

Developing a Psychometrically Sound Assessment of School Leadership: The VAL-ED as a Case Study

Educational Administration Quarterly
46(2) 135–173
© The University Council for
Educational Administration 2010
Reprints and permission: <http://www.sagepub.com/journalsPermissions.nav>
DOI: 10.1177/1094670510361747
<http://eaq.sagepub.com>



Andrew C. Porter,¹
Morgan S. Polikoff,¹
Ellen Goldring,² Joseph Murphy,²
Stephen N. Elliott,² and
Henry May¹

Abstract

Research has consistently shown that principal leadership matters for successful schools. Evaluating principals on the behaviors shown to improve student learning should be an important leverage point for raising leadership quality. Yet principals are often evaluated with the use of instruments with no theoretical background and little, if any, documented psychometric properties. To address this need, a team of researchers in principal leadership, assessment development, and psychometrics developed the Vanderbilt Assessment of Leadership in Education (VAL-ED). The purpose here is to report on iterative development work where the instrument was tested and revised across several cycles. Future work to investigate the instrument's psychometric properties is identified. After an extensive item writing and instrument

¹University of Pennsylvania, Philadelphia, PA, USA

²Vanderbilt University, Nashville, TN, USA

Corresponding Author:

Morgan S. Polikoff, University of Pennsylvania, 1923 Christian St., Apt B., Philadelphia, PA 19146
Email: Polikoff@dolphin.upenn.edu

development phase, the authors embarked on a series of studies designed to guide improvements to the instrument. These studies include a sorting study, two rounds of cognitive interviews, a bias review, and two rounds of small-scale pilot tests. Results and implications from each study are discussed. The iterative development process helped improve the clarity of instructions and items while building a growing collection of preliminary validity and reliability evidence. At the end of the development process, the VAL-ED represents a promising instrument for assessing principal instructional leadership. The VAL-ED also represents a tool for possible use by principal leadership researchers in measuring the effectiveness of school principals.

Keywords

instructional leadership, assessment, test development, validity, psychometrics


As standards-based reform has increased the accountability pressure on schools to raise student achievement, it has become increasingly apparent that principal leadership matters. Researchers studying effective schools have found that principal leadership is a key element driving school success defined in terms of student achievement (Leithwood, Louis, Anderson, & Wahlstrom, 2004; Marzano, Waters, & McNulty, 2005). Leadership may be especially important in difficult times or in times of organizational change and may separate those institutions that persevere and succeed from those that fail (Murphy, Elliott, Goldring, & Porter, 2006). Highlighting the importance of principal leadership, researchers and policy makers have sought to establish a means to improve the quality of leadership through five key leverage points: standards, licensure, program accreditation, professional development, and leadership evaluation and consequences. Work on the first four leverage points has been robust, as evidenced by the creation, adoption, and revision of standards for school leadership through the Interstate School Leaders Licensure Consortium (ISSLC); the improvement of accreditation through the National Council for Accreditation of Teacher Education; the development of a national licensure examination through the School Leaders Licensure Assessment by Educational Testing Service (ETS); and the establishment of professional development programs tied to the standards (Murphy, Elliott, Goldring, & Porter, 2007). However, leadership evaluation and consequences have seen little traction.

Examination of the leadership assessment field has highlighted weaknesses and the need for more theoretically and psychometrically sound work (Ginsberg & Berry, 1990). Almost every school and district in the country requires some form of principal evaluation, but a recent review of 65 principal evaluation instruments used by districts and states concluded that the instruments almost

universally lacked conceptual frameworks tied to the literature on effective principal leadership. Furthermore, just two of the instruments included information about the psychometric properties of the instruments in their instruction (Goldring, Cravens, et al., 2009). In short, with few valid and reliable methods for evaluating principal leadership, districts and states often turn to homemade instruments with unknown properties.

The paucity of sound principal evaluation instruments is of concern for several reasons. First, the use of evaluation instruments that are not tied to the evidence on effective leadership may encourage principals to modify their behaviors in ways that have little to no effect on school success. Second, if the instruments are being used for high-stakes purposes in principal evaluation, unreliable or invalid measurement of principal effectiveness may result in personnel decisions that negatively affect the school. Third, if designed properly and used correctly, a valid and reliable assessment of principal leadership could be an integral component in a standards-based accountability system. Such an assessment could be used for formative and for summative purposes to help ensure professional growth and establish school and individual growth targets for principals. In other words, focusing principal evaluation on principal behaviors known to be associated with student achievement gains might focus principals on improving their behaviors in these key areas.

With these points in mind, a group consisting of two school leadership researchers, a school psychologist, and a psychometrician began a 3-year project to develop and test an education leadership performance assessment system for measuring the effectiveness of principal leadership behaviors. The core of the assessment system is an instrument that measures leadership behaviors whose conception and foundations have been addressed in earlier reports (Goldring, Porter, Murphy, Elliott, & Cravens, 2009; Murphy, Elliott, Goldring, & Porter, 2006, 2007). The resulting instrument, the Vanderbilt Assessment of Leadership in Education (VAL-ED) is a paper and online assessment that uses a multirater, evidence-based approach to measure the effectiveness of school leadership behaviors known to influence teacher performance and student learning (see the instructions and cover page in Figure 1). The VAL-ED is a 360-degree assessment: teachers, the principal, and the principal's supervisor respond to the behavior inventory. The VAL-ED measures core components and key processes. Core components refer to characteristics of schools that support the learning of students and enhance the ability of teachers to teach. Key processes refer to how leaders create and manage those core components. Effective learning-centered leadership is at the intersection of the two dimensions: core components created through key processes. Thus, the items used to assess the core components are the same items used to assess the key processes.



VANDERBILT ASSESSMENT of LEADERSHIP in EDUCATION™

Teacher Response – Form C

Name of Principal Assessed: Date:

School District: School:

Number of Years Teaching at this School:

Are you a regular classroom teacher in this school? Yes No

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

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

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

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

Completion Tips:

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

Figure 1. Vanderbilt Assessment of Leadership in Education cover page

Results of the VAL-ED are reported on two sets of six scales (core components and key processes) and total score. Results are displayed graphically and in tabular form and include principals', supervisors', teachers', and aggregated mean effectiveness ratings on each scale and total score. National norms are provided, so principals' scores can be reported in norm-referenced ways. Also, performance standards have been created for criterion-referenced evaluation

(Porter et al., 2008). Together with scale scores, norms, and performance levels, supporting text highlights key areas of strength and needed growth.

VAL-ED was designed and developed to be both reliable (i.e., provide accurate measurement) and valid (i.e., measure leadership behaviors that lead to improved student achievement) for use in elementary, middle, and high schools in urban, suburban, and rural settings. The instrument was constructed to (a) work well in a variety of settings and circumstances, (b) be construct valid, (c) be reliable, (d) be unbiased, (e) provide accurate and useful reporting of results, (f) yield diagnostic profiles for formative purposes, (g) be used to measure progress over time in the development of leadership, and (h) predict important outcomes. To accomplish these goals, the research team has followed a multistage development process that involved multiple sources of validity and reliability evidence. At each stage of the design and development process, the properties of the instrument were investigated through empirical study and expert review. The process is guided by the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education [NCME], 1999).

The purpose of this article is to report on the development of the VAL-ED. Investigating the psychometric properties of the instrument in actual use lies ahead.¹ First, the conceptual framework is briefly described. Next, the initial instrument drafting phase is described, with a focus on the establishment of content validity. Third, the results from a series of studies are presented, with a focus on validity and reliability evidence for the instrument. To conclude, we highlight the iterative nature of the work we have done in producing the assessment and reporting of assessment results.

This article contributes to the literature in several ways. As described earlier, the current field of principal leadership evaluation tools is notably lacking. This analysis provides a presentation of initial validity and reliability evidence for a research-based principal instructional leadership assessment for use in elementary, middle, and high schools. The analysis also provides an illustration of a leadership assessment development process guided by the Standards for Educational and Psychological Testing (AERA, American Psychological Association, & NCME, 1999). Finally, the analysis provides information to policy makers, practitioners, and researchers looking for a new way to assess instructional leadership.

Conceptual Framework

The conceptual framework for the instrument is shown in Figure 2 (for a complete description of the conceptual framework, see Goldring, Porter, Murphy,

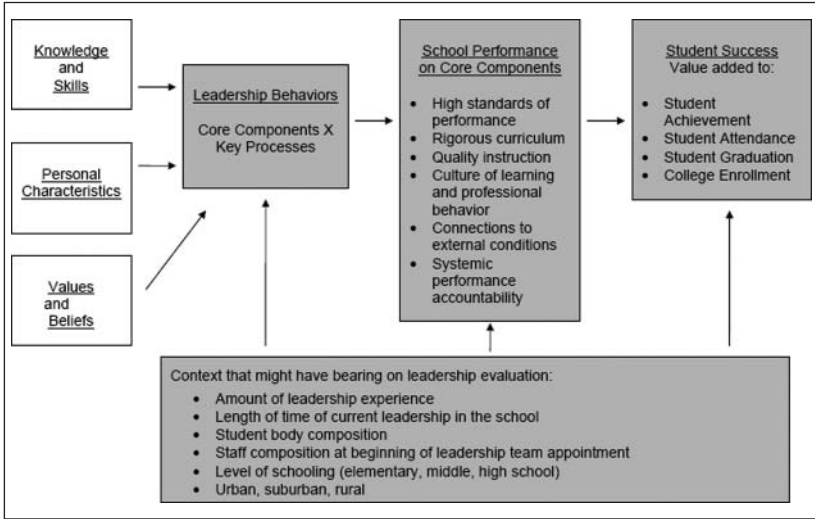


Figure 2. Conceptual model

Elliott, & Cravens, 2009). The model shows leaders’ background skills, characteristics, and beliefs intersecting with school and local context to affect the leadership behaviors at the heart of the instrument. These leadership behaviors, in turn, interact with context to affect the school’s performance in terms of the core components (to be described later). These core components lead to student success in terms of value added to student achievement and other value-added student variables. Although this model envisions indirect effects of principals’ characteristics, attitudes, and values on value added to student achievement, we do not assess those. Instead, we focus on the shaded variables in the model and especially on the key leadership behaviors that affect value added to student achievement through the core components.

