

MORGAN S. POLIKOFF
HENRY MAY
ANDREW C. PORTER
STEPHEN N. ELLIOTT
ELLEN GOLDRING
JOSEPH MURPHY

An Examination of Differential Item Functioning on the Vanderbilt Assessment of Leadership in Education

ABSTRACT: The Vanderbilt Assessment of Leadership in Education is a 360-degree assessment of the effectiveness of principals' learning-centered leadership behaviors. In this report, we present results from a differential item functioning (DIF) study of the assessment. Using data from a national field trial, we searched for evidence of DIF on school level, geographic region, and urbanicity. We found evidence of intercept DIF for urbanicity on four items and slope DIF for urbanicity on one item, although all magnitudes were small to moderate. We conclude that the Vanderbilt Assessment of Leadership in Education's items are not biased on the basis of these school characteristics, bolstering its use in schools around the country.

Principal leadership is widely believed to be an important factor in effecting school improvement. Researchers studying the characteristics of effective schools that raise achievement and narrow achievement gaps have found that principal leadership is among the most important factors (Beck & Murphy, 1996; Leithwood, Louis, Anderson, & Wahlstrom, 2004; Marks & Printy, 2003; Marzano, Waters, & McNulty, 2005; Murphy & Hallinger, 1988; Robinson, Lloyd, & Rowe, 2008; Waters, Marzano, & McNulty, 2003). Although there has been important work in research and policy to improve the quality of principal leadership nationwide, the focus of this work has

Address correspondence to Morgan S. Polikoff, University of Southern California, Rossier School of Education, Waite Phillips Hall, Los Angeles, CA 90089. E-mail: morgan.polikoff@gmail.com.

been imbalanced. Much of the work to date has focused on the leverage points of standards—through the Interstate School Leaders Licensure Consortium and Educational Leadership Constituents Council standards, through professional development tied to the standards, through accreditation via the National Council for Accreditation of Teacher Education, and through licensure and certification via the School Leaders Licensure Assessment (Murphy, Elliott, Goldring, & Porter, 2007). However, little work has been done with regard to a critical fifth leverage point—leadership evaluation and consequences.

Many schools and districts use some form of principal evaluation tool (Goldring, Cravens, et al., 2009). However, research has highlighted that these tools often lack conceptual (Ginsberg & Berry, 1990) and empirical grounding (Goldring, Cravens, et al., 2009). To address this important gap, a team of researchers at Vanderbilt University and the University of Pennsylvania set out in 2005 to construct a theoretically grounded, reliable, and valid assessment of principal instructional leadership. The first phase of work involved synthesizing the literature on effective principal leadership, constructing a conceptual framework based on the literature, and drafting an instrument (Goldring, Porter, Murphy, Elliott, & Cravens, 2009; Murphy et al., 2007; Porter, Goldring, Murphy, Elliott, & Cravens, 2006). Next, a series of small-scale psychometric studies was conducted to provide initial validity and reliability evidence and to make improvements to the instrument (Porter et al., 2010). Finally, a national field trial was conducted in elementary, middle, and high schools across the country, with the results lending important support to the conceptual framework (Porter et al., 2009). Additional validation work is ongoing through a grant from the Institute of Education Sciences. The instrument, called the Vanderbilt Assessment of Leadership in Education (VAL-ED), is now available for use by schools and districts around the country as a means to measure the effectiveness of principal instructional leadership and to provide growth targets for professional development activities.¹

A critical component of the analysis of field trial data was the consideration of item bias. One definition of item bias involves responses on an item that are systematically inconsistent with the overall scores of the respondents in subgroups (Camilli, 2006). Whereas a previous qualitative study had been conducted to remove potentially problematic items from the instrument (Porter et al., 2010), it was important to test whether the qualitative examination had removed all biased items. To that end, we conducted an item bias study, described here. The research question guiding the study was as follows: To what extent is there bias in the VAL-ED items based on grade level, urbanicity, or geographic region?² The variables chosen for differential item functioning (DIF) analysis were selected because they were the variables on which the sample was stratified for selection

into the national field trial. The purpose of this analysis was to detect evidence that the VAL-ED items—as written by leadership experts and subject to numerous qualitative and quantitative reviews—were operating equally for principals in all types of U.S. public schools.

