership*. Nashville, TN: Vanderbilt University.
- Porter, A. C., Polikoff, M. S., Goldring, E., Murphy, J., Elliott, S. N., & May, H. (2010). Building a psychometrically sound assessment of school leadership: The VAL-ED as a case study. *Educational Administration Quarterly*, *46*(2), 135–173.
- Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. *Elementary School Journal*, *83*(4), 426–452.
- 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.
- Robinson, V. M. J., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential impact of leadership types. *Educational Administration Quarterly*, *44*(5), 635–674.
- Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. Cambridge, MA: Harvard University Press.
- Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometric Monograph*, No. 18.
- Stanley, J. C. (1967). General and specific formulas for reliability of differences. *Journal of Educational Measurement*, *4*, 249–252.