The assessment model (see Figure 3) concentrates on two dimensions of leadership behaviors: core components and key processes (Goldring, Porter, Murphy, Elliott, & Cravens, 2009). *Core components* refer to the features of schools that support student learning and teachers’ ability to teach (Marks & Printy, 2003; Sebring & Bryk, 2000). In our model, these are high standards for student learning, rigorous curriculum, quality instruction, culture of learning and professional behavior, connections to external communities, and performance accountability. *Key processes* refer to the leadership behaviors that leaders use to produce the core components (Burns, 1978; Conley & Goldman, 1990; Leithwood, 1994). These are planning, implementing,

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

Figure 3. Conceptual framework

supporting, advocating, communicating, and monitoring. Thus, we envision principals enacting core components through key processes (e.g., planning a rigorous curriculum, supporting performance accountability, monitoring high standards for student learning).

Also playing an important role in the evaluation of principals in our model is the context. In the bottom frame in Figure 2, we highlight some of the key contextual variables that could affect principal leadership. For instance, a brand-new principal might not have the same effect on enactment of the core components as a seasoned veteran who has been in the school for many years. Or a leader in a school with many students with special needs may have more challenges in bringing about certain core components than another principal. Although we recognize that there are legitimate context considerations that should be taken into account, we do not propose that context ever become an excuse for poor-quality leadership behaviors (Goldring, Porter, Murphy, Elliott, & Cravens, 2009). Differences in context do not alter the desirable leadership behaviors, but they could alter how the results of the assessment should be interpreted for the evaluation of a particular principal in a particular school at a particular time.

These conditional interpretations taking into account context are what supervisors are required to do routinely. We do not suggest that a principal's performance should be judged solely on a single assessment of the principal's behaviors.

Instrument Development and Content Validity

In his chapter on validity in the third edition of *Educational Measurement*, Messick (1989) articulates the evolution of definitions for validity and offers a comprehensive update. Messick describes validity as composed of three categories: construct, concurrent, and predictive. Of these three forms of validity, construct validity subsumes two other forms: content and criterion. Focusing on these two subtypes, Messick states that content validity is founded on relevance between the content of the survey and the representativeness with which it covers the domain. Criterion-related validity, Messick argues, is "pointed toward selected relationships with measures that are criterial for a particular applied purpose in a specific applied setting" (Messick, 1989, p. 17). The work reported here focuses on the development of the instrument to be content valid.

In the most recent edition of *Educational Measurement* (4th ed.), Kane's (2006) chapter on validation argues that although this "unified version of construct validity" was well accepted and attractive, it has not offered a clear process for the validation of measures or the purposes for which they would be used. Thus Kane offers the "argument-based" validity approach, which does not negate the unified form of construct validity but adds the necessity of analyzing and testing the interpretive argument of the test or measure by "laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances" (Kane, 2006, p. 23). Investigating this type of validity requires the instrument to be in real use, not administered for research purposes only, as was the case here.

Finally, in a recent issue of *Educational Researcher* devoted to validity, Lissitz and Samuelson (2007) criticize the unitary notion of validity, calling for a change in terminology and emphasis. Their argument points to two critical components for establishing validity: test definition and development (currently known as content validity) and test stability (currently known as reliability). In this conception of validity, "thinking clearly about the content of the assessment is the first step and the most basic step" (Lissitz & Samuelson, 2007, p. 446). Although the process we describe here for validation of the VAL-ED was an iterative process involving multiple steps, we agree that the test development phase is the most basic and essential step in establishing validity.

Instrument Development

The first phase of instrument development began with a thorough examination of the research literature (Murphy, Elliott, Goldring, & Porter, 2006) and creation of the conceptual framework. (The conceptual framework was briefly outlined earlier; for a complete report, see Goldring, Porter, Murphy, Elliott, & Cravens, 2009.) From the 36-cell framework pictured in Figure 3, the process of item writing began. For each cell in the framework, one of the test's two leadership experts first wrote a set of leadership behaviors intended to be exhaustive. The other leadership expert examined several extant principal leadership evaluations to cull additional items that fit into specific cells in the framework. From the first comprehensive list of items, both original and assembled from other instruments, item editing continued with the goal of developing a census of all important leadership behaviors in each cell.

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

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

The item-writing process took place during a span of 7 months and produced an item set with several important characteristics. First, every item written was at an appropriate grain size—neither too broad nor too narrow. Second, the items were a census of the possible items that fit into the two-dimensional framework, with no redundancies. Third, every item fit clearly into a specific cell on the basis of the definitions of the core component and key process that corresponded to that cell.

In addition to the item-writing process, this period saw work on the creation of other features of the assessment. First, the decision was made to have an instrument for each of three response groups: teacher, principal, and supervisor. The rating scale was discussed at length, with proposals for a 5-point scale and a 4-point scale considered. The target of the rating scale was chosen to be "effectiveness," rather than frequency, and the items were specifically worded to point to effectiveness, because of the belief that some behaviors might be important but infrequent. After thorough consideration, the rating scale chosen was a 5-point scale with 1 = *ineffective or not done*, 3 = *moderately effective*, and 5 = *highly effective*. Options 2 and 4 were unlabeled. Finally, for the teacher and supervisor forms only, an option of *don't know* was created, so that respondents who genuinely did not know whether a behavior was done could have an option. *Don't know* was not included on the principal form because principals should know whether they performed any behavior.

Another set of discussions during the instrument development phase focused on "sources of evidence." For each item, respondents are required to indicate the sources of evidence they used to come to their effectiveness rating. The choices are "personal observation," "reports from others," "school documents," "school projects or activities," "other sources," and "no evidence." These sources of evidence are reported prior to making effectiveness ratings but are not included in the calculation of effectiveness scores. The purpose is to have each respondent think hard about each behavior item before making an indication of effectiveness. Because we did not want teachers or supervisors giving effectiveness ratings for items for which they had no evidence, the decision was made that a teacher or supervisor respondent marking "no evidence" would be forced to select either *ineffective or not done* or *don't know*. For principals, selecting "no evidence" required them to mark *ineffective or not done*, because the only way a principal could have no evidence for a behavior would be if he or she did not perform that behavior.

Also important in the instrument development was choosing an appropriate stem. The principal does not have to perform a behavior himself or herself for the behavior to be done. For instance, a principal might not conduct regular classroom observations of every teacher but might work together with other administrators to make sure that routine observations are undertaken. In these instances, we believe that the principal should still be credited for having done the behavior, because he or she ensured that it was done. Hence, the stem chosen was "The principal ensures the school . . ." The stem was not included in every item but, rather, at the top of each page of the instrument.

With these key decisions made, the first complete draft of the instrument and items was ready for examination. Parallel Forms A and C were constructed

so schools could use the instrument in consecutive years without seeing the same items twice. The goal was to focus attention on the domains of behavior represented by each of the 36 cells in the conceptual framework, not the 108 behaviors in one form of the instrument. There were enough items in each cell to allow for random sampling of items from each cell to the forms, initially three per cell.

The item construction and test development phase was the beginning and most important step in building instrument content validity. Again, items were specifically written for each of the 36 cells in the conceptual framework. Items were repeatedly revised by researchers and corrected for grain size, redundancy, clarity, and cell fit.

Additional Tests of Validity and Reliability

Although the item and instrument development phase had established sufficient content validity, we then began an iterative process of detailed psychometric evaluation of the items and the instrument. In this section, we describe each of the tests of validity and reliability and discuss how the results helped improve the instrument.

Sorting Study

A sorting study further investigated the content validity of our assessment. The purpose of the sorting study was to see whether school principals could accurately place items into the 36 cells defined by the intersection of the six core components and six key processes (Figure 3). Nine principals were recruited to the task. Each was provided with the definitions of each core component and each key process and the 36 cell matrix in Figure 3. The pool of 294 items was divided into three random sets stratified by cell. Each set of 98 items was independently sorted by three principals. Items were presented in a random order with no identification as to core component or key process. Principals completed the task off site and on their own timeline.

Results

Eighty-six percent of the classifications into cells of the 294 items resulted in the correct cell identified by at least one of the three principals assigned the item. Fifty-nine percent of the classifications of items were in the exact correct cell by two of three principals assigned the item. Placement in the correct cell is a demanding criterion. When the criterion for classification

was relaxed to ask whether the principal identified the item's correct core component, 75% of the placements were correct. For key processes, 76% of the placements were correct.