To address the research question, we use an item response theory framework by employing DIF analyses (Holland & Wainer, 1993). We supplement these analyses with examinations of the item descriptive statistics to provide interpretations of the degree and nature of item bias. We conclude with a discussion of the ways in which the results are used to guide further test development and item refinement.³

BACKGROUND

IMPROVING PRINCIPAL EFFECTIVENESS

For at least several decades, improving the effectiveness of principal leadership has been a focus of education research. Although principal evaluation has been seen as an important lever for bringing about these improvements, evaluation has been seen as a challenge because of the many perspectives on the leadership behaviors that are important to assess (see, for instance, Glasman & Heck, 1992; Hart, 1992; Hoyle, English, & Steffy, 1985; Oyinlade, 2006). Despite these challenges, assessment of leaders has become widespread in schools and districts (Goldring, Cravens, et al., 2009). However, research on the instruments currently in use in states and urban districts suggests that the instrumentation used for principal evaluation varies widely in terms of content and methodology (Goldring, Cravens, et al., 2009). For instance, the number of items on a selection of principal leadership evaluation tools created and used by states and urban districts varies from fewer than 10 to more than 180. The proportion of the assessments focused on school and instruction varies from less than 30% to more than 75%. Furthermore, all but 2 of the 66 instruments analyzed have no documented psychometric properties. In short, the instruments used by states and urban districts to evaluate leadership effectiveness are largely unconnected to the research on effective leadership and unsupported by validity and reliability evidence.

Researchers have created a number of other principal leadership assessments for use in research or practice. Among these are the Principal Instructional Management Rating Scale (Hallinger, 1983; Hallinger & Murphy, 1985), McREL's Balanced Leadership Profile (Marzano et al., 2005; Waters et al., 2003), and Heck and colleagues' survey instrument (Heck & Marcoulides, 1996). The conceptual models underlying these instruments, like that underlying the VAL-ED, assume that principal leadership has indirect effects on student outcomes (see Figure 1 for our conceptual model). That is, principals

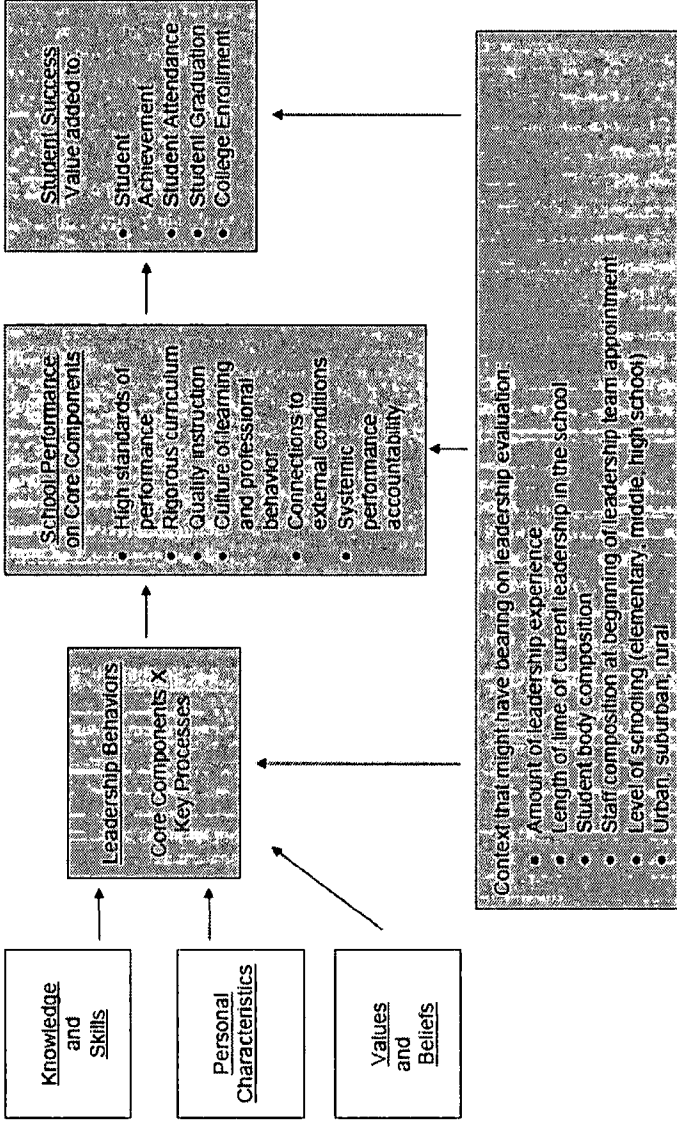


Figure 1. Vanderbilt Assessment of Leadership in Education conceptual model.

influence the effectiveness of school operations, such as the quality of teaching and learning, which affects student learning.

An important frame for thinking about how to improve the effectiveness of principal leadership is an instruction-focused frame (Knapp, Copland, & Talbert, 2003; Murphy, 1990). Instructional leadership is leadership that is intensely focused on teaching, learning, curriculum, and assessment and that makes all other features of the school work in support of the school's instructional mission (Murphy et al., 2007). Specifically, our review of the literature found that effective instructional leaders focus their effort on six core components of schooling (for detail, including definitions, see Goldring, Porter, et al., 2009): high standards for student learning (Brookover & Lezotte, 1977; Hallinger & Heck, 2002; Murphy et al., 2007; Purkey & Smith, 1983), rigorous curriculum (Marzano et al., 2005; Newmann, 1997; Russell, Mazzarella, White, & Maurer, 1985), quality instruction (Clark, Lotto, & McCarthy, 1980; Conley, 1991; Leithwood & Jantzi, 1990), culture of learning and professional behavior (Clark et al., 1980; Louis, Marks, & Kruse, 1996), connections to external communities (Corcoran & Wilson, 1985; Goldring & Hausman, 2001; Goldring & Sullivan, 1996; Russell et al., 1985), and performance accountability (Adams & Kirst, 1999; Bryk & Schneider, 2002). The six core components are enacted by effective leaders through six key processes: planning, implementing, supporting, advocating, communicating, and monitoring (Goldring, Porter, et al., 2009). See Figure 2 for our framework. In our conceptual framework, as supported by the research on effective principal leadership and closely tied to the Interstate School Leaders Licensure Consortium standards, high-quality instructional leadership takes place at the intersection of core components and key processes.

The conceptual framework guiding development of the VAL-ED contains many elements found in existing leadership assessments. For instance, McREL's Balanced Leadership Profile contains 21 leadership responsibilities and associated behaviors, including culture; curriculum, instruction, and assessment; outreach; and monitoring/evaluating (Marzano et al., 2005). The majority of the other leadership responsibilities in that framework fit in our conceptual framework; however, that framework includes principal knowledge and attitudes, which we believe are precursors to effective principal behaviors. Hallinger's Principal Instructional Management Rating Scale (1983; Hallinger & Murphy, 1985) has strands for framing and communicating school goals, supervising and evaluating instruction, coordinating the curriculum, monitoring student progress, protecting instructional time, and others. Again, these behaviors are included in our framework, but the VAL-ED focuses on the effectiveness with which the behaviors are implemented, rather than on their frequency. In short, the framework that

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

Figure 2.

guided the development of the VAL-ED has been extensively studied, and it is closely connected to professional standards and theories of effective leadership embodied in existing instrumentation. However, our framework organizes the key leadership behaviors in a way that emphasizes core features of instructionally effective schools and the ways in which principal behaviors can bring about school improvement.

THE VAL-ED

The VAL-ED is an evidence-based multirater rating scale that assesses principals' learning-centered leadership behaviors known to directly influence teachers' performance and, in turn, students' learning. The VAL-ED measures critical learning-centered leadership behaviors for the purposes of diagnostic analyses, performance feedback, progress monitoring, and professional development planning.

With support from the Wallace Foundation, work began in 2005 to develop a 360-degree instrument to assess the effectiveness of school leaders as evaluated by teachers, supervisors, and principals themselves. The resulting assessment is available in paper form or online, and it utilizes a multirater evidence-based approach to measure the effectiveness of leadership behaviors. There are two parallel forms of the instrument (Forms A and C) to facilitate repeated assessments over time. The VAL-ED measures core components and key processes, as described previously. Core components refer to characteristics of schools that support the learning of students and enhance the ability of teachers to teach. Key processes refer to how leaders create and manage those core components. Effective learning-centered leadership is at the intersection of the two dimensions: core components created through key processes. The outcomes of the assessment are profiles, interpretable from both norm-referenced and standards-referenced perspectives, and suggested clusters of behaviors for improvement. The vision for the VAL-ED is a larger assessment system, with the VAL-ED measuring not only leadership effectiveness but also direct measures of the core components of effective schools. Such a system would allow for detailed investigation of the relationships among leadership effectiveness, school effectiveness on the core components, and value-added measures of student learning.

The first step in creating the VAL-ED was item and instrument development based on the conceptual framework. Once an item pool was finalized, items were randomly assigned to Form A and Form C by cell to ensure form equivalence. Following this, we conducted a series of quantitative and qualitative studies designed to make iterative improvements to the instrument and to marshal validity and reliability evidence. These studies included a sorting study, cognitive interviews, a qualitative bias study,

and two small-scale pilot tests (for information on the early development of the VAL-ED, see Porter et al., 2010).

Following the development phase, we conducted a national field trial of the VAL-ED in the spring of 2008. The purposes of the field trial were several. Among them were further investigation of the instrument's psychometric properties, establishment of norms and performance standards, and investigation of possible differences in VAL-ED performance by design effects. The intended sample was all teachers, the principal, and the principal's supervisors in 300 schools nationwide. The sample was to include 100 elementary schools, 100 middle schools, and 100 high schools. By region, we sought schools in the Midwest, West, Northeast, and South: 75 schools per region. We also sought 150 urban, 100 suburban, and 50 rural schools. Of the 150 urban schools, 50 were to be drawn from Wallace Foundation grantee districts, 50 from districts in Wallace grantee states, and 50 from non-Wallace-affiliated districts and states. The final sample included responses from 276 schools, with 218 complete sets of responses. For more information on the national field trial, see the *VAL-ED Technical Manual* (Porter et al., 2009).