Table 1 provides detailed results on the percentage of accurate classifications at the cell, core component, and key process levels for the items in each of the 36 cells in the conceptual framework. Results reveal that some core components and some key processes were easier to classify accurately than were others. High percentage accurate classifications were found for some specific cells: advocating high standards, planning or communicating rigorous curriculum, monitoring quality instruction, communicating culture of learning and professional behavior, and supporting connections to external communities. Each of these five cells had 70% or greater correct classification for the items in that cell. At the other extreme, implementing quality instruction (37%) and implementing performance accountability (36%) were more difficult combinations of key processes by core components to classify. Comparing the first entry in each cell (i.e., percentage of accurate classification at the cell level) to each of the other two entries in the cell identifies whether it was primarily the core component or the key process that created a difficulty in accurate classification. For example, in implementing rigorous curriculum, there was 46% accurate classification at the cell level, 92% accurate classification for the core component, and only 46% accurate classification for the key process. Clearly, it was the key process of implementing that principals had difficulty detecting.

For the key processes, averaged across all core components, planning had 72% accurate classification; implementing, 51%; supporting, 76%; advocating, 86%; communicating, 85%; and monitoring, 88%. Similarly, for core components averaging across key processes, high standards for student learning had 68%; rigorous curriculum, 83%; quality instruction, 71%; culture of learning and professional behavior, 82%; connections to external communities, 81%; and performance accountability, 72%.

Overall, the results of the sorting study indicated that at least for school principals, the behaviors captured by the 294 items were content valid when judged against the conceptual framework of core components by key processes against which the items were written. Several items were revised as a result of the sorting study. The respondents had particular difficulty sorting implementing items correctly, often sorting them into planning or supporting. To address this problem, all planning items were edited to include the words *plan* or *planning*. Additionally, each core component was assigned to a study team member to examine items with significant sorting issues (zero or only one respondent placed the item in the correct cell). The team member

Table 1. Principals' Classification of Items in the Conceptual Framework (in percentages)

Core component	Key processes							Marginals for Core Components
	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring		
High standards for student learning	42	44	63	80	57	37 Total	68	
	63	83	79	87	57	52 CC		
	67	61	79	93	100	74 KP		
Rigorous curriculum (content)	71	46	67	60	78	89 Total	83	
	92	92	67	67	78	89 CC		
	79	46	87	93	100	100 KP		
Quality instruction (pedagogy)	52	37	58	47	67	71 Total	71	
	57	63	79	47	94	81 CC		
	81	57	71	93	67	86 KP		
Culture of learning and professional behavior	53	47	63	58	76	86 Total	82	
	83	80	89	67	90	86 CC		
	60	60	74	88	86	100 KP		
Connections to external communities	67	41	71	63	50	76 Total	81	
	90	82	90	85	62	95 CC		
	71	44	76	78	81	81 KP		
Performance accountability	50	36	63	58	62	64 Total	72	
	63	74	83	67	71	72 CC		
	75	41	75	75	86	92 KP		
Marginals for key Processes	72	51	76	86	85	88		

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

suggested appropriate changes to ensure better fit to the target cell. If no appropriate remedy could be reached, the item was deleted or reworded substantially and assigned to another cell. In all, 67 items were revised or moved, 20 items were deleted, and two items were created as a result of the sorting study. The full study team signed off on each item change. No sorting study of the revised items was done.

Cognitive Interviews

The first examination of the full VAL-ED instrument took place in two rounds of cognitive interviews. Cognitive interviews are helpful in addressing some common threats to survey validity, including the possibility of socially desirable responses and the likelihood of unintentionally misleading directions (Biemer, Groves, Lyberg, Mathiowetz, & Sudman, 1991; Desimone & LeFloch, 2004). There are two stages to a cognitive interview. In the first stage, respondents are asked to “think aloud” as they answer questions or read directions. Here, respondents are asked to describe their thought process as it occurs, providing as much detail as possible. Whether the item is clear or ambiguous, respondents speak whatever is on their mind. In the second stage, interviewers ask specific questions of respondents about item or response choice interpretation (for a full description of the cognitive interview methodology, see Desimone & LeFloch, 2004).

Two rounds of cognitive interviews with three sets of interviews each were conducted. In the first round, the interviews were conducted in one school each in three urban districts—a middle school in St. Louis, a high school in Louisville, and an elementary school in Nashville. In each district, there were three respondents: a principal, one of the principal’s teachers, and a supervisor of principals. Both forms of the assessment instrument were studied.

First Round

For the first round of cognitive interviews, participants were introduced to the cognitive interview “think-aloud” methodology with an example. Next, respondents were asked to read aloud and examine the study’s cover page and directions and comment on language, aesthetics, and clarity. They were also asked probing questions about particular phrases and words the investigators anticipated being problematic. In the next step, respondents read the survey aloud, item by item, describing their thought process as they identified sources of evidence and checked effectiveness ratings. The 108 items were randomly arranged. Respondents were asked to complete the assessment as if they were

actually using it to evaluate a principal. Periodically, the researcher would stop the respondent to ask questions about the respondent's interpretation of key words and phrases. Finally, at the end of the interview, respondents were asked several questions about the instrument and its likely utility in the field. After the first round of cognitive interviews, the research team met and examined cognitive interview data to make improvements to instructions, formatting, and individual item wording.

The primary results from the first round of cognitive interviews had to do with instrument instructions, the response scale, sources of evidence, and the item stem. One respondent was confused by the part of the instructions that read, "In some cases, the principal may not have actually performed the behavior described, but he or she ensured that it was performed." This critical piece of the instructions aligned with the item stem "The principal ensures the school . . ." and reinforced the idea that the principal should not be expected to personally carry out every behavior in the behavior inventory but could delegate a particular responsibility and ensure the task was performed. Respondents also had difficulty with the format and layout of the instructions page—many thought it was "wordy" and several were confused by the example items. As a result of these challenges, the survey was reformatted after the first round with bullet points for directions and a clearer example.

Respondents expressed few concerns about the effectiveness rating scale in the first round. The scale was presented in a 1-to-5 format, with 1 representing *ineffective or not done*, 3 representing *moderately effective*, and 5 representing *highly effective*. There were no labels defining the 2 and 4 ratings, and two respondents felt that labels would be an improvement. However, these respondents still used Categories 2 and 4 frequently. Other respondents felt a *not applicable* response was needed for certain items that were not relevant to their school (for example, one teacher said her school did not have any English language learners (ELLs), so the item referencing ELLs was irrelevant). Respondents have the choice of *don't know*, so they should opt for this response if they believe an item is not applicable. Some respondents noted that 1 = *ineffective or not done* was double-barreled. Responding to this concern, we chose to change to a 0-to-5 scale, with 0 representing *not done* and 1 representing *ineffective*.

Other results from the first round of cognitive interviews focused on concerns about sources of evidence. One respondent said he was using the number of sources of evidence checked as a gauge of effectiveness. Although the sources of evidence were intended to help respondents consider the support they had for their effectiveness rating, no correlation should necessarily exist between the number of sources marked and the effectiveness rating for that

item. Some respondents thought source-of-evidence choices were missing from the instrument, but these respondents indicated that they used the “other sources” category to indicate missing response categories. When asked about the inclusion of sources of evidence and their placement before the effectiveness rating scale, all first-round respondents indicated that the design was useful. Several of the respondents labored over the sources of evidence, however, thinking carefully through each one before marking their response. This appeared to indicate that the respondents were thinking carefully about their effectiveness rating. However, the amount of effort many respondents were putting into the sources of evidence greatly raised the cognitive demand of the task, and at 108 items, respondents were already concerned about length. No changes to the instrument were made as a result of these comments.

Finally, respondents in the first round seemed to periodically forget the stem, “The principal ensures the school . . . ,” focusing instead on whether the principal performed the behavior directly. This problem led to adding the stem to each item after the first round to ensure respondent understanding.

Second Round

The second round of cognitive interviews was conducted with three respondents each in a Chicago elementary school and a Fairfax County, Virginia, middle school, and two respondents (no supervisor could participate) in a Nashville high school. In this second round, respondents first completed the assessment on their own, making notes by items they wanted to discuss. A change in the format of the survey items was also included in the interviews—the 108 items were organized by core component and, within core component, key process. Respondents completed the form without interruption. When respondents were done with the survey, the researcher probed them on key words and phrases that still seemed potentially unclear after the first round of edits. This modified interview methodology was used to give the research team a better idea of whether respondents could successfully complete the instrument without additional support.

There were no new problems that arose during the second round of interviews. As in the first round, there was a propensity of the interviewee to defer to outcomes when determining an effectiveness rating. For example, when an item indicated that the principal ensures that the school plans a rigorous curriculum, the rating was given on whether a rigorous curriculum existed, which reveals some combination of good planning and good implementation, two separate key processes. This “bleeding” of processes caused us to analyze and revise items in many instances to more fully distinguish between the

key processes. Such bleeding of categories also occurred at times between the core component of performance accountability and the key process of monitoring crossed with other core components. Once again, potential revisions were debated and changes were implemented at the conclusion of the second round of cognitive interviews.