After the national field trial, a series of psychometric studies using the field trial data established that, when used as designed, the VAL-ED works well in a variety of settings and circumstances; is unbiased, construct valid, reliable, and feasible for widespread use (both the online and paper-and-pencil versions); provides accurate and useful reporting of results; and yields a diagnostic profile for formative purposes (Porter et al., 2009; Porter et al., 2010).

Subsequent to the national field trial, we began work on a second grant to further investigate the psychometric properties of the instrument. This grant, currently in its 2nd year, consists of several reliability and validity studies, including a test-retest reliability study, a convergent/divergent validity study, an evidence study to examine the utility of the VAL-ED's sources of evidence, a consequences study, a known-groups study, and a longitudinal study of the relationship between VAL-ED scores and value added to student achievement. Thus, although it is important to demonstrate that the VAL-ED is free from substantial bias, there is validation work to be done and it is under way. When completed, this work will provide important evidence about the relationship of principal effectiveness (as measured by the VAL-ED) and student learning outcomes.

ITEM BIAS

There has been little work in the literature on item bias on assessments of principal leadership. The relevant results are summarized as follows. First, we summarize results from a previous study of item bias on the VAL-ED (Porter et al., 2010).

Item bias on the VAL-ED. During the test development phase, we conducted a qualitative fairness review of the VAL-ED instructions and items against the test fairness guidelines published and used by the Educational Testing Service (2000). Nine respondents with knowledge of testing methods, including eight school personnel, participated in the study. Panel members independently reviewed the two forms of the VAL-ED and indicated potential violations of the fairness guidelines. The results indicate four items across the two forms that were deemed problematic by three or more panelists. Three of the four changed items concerned the key process of advocating. Two of the language changes involved emphasizing person-first language. The panel discussed and modified these four items, and the resulting modifications were included in updated versions of the VAL-ED. With the completion of the item bias study, the items and instructions for the VAL-ED were believed to meet the Educational Testing Service fairness criteria for tests.

Item bias on other leadership assessments. There is little in the literature with regard to studies of item bias on other assessments of leadership. A few studies provided DIF analyses of 360-degree leadership assessments in fields other than education (e.g., Fecteau & Craig, 2001; Penny, 2003). These are generally studies of rater effects rather than other contextual or background variables. In a study of a multisource appraisal form (Fecteau & Craig, 2001), roughly 10% of items showed DIF on rater group, but all magnitudes were quite small. In a study of the Benchmarks instrument (Penny, 2003), the DIF focused on just three items and was deemed small in magnitude. Both studies concluded that rater effects are not a serious problem for the respective instruments. Several studies of Hallinger's Principal Instructional Management Rating Scale (1983) involved the comparison of ratings based on background variables of the principal or school (e.g., Collins, 1998; Delano, 1985; Gallon, 1998). However, these were not studies of item bias but, rather, differences between groups. In short, there is no good evidence from the literature to inform the discussion of item bias on assessments of principal leadership.

METHOD

Item response data from the VAL-ED national field trial including principals, teachers, and supervisors (usually, the superintendent) from 218 schools were subjected to DIF analyses using PARSCALE 4.1 (Muraki & Bock, 1996). Samejima's (1969) graded-response item response theory model with a logit link function was used to model the cumulative probability of responses to the 72 VAL-ED survey items employing a 5-point Likert-type scale of leadership effectiveness. This model leverages the ordinal nature of the item responses by extending the traditional two-parameter

logit item response theory model to include a set of four ordinal category parameters that separate the five response categories. In the analyses presented here, each item had distinct intercept (i.e., difficulty) and slope (i.e., discrimination) parameters, but the category parameters were constrained to be the same for all items. Thus, although it may have been easier or harder to earn high ratings on various items and although the relative contribution of items to the total score varied, the spacing between response categories was assumed to be consistent from one item to the next.

The DIF analyses assessed item bias by allowing the difficulty and discrimination parameters for each item to vary across specific subgroups. In the analyses presented here, we tested for differences between subgroups defined by grade level (elementary, middle, high), urbanicity (rural, suburban, urban), and region (Midwest, Northeast, South, West). DIF was identified with a chi-square test of the standardized difference between the item parameters of two groups. Because of the large number of comparisons, we expected a sizable number of items to be identified as displaying DIF by chance; thus, we set a slightly more conservative critical value of .01 for all analyses. The reference categories for these analyses were elementary, urban, and Northeast.

Because the primary unit of analysis and interpretation of scores from the VAL-ED is the individual principal, sample weights were used in PARSCALE to adjust for the multiple survey responses collected for each principal. The weighting scheme for the DIF analyses mirrored the weighting scheme used for scoring the VAL-ED as used in the field. Each principal's and each supervisor's responses were weighted one third, and the remaining one-third weight for each school was divided evenly among the teachers responding from that school. By setting the sum of the weights to equal the number of principals in the analysis sample, the effective sample size used for the calculation of standard errors is equal to the sample size for the primary unit of analysis—the principal. The effective sample sizes for the DIF analyses were $n = 103$ and $n = 115$ for Forms A and C, respectively. These effective sample sizes yield standard errors between .20 and .40 for item statistic contrasts, suggesting that the power to detect item bias is sufficient (i.e., greater than 80%) for medium to large effect sizes ($\delta = .50$ to $.80$).

The primary disadvantage of DIF analyses is that it is somewhat difficult to interpret differences in item parameters when DIF is present. This is especially true regarding slope DIF, given that the slope parameter in an item response theory model has no easily interpretable units. To aid in interpretation of results, we conducted additional analyses using classical item statistics to enhance the interpretation of the DIF results. For those items showing intercept bias, we present the average item score for the different

groups, along with the average total score for the different groups on the original 5-point Likert-type scale. By comparing the difference in item scores to the difference in total scores, we can benchmark the position and size of the item bias against the original response metric. Likewise, for those items showing slope bias, we present the item–total correlation for the different groups. By comparing *item–total correlations*, we can evaluate the relative size of the bias against the familiar correlation metric instead of the logistic slope discrimination metric from the item response theory model.

RESULTS

Analyses for region and school level revealed no significant slope or intercept DIF on any items. Table 1 shows results of DIF analyses by urbanicity. Four items showed statistically significant DIF: Items 7 and 49 for the urban–suburban comparison and Items 38 and 40 for the urban–rural comparison. (To see the core-component–key-process combination for each item, refer to Table 2; to see the text for each item, refer to Table 3.) The DIF is also evident by examining the item means for the various groups. For instance, the suburban principal mean for Item 7 was 0.27 points higher than the urban principal mean, as compared to a total score difference of just 0.05. In contrast, the suburban principal mean for Item 49 was 0.31 points lower than the urban principal mean for that item. Similarly, for the urban–rural comparison, one item was significantly easier for the urban principals (Item 38), and one was significantly more difficult for urban principals (Item 40). All subgroup differences were between 0.16 and 0.33 on the VAL-ED 5-point effectiveness scale. Given an average item standard deviation of 0.72 for Form C, these subgroup differences translate into standardized effect sizes between 0.22 and 0.46 standard deviations.

Table 1. Intercept Differential Item Functioning for Urbanicity: Form C

Item	Comparison Group	DIF Parameter	p	M		
				Urban	Suburban	Rural
7	Suburban	–2.717	.003	3.54	3.81	
49	Suburban	4.293	< .001	3.61	3.30	
38	Rural	2.708	.003	3.75		3.42
40	Rural	–2.709	.003	3.91		4.07
Total				3.63	3.68	

Note. DIF = differential item functioning.

Table 2. Vanderbilt Assessment of Leadership in Education: Item Numbers

<i>Core Components</i>	<i>Key Processes</i>					
	<i>Planning</i>	<i>Implementing</i>	<i>Supporting</i>	<i>Advocating</i>	<i>Communicating</i>	<i>Monitoring</i>
High standards	1, 2	3, 4	5, 6	7, 8	9, 10	11, 12
Rigorous curriculum	13, 14	15, 16	17, 18	19, 20	21, 22	23, 24
Quality instruction	25, 26	27, 28	29, 30	31, 32	33, 34	35, 36
Culture of learning	37, 38	39, 40	41, 42	43, 44	45, 46	47, 48
Connections to external communities	49, 50	51, 52	53, 54	55, 56	57, 58	59, 60
Performance accountability	61, 62	63, 64	65, 66	67, 68	69, 70	71, 72

To help interpret these DIF results, we can examine the text of the items, as displayed in Table 3. For the urban–suburban comparison, the two items exhibiting DIF were Item 7, which reads, “How effective is the principal at ensuring the school advocates for students with special needs when making decisions about high standards for student learning?” and Item 49, which reads, “How effective is the principal at ensuring the school plans for the use of external community resources to promote academic and social learning goals?” For Item 7, the item mean was higher than that expected for suburban principals, indicating that urban principals are perceived as not focusing as directly on high standards for students with special needs. For Item 49, the item mean was higher than that expected for urban principals, indicating that they are perceived as planning for external partnerships more effectively than are suburban principals.