Postsurvey questions addressed concerns about the best way to introduce and implement the VAL-ED instrument to our key users of teachers, principals, and supervisors of principals. We asked interviewees how much time we could reasonably expect future participants to spend on a survey such as ours; the vast majority agreed on a time of approximately 30 min. When we asked about preference between paper-and-pencil or online format, most interviewees indicated a preference for online, despite the fact that they saw only a paper-and-pencil version; they were split on whether others would prefer one format to the other. When asked how teachers would react to taking a survey such as ours, nearly every individual felt the survey was comprehensive, touching on all of the right behaviors, but length was an issue. When interviewees were asked whether anything was missing from the survey, a few suggested that having data on some traditional outcomes, such as student achievement, would have helped them with their ratings. This indicated a tendency by some individuals to again defer to outcomes regardless of the key process the item was seeking to highlight. Overall, however, the response was that the instrument was inclusive, sometimes even redundant, and that it seemed to capture key principal leadership behaviors.

Finally, the remainder of the feedback had to do with specific items and phrasing. A problem that persisted across both rounds was with the term *leaders* in items such as “The principal ensures the school allocates leaders’ time to support a system that holds students accountable for their learning.” Some respondents thought that administrators were leaders, but one principal thought that every teacher in the school was a leader. Forceful terms such as *ensure* and *cause* often created problems for respondents across rounds, who felt that, for instance, nothing could “ensure students would meet high standards.” This concern indicated a need to soften the language to more closely approximate the intended meaning of the item. As with concerns about item bleeding, these concerns were addressed by a small set of item revisions after the second round.

Overall, the first round of cognitive interviews provided important information about the sources of evidence, the instrument’s instructions, and the item stem. Modifications made to the instrument between rounds were examined in the second round of interviews, with the changes adequately addressing most of the previous concerns. Although the second round did not

provide evidence of substantial new concerns, it did provide an important check of the changes made to the instrument. Importantly, the second round also provided evidence of respondents' being able to complete the task independently on their own.

Item Bias Study

To consider bias, a fairness review of the VAL-ED instructions and items was conducted. The purpose was to identify and remove aspects of items or directions that might hinder respondents from various groups from completing the instrument as intended and might lead to inappropriate inferences about a principal's behavior. The fairness review was based on the test fairness guidelines published and used by ETS (2000):

Guideline 1. Treat people with respect in test materials.

Guideline 2. Minimize the effects of construct-irrelevant knowledge or skills.

Guideline 3. Avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting.

Guideline 4. Use appropriate terminology.

Guideline 5. Avoid stereotypes.

Guideline 6. Represent diversity in depictions of people.

The fairness review was conducted via individual electronic surveys to each panelist followed by a Webex conference after all surveys were returned. Nine individuals with knowledge of testing and rating scale methods were selected to participate on the panel. Of the nine members, six were female and three male, and all but one currently worked in public schools as either a teacher, behavior specialist, or administrator. The non-school-based person worked in the testing industry as an editor. The panel members self-identified themselves as four Caucasians, two Hispanics, two African Americans, and one Asian American. Three respondents had a PhD, two had a master's degree, three had a bachelor's degree, and one had a high school degree. Collectively, the panel members represented six regions of the country.

The respondents were trained about the six ETS (2000) fairness guidelines using a 21-slide PowerPoint show. The PowerPoint presentation was reviewed independently by all individuals, then reviewed and discussed briefly by the group on a conference call. At the end of this training phase, all panel members reported that they understood the fairness guidelines and felt confident that they could apply them to the review of rating scale items.

Item Number	Item Content	# of Panelists Who Indicated Problems	Fairness Guideline Cited
Form A Item 8	... challenges low expectations for special needs students. ... challenges low expectations for students with special needs.	7	3, 4, 5
Form A Item 56	... challenges teachers to work with community agencies to support students at risk. ... challenges teachers to work with community agencies to support students' needs.	3	1, 4, 5
Form A Item 67	... challenges faculty who blame others for student failure. ... challenges faculty who attribute student failure to others.	5	1, 2, 3, 4
Form C Item 17	... supports teachers to participate in professional development that deepens their understanding of the rigorous curriculum ... supports participation in professional development that deepens teachers understanding of a rigorous curriculum.	3	1, 2, 4

Figure 4. Vanderbilt Assessment of Leadership in Education items identified as potentially unfair and suggested revisions

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

Results

The panelists worked independently through both forms of the VAL-ED and recorded fairness guideline violations for the instrument’s instructions and each item. The aggregated results of all nine panelists indicated no fairness concerns with the VAL-ED instructions or introductory content. With regard to Form A, two or more panelists identified 13 items that raised a fairness concern and possible violation. On Form C, the panelists identified 14 items that raised a fairness concern and possible violation. From this total pool of 27 items, four items were perceived to be a serious concern for three or more panelists (see Figure 4). These items and the identified type of violation were

discussed on a conference call with all panelists together. The end result of the discussion was suggested revisions for each of these items. These revisions are documented in Figure 4 in boldface type. A review of the items indicates that three of them concerned the leadership behavior process of advocating. The subtle, but meaningful, suggested changes for these items emphasized person-first language. The authors of the VAL-ED reviewed the panelists' suggested item revisions and accepted them.

Nine-School Pilot Test

Process

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

All contact with schools was coordinated through a designated liaison. Survey forms were sent to the liaison, and she sent them to each school to be completed. No instructions were given as to the setting in which the assessment was to be completed. Members of the VAL-ED research team traveled to the schools to collect the forms 2 to 3 days after the schools received the forms. Respondents were also provided with postage-paid envelopes if they wished to mail back additional completed forms. In each participating school, the principal, his or her supervisor, and all teachers in the school were requested to participate. Teachers were assured of confidentiality. To encourage high response rates, a graded system of incentives was implemented. Schools received \$500 for participating, but the incentive increased to \$750 for 75% teacher response rate and to \$1,000 for 90% teacher response rate.

An important issue that arose in the pilot study related to the supervisor's ratings. Only one supervisor evaluated the principals from each level of school. The elementary school supervisor rated each of his or her three principals as highly effective on all items, for overall ratings of 5.00 for the elementary school principals. These data suggest that the supervisor did not take the exercise of rating the principals seriously. This may be because of the fact that the pilot study was not taken under "high-stakes" conditions. That is, no accountability was associated with the ratings provided, so supervisors

(and other respondents) may not have given the same ratings they would actually give under conditions of regular use.

Results

Feasibility. The first element of feasibility simply asks whether respondents completed the assessment. Response rates from the pilot study suggest that teachers and supervisors are willing to complete the VAL-ED. Of the nine schools, two had 100% teacher response rates, and three others had teacher response rates of greater than 90%. One school had a response rate between 75% and 90%, and the remaining three schools had response rates of 39%, 41%, and 58%. The overall teacher response rates were 70% for Form A and 75% for Form C (72.5% overall). Response rates were 70% or greater for each level of school. Nine of nine supervisor forms were completed, and eight of nine principal forms were completed. A total of 319 teacher responses were collected: 153 on Form A and 166 on Form C. Possible response bias from teachers was not investigated; to ensure teachers' anonymity, no demographic information was collected from respondents.

A second element of feasibility concerns whether respondents completed individual items. There are two ways in which respondents could choose to not rate a principal: They could leave an item blank (missing data), or they could select the *don't know* option. Principals did not have the option of selecting *don't know*, but they left 0% of items blank. Supervisors selected *don't know* 4% of the time and left no items blank. For teachers, 1.7% of items were left blank, and 6.1% of items were marked *don't know*. Results at the scale level, shown in Table 2, reveal that certain scales had higher-than-average rates of *don't know* responses. All 12 scales had low missing-data rates. However, two core components—connections to external communities and performance accountability—and two key processes—advocating and monitoring—had higher proportions of *don't know* ratings, with proportions greater than 10% on one or both forms. At the item level, no items had more than 6% missing data. Six items on Form A and one item on Form C had more than 25% *don't know* ratings, but the majority of items on both forms had less than 10% *don't know* ratings. In short, missing data were not a problem at the item, scale, or form level for any respondent group. Clearly, there are some core components and some key processes for which supervisors and teachers were not in a position to evaluate the effectiveness of the principal's behavior. When an item is answered *don't know*, the item for that respondent is deleted from the scales in which it fits, and the mean item response is based on the reduced number of items for that scale.

Table 2. Teacher Missing Data and Don't Know Responses by Scale (in percentages)

Scale	Teacher Rating Distribution by Form, Nine-School Pilot, Spring 2007							Missing/ Not Entered
	0	1	2	3	4	5	Don't Know	
Form C								
High standards	0.3	0.7	2.3	10.1	28.5	55.5	1.3	1.3
Rigorous curriculum	0.2	0.9	2.0	11.7	28.1	52.1	3.0	2.0
Quality instruction	0.1	0.7	1.8	10.6	28.0	52.9	3.4	2.4
Culture of learning	0.4	0.4	2.1	10.3	29.5	53.1	2.2	2.1
Connections to communities	0.4	0.7	2.9	14.6	27.0	40.9	11.5	1.9
Performance Accountability	0.5	0.6	2.4	12.2	26.9	47.9	7.9	1.7
Planning	0.3	0.6	2.2	11.5	28.9	50.4	3.7	2.3
Implementing	0.1	0.6	2.4	11.9	27.8	51.1	4.0	2.1
Supporting	0.2	0.7	2.6	10.1	28.6	52.6	2.9	2.3
Advocating	0.4	0.7	2.5	13.0	29.2	46.5	6.2	1.4
Communicating	0.1	0.8	1.9	11.6	27.4	52.5	4.2	1.5
Monitoring	0.7	0.7	2.0	11.2	26.2	49.3	8.2	1.8
Total	0.3	0.7	2.3	11.6	28.0	50.4	4.9	1.9
Form A								
High standards	0.3	0.4	1.0	9.5	33.0	50.9	3.1	1.7
Rigorous curriculum	0.1	0.4	2.0	10.3	34.0	48.8	3.7	0.7
Quality instruction	0.5	0.6	2.0	9.7	29.1	51.7	4.8	1.5
Culture of learning	0.6	0.7	1.8	9.6	27.7	52.4	4.7	2.4
Connections to communities	1.5	0.6	2.2	12.4	28.3	36.5	15.9	2.6
Performance Accountability	1.2	0.8	2.0	11.6	29.1	40.5	12.4	2.5
Planning	0.7	0.8	1.7	10.2	32.0	47.2	6.0	1.4
Implementing	0.7	0.5	1.9	10.4	31.2	49.1	4.3	1.9
Supporting	0.4	0.5	1.5	8.9	29.5	52.9	4.4	1.9
Advocating	1.0	0.7	2.1	12.5	29.1	42.3	10.1	2.2
Communicating	0.3	0.3	2.1	10.3	30.3	48.2	6.7	1.7
Monitoring	1.1	0.7	1.6	10.5	28.8	41.8	13.2	2.4
Total	0.7	0.6	1.8	10.5	30.2	46.9	7.4	1.9

Analysis of the sources of evidence used in the pilot study is provided in Table 3. Results show that all respondent groups were most likely to indicate personal observation; roughly 70% of items had personal observation as evidence. Principals and supervisors selected more sources of evidence than

Table 3. Sources of Evidence Used: Mean Percentage of Items With Each Kind of Evidence by Respondent, Form, School Type, Nine-School Pilot, Spring 2007

	Teacher	Supervisor	Principal	Form		Elementary	Middle	High	Overall
				A	C				
Reports from others	28.0	48.1	43.5	28.7	28.8	26.4	27.2	31.8	28.8
Personal observation	71.0	69.7	73.6	70.1	71.7	71.4	74.1	67.3	70.9
School documents	45.3	79.9	74.5	45.8	47.4	49.1	52.8	38.6	46.6
School projects or activities	32.7	42.9	42.7	36.3	29.9	36.3	34.6	29.3	33.0
Other sources	11.4	1.4	28.6	10.1	12.7	15.8	11.2	9.1	11.5
No evidence	2.7	2.7	0.7	3.1	2.2	2.5	2.1	3.3	2.6
Average number of sources	1.91	2.45	2.64	1.94	1.93	2.02	2.02	1.79	1.93

teachers, especially school documents. Additionally, elementary and middle school respondents marked more sources of evidence than did high school respondents. Again, sources of evidence are included in the assessment to facilitate the respondents' thinking carefully about each item; sources of evidence are not used in calculating the effectiveness scores.

Data on the distribution of effectiveness ratings also provide evidence about the feasibility of the response scale. The percentages shown in Table 4 reveal that ratings were high. On the 0-to-5 scale, with 0 representing *not done* and 5 representing *highly effective*, most scales had roughly 80% of teacher ratings at the 4 or 5 levels. Overall, roughly 30% of items were rated a 4, and 47% of items were rated a 5. Approximately 10% of items were rated a 3, and 3% of items were rated a 0, 1, or 2. Teacher item-level means had a roughly normal distribution, with a mean item response of approximately 4.4 on each form. Except for one outlier item, item means ranged from 3.9 to 4.7. Teacher item standard deviations ranged from 0.6 to 1.3, with a mean item standard deviation of 0.95. The item distribution results suggest either that the principals were extremely effective or that the VAL-ED forms used in this pilot study experienced a common issue with behavior rating scales—the tendency of respondents to give very high ratings overall, possible evidence of the presence of construct-irrelevant variance (Messick, 1994). Both explanations seemed likely in this case; the district had a reputation for successful leadership reform.

The fifth and final component of feasibility to be discussed is the respondents' reactions to questions about the VAL-ED's feasibility. Respondents

Table 4. Teacher Rating Distributions by Form, Nine-School Pilot, Spring 2007
(in percentages)

Scale	Teacher Ratings					
	0	1	2	3	4	5
Form C						
High standards	0.3	0.7	2.3	10.1	28.5	55.5
Rigorous curriculum	0.2	0.9	2.0	11.7	28.1	52.1
Quality instruction	0.1	0.7	1.8	10.6	28.0	52.9
Culture of learning	0.4	0.4	2.1	10.3	29.5	53.1
Connections to communities	0.4	0.7	2.9	14.6	27.0	40.9
Performance	0.5	0.6	2.4	12.2	26.9	47.9
Accountability						
Planning	0.3	0.6	2.2	11.5	28.9	50.4
Implementing	0.1	0.6	2.4	11.9	27.8	51.1
Supporting	0.2	0.7	2.6	10.1	28.6	52.6
Advocating	0.4	0.7	2.5	13.0	29.2	46.5
Communicating	0.1	0.8	1.9	11.6	27.4	52.5
Monitoring	0.7	0.7	2.0	11.	26.2	49.3
Total	0.3	0.7	2.3	11.6	28.0	50.4
Form A						
High standards	0.3	0.4	1.0	9.5	33.0	50.9
Rigorous curriculum	0.1	0.4	2.0	10.3	34.0	48.8
Quality instruction	0.5	0.6	2.0	9.7	29.1	51.7
Culture of learning	0.6	0.7	1.8	9.6	27.7	52.4
Connections to communities	1.5	0.6	2.2	12.4	28.3	36.5
Performance	1.2	0.8	2.0	11.6	29.1	40.5
Accountability						
Planning	0.7	0.8	1.7	10.2	32.0	47.2
Implementing	0.7	0.5	1.9	10.4	31.2	49.1
Supporting	0.4	0.5	1.5	8.9	29.5	52.9
Advocating	1.0	0.7	2.1	12.5	29.1	42.3
Communicating	0.3	0.3	2.1	10.3	30.3	48.2
Monitoring	1.1	0.7	1.6	10.5	28.8	41.8
Total	0.7	0.6	1.8	10.5	30.2	46.9

Percentages going across add to 100% when missing and *don't know* values from Table 2 are added. Percentages may not add to 100% because of rounding.

were asked to answer six items on the final page of the assessment, with response categories of 1, *strongly disagree*; 2, *disagree*; 3, *agree*; and 4, *strongly agree*. Results appear in Table 5. The three most important items

Table 5. Responses to Feasibility Questions by Respondent, Nine-School Pilot, Spring 2007

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

1 = *strongly disagree*, 4 = *strongly agree*.

from a feasibility and validity standpoint are Items 1, 2, and 6. Teachers and supervisors leaned toward agreement that the response form was easy to use, whereas principals were neutral. Teachers, principals, and supervisors also leaned toward agreement that (a) the items focused on important leadership behaviors and (b) they understood the items. All three respondent groups were neutral in their views of (a) using the instrument every year and (b) supporting the instrument's use in their district. As for checking sources of evidence, all three groups were just above neutral as to usefulness. Additional space on the form was left for respondent comments. Ninety-nine respondents left comments; 81 suggested that the form was too long or too repetitive. Given the complaints about time required for completion, the neutral assessment of use is surprisingly positive.

Reliability. An important component of any assessment instrument is its reliability. There are many forms of reliability; in the pilot study, only

Table 6. Estimates of Internal Consistency Reliability, Nine-School Pilot, Spring 2007

	Cronbach's α	
	Form A	Form C
High standards for student learning	.95	.97
Quality instruction	.94	.95
Rigorous curriculum	.95	.97
Culture of learning	.93	.96
Connections to external community	.95	.97
Performance accountability	.95	.97
Planning	.92	.95
Implementing	.94	.95
Supporting	.93	.96
Advocating	.94	.96
Communicating	.94	.97
Monitoring	.93	.96
Total	.99	.99

internal-consistency reliability could be estimated. Reliabilities for both forms and all scales were high. Cronbach's alpha reliabilities for teacher scores are presented in Table 6. For all 12 scales on both forms, reliabilities were near perfect. For the total score, reliabilities were greater than .98 on both forms. Reliabilities tended to be somewhat higher for core components than for key processes. The sample size of schools in the pilot was too small to accurately estimate reliability for supervisors or principals.

Validity. Confirmatory factor analysis using teacher data was done to investigate data fit to our conceptual model. Again, the pilot study did not have a sufficient sample of schools ($n = 9$) to investigate factor structure for supervisors or principals. The factor analytic model was designed to parallel the conceptual framework for the VAL-ED by incorporating higher-order factors for core components, key processes, and an overall score. Thus, the hierarchical factor analytic model had four levels. The first level involved the 108 individual items, which were endogenous to latent factors for the 36 cells representing six core components crossed with six key processes at the second level. At the third level were latent factors for the six core components or key processes. At the fourth level was a single latent trait representing overall principal leadership (i.e., the total score). Because each item contributed to both a core component and a key process, the factor analytic model was split into two separate analyses: one on core components and the other on key processes.