For the urban–rural comparison, the two items exhibiting DIF were Item 38, which reads, “How effective is the principal at ensuring the school plans for a culture of shared responsibility for the social and academic learning of students?” and Item 40, which reads, “How effective is the principal at ensuring the school builds a school environment that is safe and orderly for all students?” In the case of Item 38, the results reveal that urban principals are perceived as being more effective at developing a school culture of responsibility than are rural principals. In the case of Item 40, the results suggest that urban principals are seen as being less effective in terms of ensuring school safety than are rural principals. All these differences, urban–rural and urban–suburban, are small to moderate in size.

Finally, one item exhibited slope DIF for urbanicity, as seen in Table 4. Item 26, which reads, “How effective is the principal at ensuring the school plans opportunities for teachers to improve their instruction through professional development?” had a higher item–total correlation for urban principals (.76) than for rural principals (.64). Thus, this item appears to provide less information about overall principal effectiveness for rural principals than for suburban principals.

Table 3. Text of Items Exhibiting Differential Item Functioning: Form C

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

Table 4. Slope Differential Item Functioning for Urbanicity: Form C

Item	Comparison Group	DIF Parameter	p	Item-Total Correlation		
				Urban	Suburban	Rural
26	Rural	-2.629	.004	.76		.64
Average				.67	.69	.66

Note. DIF = differential item functioning.

DISCUSSION

The purpose of this analysis was to investigate data from the VAL-ED field trial for evidence of item bias. An item response theory DIF model was used. Overall, little item bias was found in the data. No evidence of DIF was found for school region or level. Five items were found to exhibit DIF on urbanicity for Form C. However, the degree of DIF was small to moderate when item means or item total correlations were compared—less than 0.50 standard deviations for all instances of intercept DIF and a 0.10 difference in correlation coefficients for the one instance of slope DIF.

We will replicate this analysis on real-user data when a robust and representative national sample accumulates to inform inspection of potential future DIF. Even if DIF is found for region, urbanicity, or school level, it is not necessarily indicative that an item should be removed for bias. For instance, elementary school principals are expected to score better than high school principals on items assessing school–parent relations, based on the size of elementary schools and the relative lack of independence of elementary school students compared to high school students. Nevertheless, school–parent relations are important no matter the level of the school, so even if there were DIF on such an item, it would not necessarily be grounds for removing the item from the assessment. All cases of DIF will be reviewed by trained bias review panels to determine if the DIF merits removing or editing VAL-ED items.

There are a number of potential explanations for the few instances of DIF we identified. Certainly, there is the potential for real item bias. Certain leadership activities may be more or less common for principals at different types of schools. Another potential explanation has to do with the respondents who participated in the field trial. Although response rates were generally high, there did appear to be different response rates for certain groups. For instance, of the 218 schools with data from all three response groups, 39% were elementary schools, 32% were middle schools, and 28% were high schools. Twenty-three percent of the schools were from the West, 30% from the South, 22% from the Midwest, and 25% from the

Northeast. There were 39% urban schools, 39% suburban schools, and 22% rural schools. Thus, there was an apparent underrepresentation of urban schools in the obtained sample relative to the desired sample, as well as an overrepresentation of suburban, rural, and elementary schools. The low urban response rate was concentrated in three districts, which had 0 of 11, 5 of 13, and 4 of 16 schools responding. There was also differential response by form, with 115 of the 218 complete schools filling out Form C. It is certainly possible that the DIF we identified was due to characteristics of the sample that differed across groups rather than to true bias. Again, this argument will have to be evaluated with operational data from future VAL-ED districts.

In future analyses, it will be useful to consider the possibility of item bias on respondent characteristics (e.g., race, gender) in addition to the variables discussed here. Although this study provides additional evidence that the VAL-ED is fair in terms of being free from substantial bias (thereby supplementing the qualitative item bias study), there is additional verification to be done.

Finally, although the VAL-ED appears to be free from substantial bias, there is a great deal more work with the VAL-ED and other instruments to critically investigate how leadership affects school performance on the core components and, ultimately, student learning. We will do some of this work in our planned longitudinal study, where we will correlate VAL-ED scores with value-added measures of student achievement. Additional studies could use targeted professional development interventions to improve leaders' effectiveness and evaluate effects on core components and student learning. Thus, a reliable and valid VAL-ED, free from bias, can be a useful tool in expanding and improving the quality of research on leadership effects.