Item	Level 1 Factor Loading	Level 1 Factor	Level 2 Factor Loading	Level 2 Factor (Core Components)	Level 3 Factor Loading	Level 3 Factor	Level 3 Factor Loading	Level 3 Factor (Core Components)	Level 2 Factor Loading	Level 2 Factor	Level 1 Factor	Level 1 Factor Loading	Item
Item 1	.76	HighStandards_Planning	.83	High Standards	.90	Overall Score	.89	Culture of Learning	1.08	Culture of Learning_Planning	.81	Item 55	
Item 2	.84									.76	Item 56		
Item 3	.72									.55	Item 57		
Item 4	.81									.55	Item 58		
Item 5	.77									.55	Item 59		
Item 6	.85	HighStandards_Implementing	.83	High Standards	.90	Overall Score	.89	Culture of Learning	1.08	Culture of Learning_Implementing	.85	Item 60	
Item 7	.76									.74	Item 61		
Item 8	.80									.83	Item 62		
Item 9	.76									.76	Item 63		
Item 10	.80									.76	Item 64		
Item 11	.85	HighStandards_Supporting	1.00	High Standards	.90	Overall Score	.89	Culture of Learning	1.08	Culture of Learning_Supporting	.83	Item 65	
Item 12	.80									.80	Item 66		
Item 13	.79									.80	Item 67		
Item 14	.76									.82	Item 68		
Item 15	.80									.80	Item 69		
Item 16	.81	HighStandards_Advocating	.84	High Standards	.90	Overall Score	.89	Culture of Learning	1.08	Culture of Learning_Advocating	.80	Item 70	
Item 17	.80									.76	Item 71		
Item 18	.76									.80	Item 72		
Item 19	.87									.72	Item 73		
Item 20	.80									.79	Item 74		
Item 21	.76	.76	Item 75										
Item 22	.75	RigorousCurriculum_Planning	.83	Rigorous Curriculum	.94	Overall Score	.89	External Communities	1.08	ExtCommunities_Planning	.80	Item 76	
Item 23	.73									.87	Item 77		
Item 24	.80									.76	Item 78		
Item 25	.81									.79	Item 79		
Item 26	.81									.80	Item 80		
Item 27	.89	RigorousCurriculum_Implementing	.87	Rigorous Curriculum	.94	Overall Score	.89	External Communities	1.08	ExtCommunities_Implementing	.85	Item 81	
Item 28	.80									.79	Item 82		
Item 29	.85									.80	Item 83		
Item 30	.76									.82	Item 84		
Item 31	.83									.80	Item 85		
Item 32	.87	RigorousCurriculum_Supporting	.86	Rigorous Curriculum	.94	Overall Score	.89	External Communities	1.08	ExtCommunities_Supporting	.81	Item 86	
Item 33	.83									.86	Item 87		
Item 34	.86									.79	Item 88		
Item 35	.76									.73	Item 89		
Item 36	.80									.80	Item 90		
Item 37	.83	RigorousCurriculum_Advocating	1.00	Rigorous Curriculum	.94	Overall Score	.89	External Communities	1.08	ExtCommunities_Advocating	.84	Item 91	
Item 38	.80									.80	Item 92		
Item 39	.80									.80	Item 93		
Item 40	.82									.71	Item 94		
Item 41	.82									.84	Item 95		
Item 42	.88	.73	Item 96										
Item 43	.84	RigorousCurriculum_Communicating	1.00	Rigorous Curriculum	.94	Overall Score	.89	External Communities	1.08	ExtCommunities_Communicating	.80	Item 97	
Item 44	.79									.86	Item 98		
Item 45	.76									.80	Item 99		
Item 46	.87									.80	Item 100		
Item 47	.70									.85	Item 101		
Item 48	.77	.85	Item 102										
Item 49	.82	QualityInstruction_Planning	.82	Quality Instruction	.97	Overall Score	.95	Performance Accountability	1.08	AccAccountability_Planning	.84	Item 103	
Item 50	.83									.75	Item 104		
Item 51	.77									.83	Item 105		
Item 52	.72									.80	Item 106		
Item 53	.84									.85	Item 107		
Item 54	.80	.77	Item 108										
Item 55	.80	QualityInstruction_Implementing	.86	Quality Instruction	.97	Overall Score	.95	Performance Accountability	1.08	AccAccountability_Implementing	.86	Item 109	
Item 56	.80									.80	Item 110		
Item 57	.80									.80	Item 111		
Item 58	.80									.80	Item 112		
Item 59	.80									.80	Item 113		
Item 60	.80	QualityInstruction_Supporting	.81	Quality Instruction	.97	Overall Score	.95	Performance Accountability	1.08	AccAccountability_Supporting	.80	Item 114	
Item 61	.80									.80	Item 115		
Item 62	.80									.80	Item 116		
Item 63	.80									.80	Item 117		
Item 64	.80									.80	Item 118		
Item 65	.80	QualityInstruction_Advocating	1.00	Quality Instruction	.97	Overall Score	.95	Performance Accountability	1.08	AccAccountability_Advocating	.80	Item 119	
Item 66	.80									.80	Item 120		
Item 67	.80									.80	Item 121		
Item 68	.80									.80	Item 122		
Item 69	.80									.80	Item 123		
Item 70	.80	QualityInstruction_Communicating	.89	Quality Instruction	.97	Overall Score	.95	Performance Accountability	1.08	AccAccountability_Communicating	.80	Item 124	
Item 71	.80									.80	Item 125		
Item 72	.80									.80	Item 126		
Item 73	.80									.80	Item 127		
Item 74	.80									.80	Item 128		
Item 75	.80	QualityInstruction_Monitoring	.82	Quality Instruction	.97	Overall Score	.95	Performance Accountability	1.08	AccAccountability_Monitoring	.77	Item 129	
Item 76	.80									.80	Item 130		
Item 77	.80									.80	Item 131		
Item 78	.80									.80	Item 132		
Item 79	.80									.80	Item 133		

Figure 5. Core components confirmatory factor analysis for nine-school pilot data

Results from the confirmatory factor analyses reveal that both the core components and the key processes models fit the data well. The results for core components on Form A are shown in Figure 5; the three other sets of results are simply summarized here, because they are so similar. Across the four confirmatory factor analyses, goodness-of-fit indices were between .96 and .99. Even after adjusting for model complexity, the parsimonious goodness-of-fit indices (Mulaik et al., 1989) were still high, ranging from .93 to .96. All of the item factor loadings were salient, ranging from 0.41 to 0.94, with a median loading of 0.82. The second-order factor loadings were also salient, ranging from 0.60 to 1.00, with a median loading of 0.92. Last, the third-order factor loadings were salient, ranging from 0.89 to 1.00, with a median loading of 0.98. The increase in saliency across levels and the consistently high loadings at Level 3 suggest that the core components and the key processes have similar degrees of influence on the total score. In other words, the six core components and six key processes all contribute to the overall measure of principal leadership.

A second piece of validity evidence was obtained by examining the relationship of teacher ratings and principal ratings. The problem noted earlier

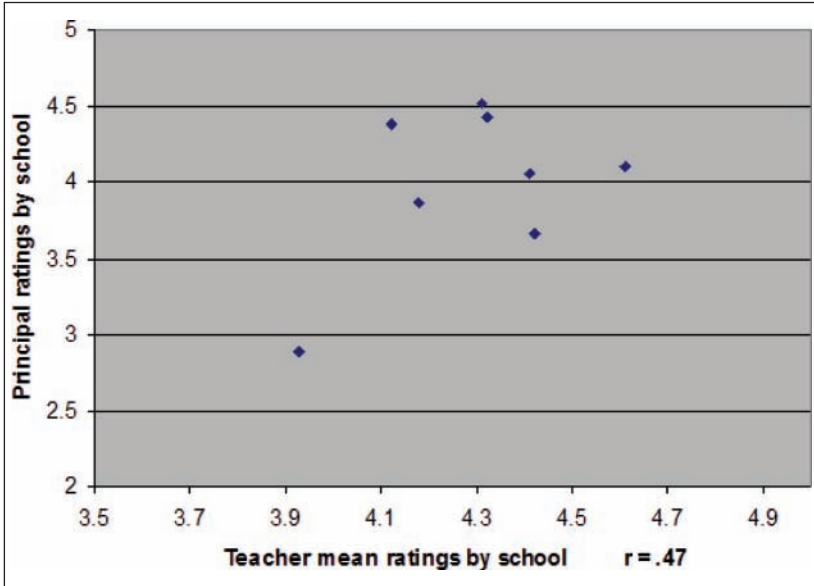


Figure 6. Scatter plot of principal ratings with mean teacher ratings for nine-school pilot

with the supervisor ratings resulted in no believed principal variance based on supervisor data. A scatter plot of teacher and principal ratings is found in Figure 6. There are only 8 data points because one principal did not complete the assessment. The scatter plot suggests that principals and teachers tended to give similar ratings of principals' effectiveness. For example, the principal who gave himself or herself the lowest score was also rated the lowest by his or her teachers. The correlation of principal and teacher ratings in these 8 data points is a moderate .47. This finding suggests that the between-principal variance for both teacher and principal data is measuring something in common, a kind of concurrent validity. Furthermore, these correlations are typical to slightly higher than between-group correlations on 360 or multi-rater assessments, which are generally .25 to .35 (Atwater, Ostroff, Yammarino, & Fleenor, 1998; Harris & Schaubroeck, 1988).

As seen in Tables 7 and 8, the correlations among core components and among key processes were high, although they appear somewhat higher for key processes. For core components, correlations ranged from a low of .73 (connections to external communities and high standards for student learning)

Table 7. Intercorrelations of Core Components, All Schools, Nine-School Pilot, Spring 2007

	High Standards	Instruction	Curriculum	Culture	Connections	Performance Accountability
High standards	—					
Instruction	0.91	—				
Curriculum	0.84	0.90	—			
Culture	0.81	0.84	0.85	—		
Connections	0.73	0.78	0.81	0.81	—	
Performance accountability	0.79	0.83	0.84	0.79	0.83	—
Total	0.91	0.95	0.94	0.92	0.90	0.92

Table 8. Intercorrelations of Key Processes, All Schools, Nine-School Pilot, Spring 2007

	Planning	Implementing	Supporting	Advocating	Communicating	Monitoring
Planning	—					
Implementing	0.93	—				
Supporting	0.91	0.92	—			
Advocating	0.91	0.93	0.90	—		
Communicating	0.92	0.92	0.92	0.92	—	
Monitoring	0.91	0.91	0.89	0.91	0.94	—
Total	0.96	0.97	0.96	0.96	0.97	0.96

to a high of .90 (quality instruction and high standards for student learning). For key processes, correlations ranged from a low of .89 (supporting and monitoring) to a high of .94 (monitoring and communicating). Correlations of core components and key processes with total score were all quite high, with none lower than .9. These high intercorrelations, along with the factor analysis results described earlier, suggest that the instrument is measuring a strong underlying construct, principal leadership. Given the high collinearity of subscales, the support for construct validity from the confirmatory factor analysis is somewhat surprising. Perhaps principals effective in one area tend to be effective across the board, much as is true for student achievement. Alternatively, respondents may find making the distinction among core components and key processes difficult.

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

Table 9. Teacher Ratings by Form and Respondent Type, Nine-School Pilot, Spring 2007

Scale	Teacher Mean	
	Form A	Form C
High standards	4.39	4.38
Rigorous curriculum	4.34	4.35
Quality instruction	4.36	4.38
Culture of learning	4.37	4.38
External communities	4.15	4.19
Performance accountability	4.21	4.30
Planning	4.31	4.33
Implementing	4.31	4.34
Supporting	4.40	4.36
Advocating	4.22	4.27
Communicating	4.34	4.36
Monitoring	4.26	4.33
Total	4.31	4.33

similarly shaped with similar ranges. Table 9 shows teacher mean ratings on scales and total score by form. The results reveal similar scores—all scale means are within 0.04 except one core component and two key processes. Mean scores on Form A are 4.31, and mean scores on Form C are 4.33. Although these are not definitive data because there were only five schools for Form A and four schools for Form C, they suggest that teacher ratings on the two forms were roughly equal. Again, sample sizes for supervisor and principal were too small to justify making the comparison.

Other results also suggest parallel forms. Missing and *don't know* data show that the four scales most likely to be marked *don't know* were the same on the two forms. Although Form C had slightly higher rates of *don't know* responses, this could be because of the fact that a much larger percentage of Form C respondents were from middle or high schools than for Form A. Internal consistency estimates in Table 6 are similar across forms. Evidence sources in Table 3 are similar across forms, with the mean number of sources of evidence used differing by just 0.01 between forms. Although none of the data reported here affirm that the forms are parallel, neither do they suggest otherwise. Given that schools were randomly assigned to forms and that there were only nine schools, the data could hardly be more supportive.

The overall message from the nine-school pilot study was straightforward. VAL-ED's items were clear to respondents, and respondents were willing to

complete them. Internal consistency reliability was excellent for each scale and total. Most importantly, the great majority of respondents from all three respondent groups agreed that VAL-ED measures key leadership behaviors.

Changes to the Instrument

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

A second change made was to relabel the Levels 3 and 5 ratings. The goal was to stretch the top end of the distribution, so *highly effective* was moved to Level 4, and Level 5 was renamed *outstandingly effective*. Level 3 was renamed from *moderately effective* to *satisfactorily effective*, and Level 2 was named *minimally effective*. To further emphasize the exceptional principal behaviors to be rated outstandingly effective, a sentence describing outstandingly effective behaviors was added on the directions page. A sentence describing ineffective behaviors was also added (these descriptions can be seen in Figure 1).

A third change was providing labels for all five of the effectiveness ratings rather than just three in the original model. Although respondents in the cognitive interviews showed that they generally understood the meanings of the unlabeled effectiveness ratings, labels were added to Levels 2 and 4. Labels were included because it was thought that in conjunction with relabeling the Levels 3 and 5 ratings, this change would result in increased spread of ratings. The set of changes to the rating scale described here was designed to have the effect of stretching the distribution to create more between-school variance in ratings of principal effectiveness.

Fourth, the number of items was reduced from 108 to 72. Respondents to the pilot study overwhelmingly indicated that the form was too long.

Additionally, our contacts in districts and states suggested that VAL-ED would be more useful to schools and districts if it took less than 30 min to complete. Randomly deleting an item from each cell and rearranging the data resulted in reliabilities for both forms and all scales largely unchanged. Scale reliabilities for teacher scores were all above or near .9, and total score reliabilities were still near perfect on these shortened forms.

The only exception to the random removal of items was the outlier item described earlier. The item, which read “The principal ensures the school uses data on parent involvement in teacher evaluations,” was not randomly selected for removal after the 36 items were removed as described. However, the item had a mean teacher rating of 3.56, roughly 0.4 lower than any other item on either form. Also, without reviewing item ratings, the item was identified by a former school principal and superintendent as a problematic item. The item was replaced with an item randomly selected from the remaining items in the pool of items for the cell (connections to external communities and monitoring).

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

Eleven-School Pilot Test

After the substantial changes made to the instrument resulting from the 9-school pilot, a second pilot study was conducted to examine the effects of the revisions. The methods for the study were identical to the 9-school pilot, except that 11 schools in four districts in a second midwestern state participated in the study. The forms used in the study were the updated 72-item forms with the modified stem and response categories. A sample set of items from the final forms is shown in Figure 7. The primary concerns for this pilot were the distribution of teacher, principal, and supervisor effectiveness ratings. We focus on results that bear on the changes made after the nine-school pilot.

Results

Results support that the changes made had the desired effects. Mean teacher responses for the 11-school pilot were 3.29 for total score, ranging from a low of 3.10 (connections to external communities) to a high of 3.37 (culture of learning and professional behavior). Comparing these results to the results

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. develops a plan for high standards of student performance that are measurable.	<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 for rigorous academic and social learning goals.	<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 7. Sample items from final form

in Table 9 for the 9-school pilot, we see that teacher scores were more than a full point lower after the revisions. Principal (3.72) and supervisor (3.77) total scores were also lower in the 11-school pilot than in the 9-school pilot. Furthermore, results were more spread. In the 9-school pilot, school-level teacher means ranged from a low of 3.93 to a high of 4.61, a spread of less than 0.7 points on the 5-point effectiveness scale. In the 11-school pilot, school-level teacher means were as low as 2.81 and as high as 3.90, a spread of more than a full point. Although it is possible that the principals in the 11-school pilot were less effective and more variable in quality than those in the 9-school pilot, it is possible that these results suggest the rescaling was effective in lowering and increasing the spread of effectiveness ratings.

A second change to arise from the 9-school pilot was a reduction from 108 items to 72 items. This change was made because 81 of approximately 350 respondents expressed concerns about length in comments at the end of the forms. Furthermore, respondents in cognitive interviews had argued that the form was too long at 108 items. In the 11-school pilot, the reduction in items had little effect on reliability; principal and supervisor scale and total score reliabilities remained above .89, and teacher scale and total score reliabilities remained above .94. Also, the number of respondents commenting on the length of the instrument decreased to 30 out of more than 500, suggesting that length was less of a concern.