NOTES

1. The Vanderbilt Assessment of Leadership in Education is published by Discovery Education Assessment (Nashville, TN).
2. Although research suggests that there may be differences in leadership effectiveness based on leader characteristics (e.g., gender; Eagly, Karau, & Makhijani, 1995), we did not examine differential item functioning based on principal characteristics, because these variables were not collected in the national field trial.
3. Porter, Murphy, Goldring, and Elliott are coauthors of the Vanderbilt Assessment of Leadership in Education and benefit financially from its sale. Although they make every effort to be objective and data based in their statements about the instrument, one should judge the facts and related information materials for himself or herself and make independent decisions regarding use of the instrument.

REFERENCES

- Adams, J. E., & Kirst, M. W. (1999). New demands and concepts for educational accountability: Striving for results in an era of accountability. In J. Murphy & K. S. Louis (Eds.), *Handbook of research on educational administration* (pp. 463–489). San Francisco: Jossey-Bass.
- Beck, L. G., & Murphy, J. (1996). *The four imperatives of a successful school*. Newbury Park, CA: Corwin.
- Brookover, W. B., & Lezotte, L. W. (1977). *Changes in school characteristics coincident with changes in student achievement*. East Lansing: Michigan State University, College of Urban Development.
- Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York: Russell Sage.
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: ACE/Praeger.
- Clark, D. L., Lotto, L. L., & McCarthy, M. (1980). *Why do some urban schools succeed?* Bloomington, IN: Phi Delta Kappa.
- Collins, F. (1998). *An investigation of perceived differences in instructional leadership and school climate of African American and Caucasian female principals in Ohio's urban public schools*. Unpublished dissertation, Bowling Green University, Bowling Green, OH.
- Conley, D. T. (1991). Lessons from laboratories in school restructuring and site-based decision making. *Oregon School Study Council Bulletin*, 34(7), 1–61.
- Corcoran, T. B., & Wilson, B. L. (1985). *The secondary school recognition program: A first report on 202 high schools*. Philadelphia: Research on Better Schools.
- Delano, K. (1985). *An analysis of ratings by teachers of elementary school principals with different classroom teaching experiences*. Unpublished dissertation, Auburn University, Auburn, AL.
- Eagly, A. H., Karau, S. J., & Makhijani, M. G. (1995). Gender and the effectiveness of leaders: A meta-analysis. *Psychological Bulletin*, 117(1), 125–145.
- Educational Testing Service. (2000). *ETS standards for quality and fairness*. Princeton, NJ: Author.
- Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings obtained from different rating sources comparable? *Journal of Applied Psychology*, 86(2), 215–227.
- Gallon, S. (1998). *A study of teachers' perceptions of the principal as an instructional leader: A comparative analysis of elementary, middle and senior high school principals*. Unpublished dissertation, Florida International University, Miami.
- Ginsberg, R., & Berry, B. (1990). The folklore of principal evaluation. *Journal of Personnel Evaluation in Education*, 3, 205–230.
- Glasman, N. S., & Heck, R. H. (1992). The changing leadership role of the principal: Implications for principal assessment. *Peabody Journal of Education*, 68(1), 5–24.
- Goldring, E., Cravens, X. 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., & Hausman, C. (2001). Civic capacity and school principals: The missing link in community development. In R. Crowson & B. Boyd (Eds.), *Community development and school reform* (pp. 193–209). Greenwich, CT: JAI Press.
- 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.
- Goldring, E., & Sullivan, A. (1996). Beyond the boundaries: Principals, parents and communities shaping the school environment. In K. Leithwood, J. Chapman, P. Corson, & P. Hallinger (Eds.), *The international handbook of educational leadership and administration* (pp. 195–222). London: Kluwer.
- Hallinger, P. (1983). *Assessing the instructional management behavior of principals*. Unpublished dissertation, Stanford University, Stanford, CA.
- Hallinger, P., & Heck, R. H. (2002). What do you call people with visions? The role of vision, missions, and goals in school improvement. In K. Leithwood, P. Hallinger, G. Furman, J. MacBeath, B. Mulford, & K. Riley (Eds.), *The second international handbook of educational leadership and administration* (pp. 9–40). Dordrecht, Netherlands: Kluwer.
- Hallinger, P., & Murphy, J. (1985). Assessing the instructional management behavior of principals. *Elementary School Journal, 68*(2), 217–247.
- Hart, A. W. (1992). The social and organizational influence of principals: Evaluating principals in context. *Peabody Journal of Education, 68*(1), 37–57.
- Heck, R. H., & Marcoulides, G. A. (1996). The assessment of principal performance: A multilevel evaluation approach. *Journal of Personnel Evaluation in Education, 10*(1), 11–28.
- Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.
- Hoyle, J. R., English, F., & Steffy, B. (1985). *Skills for successful school leaders*. Arlington, VA: American Association of School Administrators.
- Knapp, M. S., Copland, M. A., & Talbert, J. F. (2003). *Leading for learning: Reflective tools for school and district leaders*. Seattle, WA: Center for the Study of Teaching and Policy.
- Leithwood, K., & Jantzi, D. (1990, June). *Transformational leadership: How principals can help reform school cultures*. Paper presented at the annual meeting of the Canadian Association for Curriculum Studies, Victoria, British Columbia, Canada.
- Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning*. Minneapolis: University of Minnesota.
- Louis, K. S., Marks, H., & Kruse, S. (1996). Teachers' professional community in restructuring schools. *American Educational Research Journal, 33*(4), 757–798.
- 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.