A scatter plot of teacher and principal mean effectiveness ratings by school for the 11-school pilot is provided in Figure 8. The correlation was .79. For the individual scales, correlations ranged from .68 to .88. The correlation between

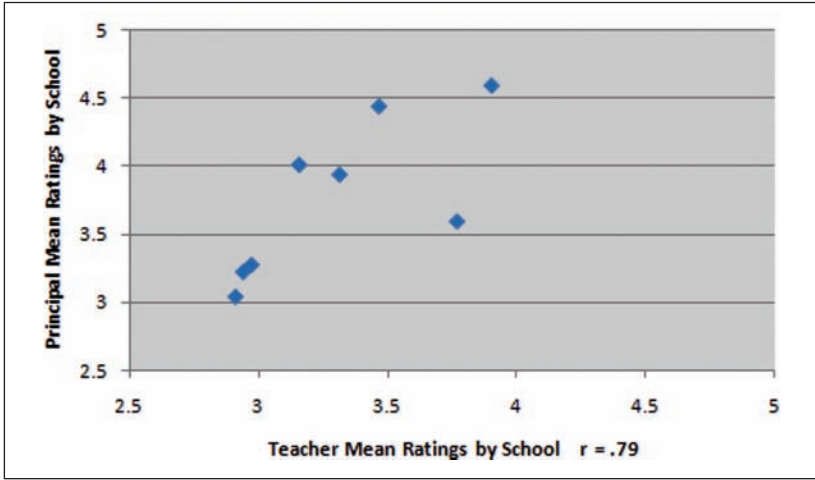


Figure 8. Scatter plot of principal ratings with mean teacher ratings for 11-school pilot

teacher and supervisor on total score was .68, and the correlation between principal and supervisor was .51. These surprisingly large correlations may be a function of the small sample size ($n = 11$).

Overall, the results of the 11-school pilot suggested that the changes made after the 9-school pilot were successful. Effectiveness ratings were lower and more variable with higher levels of agreement, and there was far less concern about the assessment's length.

Summary and Conclusions

The VAL-ED measures principals' leadership behaviors on six core components and six key processes. For each cell in the 36-cell conceptual framework, two items are included on each of the two parallel forms of the assessment available in paper and pencil and online. Respondents are the principal, the principal's supervisor, and all the teachers in the principal's school. Respondents rate the principal on a scale from 1 = *ineffective* to 5 = *outstandingly effective* on items asking, "How effective is the principal at ensuring the school"

The instrument was developed following the guidance of the Standards for Educational and Psychological Testing (AERA, American Psychological Association, & NCME, 1999). The first and most important step in establishing

validity was the item and test development phase. An iterative item-writing process was undertaken, with team members writing and revising items to fit in the conceptual framework. Care was taken to ensure consistent language, with a set of verbs chosen for each core component. Redundant items within core components or key processes were removed, and items were rewritten to ensure appropriate grain size. After an exhaustive item-writing process, 294 items remained. To examine the content validity of the items, a sorting study was conducted with three respondents for each third of the total item pool. At the marginals, all core components and key processes except implementing and high standards for student learning had greater-than-70% sorting accuracy. After the sorting study, several modifications to items were made to improve item fit.

Next, two rounds of cognitive interviews were conducted. Seventeen respondents were recruited from elementary, middle, and high schools in five urban or suburban districts in five states. Respondents, including teachers, principals, and supervisors, generally understood the directions and the items. The response scale was clear and easily understood. Respondents raised issues about certain items and phrases in the directions that were addressed in subsequent revisions. The interviews indicated ways in which the instrument was revised.

Next, a nine-school pilot study was conducted in an urban district. Elementary, middle, and high schools participated. In terms of feasibility, reliability, and validity, the pilot study provided positive evidence. Response rates were high and missing item and *don't know* rates were low. Confirmatory factor analysis revealed excellent goodness of fit, despite high intercorrelations among core components and key processes. The pilot identified the need for several revisions, the two most important of which were to change the effectiveness scale and to dramatically shorten the instrument.

To examine the changes made in light of the 9-school pilot, an 11-school pilot was conducted. Results indicated that as intended, teacher mean effectiveness ratings were both more variable and lower than the ratings given in the 9-school pilot. In some schools, teachers rated their principals as less than satisfactorily effective, whereas in others, teachers rated their principals as nearly highly effective. Furthermore, there was less concern expressed about the instrument's length after the removal of one third of the items, and reliability of scales and total score remained high for all respondent groups.

Throughout the nearly 3 years of development, the focus was on creating a reliable and valid measure of principal leadership for widescale use. Clearly, the iterative process of revise, test, revise, and test was useful in the sense that each test revealed the need for revisions while at the same time indicating the utility of prior revisions. Although there exists from the development

process more psychometric evidence than is available on any other assessment of instructional leadership, important work lies ahead.

Thus far, all of the psychometric evidence is based on data from research studies of the instrument. What is needed next is to see how the VAL-ED works in real use. A set of six studies has been planned using data from real use. A test-retest reliability study will examine the stability of VAL-ED ratings based on two proximal administrations of the instrument. A convergent-divergent validity study will compare VAL-ED ratings to ratings based on an instrument purporting to measure similar constructs and one measuring different constructs. A known groups study will address whether the VAL-ED can correctly identify principals classified as more and less effective on the basis of supervisor nominations. An evidence study will ask whether VAL-ED ratings are affected by the format with which respondents are required to provide evidence. A consequences study will use mixed methods to investigate the short- and long-term consequences of VAL-ED use on principals and schools. Finally, a longitudinal correlational study will investigate the relationship between principals' effectiveness on the VAL-ED and value added to student achievement. The most challenging study is the longitudinal investigation of the relationship between performance on the VAL-ED and value added to student achievement. We hypothesize only indirect effects, with leadership leading to better teaching leading to increase in student achievement. Throughout these additional validity studies, we will continue to investigate the psychometric properties of the instrument as we did in the 9- and 11-school pilots, including analysis of respondent group correlations, scale reliability, and factor structure, using larger samples and real-user data.

The VAL-ED was built to (a) work well in a variety of settings and circumstances, (b) be construct valid, (c) be reliable, (d) be unbiased, (e) provide accurate and useful reporting of results, (f) yield diagnostic profiles for formative purposes, (g) be used to measure progress over time in the development of leadership, and (h) predict important outcomes. The development process addressed only (a) through (d) and (f) and, even then, only on VAL-ED as a research instrument, not in actual use. Further work lies ahead; it remains to be seen how well the VAL-ED will ultimately hold up to the exacting standards to which it is being subjected.

Declaration of Conflicting Interests

The authors declared a potential conflict of interest (e.g., a financial relationship with the commercial organizations or products discussed in this article) as follows: 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.

Funding

The authors disclosed receipt of the following financial support for the research and/or authorship of this article: The authors gratefully acknowledge the generous support of the Wallace Foundation.

Note

1. The Institute of Education Sciences of the U.S. Department of Education, through Grant R305C050041-05 to the University of Pennsylvania, is supporting this further work.

References

- American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). *Standards for educational and psychological tests*. 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.
- Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., & Sudman, S. (1991). *Measurement errors in surveys*. New York: Wiley.
- Burns, J. M. (1978). *Leadership*. New York: Harper and Row.
- Conley, D. T., & Goldman, P. (1990). Ten propositions for facilitative leadership. In J. Murphy & K. S. Louis (Eds.), *Reshaping the principalship: Insights from transformational reform efforts* (pp. 237-265). Thousand Oaks, CA: Corwin.
- 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.
- Ginsberg, R., & Berry, B. (1990). The folklore of principal evaluation. *Journal of Personnel Evaluation in Education, 3*, 205-230.
- 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.

- Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*(1), 43-62.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.
- Leithwood, K. (1994). Leadership for school restructuring. *Educational Administration Quarterly, 30*, 498-518.
- Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning*. Minneapolis: University of Minnesota.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448.
- Marks, H., & Printy, S. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational Administration Quarterly, 39*, 370-397.
- Marzano, R. J., Waters, T., & McNulty, B. A. (2005). *School leadership that works: From leadership to results*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 12-103). New York: Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(1), 13-23.
- Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430-445.
- Murphy, J., Elliott, S. N., Goldring, E., & Porter, A. C. (2006). *Learning-centered leadership: A conceptual foundation*. New York: Wallace Foundation.
- Murphy, J., Elliott, S. N., Goldring, E., & Porter, A. C. (2007). Leadership for learning: A research-based model and taxonomy of behaviors. *School Leadership & Management, 27*(2), 179-201.
- Porter, A. C., Goldring, E., Elliott, S. N., Murphy, J., Polikoff, M. S., & Cravens, X. C. (2008). Setting performance standards for the VAL-ED assessment of principal leadership. (ERIC Document No. ED505799).
- Sebring, P., & Bryk, A. (2000). School leadership and the bottom line in Chicago. *Phi Delta Kappan, 81*, 440-443.

Bios

Andrew C. Porter is dean of the Graduate School of Education, University of Pennsylvania. He is an applied statistician and psychometrician who studies the measurement of education leadership, student achievement testing, curriculum policies and their effects, and professional development for teachers.

Morgan S. Polikoff is an advanced doctoral student studying Education Policy at the University of Pennsylvania's Graduate School of Education. His research focuses on state and federal policies and their impact on school processes such as instruction and leadership.

Ellen Goldring is Patricia and Rodes Hart Chair and Professor of Education Policy and Leadership. She is Chair of the Department of Leadership, Policy and Organizations at Peabody College. Professor Goldring's research focuses on improving schools with particular attention to educational leadership, school choice, and parent involvement.

Joseph Murphy is the Frank W. Mayborn Chair of Education and Associate Dean at Peabody College of Education of Vanderbilt University.

Stephen N. Elliott received his doctorate at Arizona State University in 1980 and is a Professor of Special Education and the Dunn Family Chair of Educational and Psychological Assessment in Peabody College at Vanderbilt University.

Henry May is a Senior Research Investigator at the Consortium for Policy Research in Education (CPRE) at the University of Pennsylvania.