- Muraki, E., & Bock, R. D. (1996). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks*. Chicago: Scientific Software.
- Murphy, J. (1990). Principal instructional leadership. In L. L. Lotto & P. W. Thurston (Eds.), *Advances in educational administration: Changing perspectives on the school* (Vol. 1, pp. 163–200). Greenwich, CT: JAI Press.
- 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.
- Murphy, J., & Hallinger, P. (1988). The characteristics of instructionally effective school districts. *Journal of Educational Research, 81*(3), 176–181.
- Newmann, F. M. (1997). How secondary schools contribute to academic success. In K. Borman & B. Schneider (Eds.), *Youth experiences and development: Social influences and educational challenges* (pp. 88–108). Berkeley, CA: McCutchan.
- Oyinlade, A. O. (2006). A method of assessing leadership effectiveness: Introducing the essential behavioral leadership qualities approach. *Performance Improvement Quarterly, 19*(1), 25–40.
- Penny, J. A. (2003). Exploring differential item functioning in a 360-degree assessment: Rater source and method of delivery. *Organizational Research Methods, 6*(1), 61–79.
- Porter, A. C., Goldring, E., Murphy, J., Elliott, S. N., & Cravens, X. C. (2006). *A framework for the assessment of learning-centered leadership*. Nashville, TN: Vanderbilt University.
- Porter, A. C., Murphy, J., Goldring, E., Elliott, S. N., Polikoff, M. S., & May, H. (2009). *VAL-ED technical manual*. Nashville, TN: Vanderbilt University.
- Porter, A. C., Polikoff, M. S., Goldring, E., Murphy, J., Elliott, S. N., & May, H. (April 2010). Developing a psychometrically sound assessment of school leadership: The VAL-ED as a case study. *Educational Administration Quarterly, 46*(2), 135–173.
- Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. *The Elementary School Journal, 83*(4), 426–452.
- Robinson, V. M. J., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential impact of leadership types. *Educational Administration Quarterly, 44*(5), 635–674.
- Russell, J. S., Mazarella, J. A., White, T., & Maurer, S. (1985). *Linking the behaviors and activities of secondary school principals to school effectiveness: A focus on effective/ineffective behaviors*. Eugene, OR: Center for Educational Policy and Management.
- Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometric Monograph, No. 18*.
- Waters, T., Marzano, R. J., & McNulty, B. A. (2003). *Balanced leadership: What 30 years of research tells us about the effect of leadership on student achievement*. Aurora, CO: Mid-Continent Research for Education and Learning.

Morgan S. Polikoff received his doctorate from the University of Pennsylvania's Graduate School of Education in 2010. He will be an assistant professor of K–12 policy and leadership at the University of Southern California's Rossier School of Education in the fall of 2010. His primary research interest is the implementation of standards-based accountability and its effects on schools, teachers, and students. In addition to his work on the Vanderbilt Assessment of Leadership in Education, he has written about the content of state standards and psychometrics in the context of accountability.

Henry May, a senior research investigator at the Consortium for Policy Research in Education at the University of Pennsylvania, specializes in the application of quantitative research methods for evaluating the implementation and impacts of educational policies and programs. His research interests include the evaluation of the impacts of educational policies and programs (with particular interest in *multilevel randomized experiments*), the measurement of teacher and school effects on student performance, and the identification of contextual predictors of teacher quality and turnover.

Andrew C. Porter, dean and George & Diane Weiss Professor of Education at the University of Pennsylvania's Graduate School of Education, is an applied statistician and psychometrician whose research agenda has focused on assessment and accountability, content alignment, and the effects of curriculum policies.

Stephen N. Elliott, a professor of special education and the Dunn Family Chair of Educational and Psychological Assessment in Peabody College at Vanderbilt University, received his doctorate at Arizona State University in 1980. He teaches courses on measurement and assessment of academic and social behavior. He currently codirects three U.S. Department of Education research grants concerning the assessment of learning-focused school leadership and the validity of testing modifications and alternate assessments for students with disabilities. He also directs Peabody College's new interdisciplinary program in educational psychology.

Ellen Goldring, Patricia and Rodes Hart Chair and professor of education policy and leadership at Peabody College, Vanderbilt University, received her doctorate from the University of Chicago. Before coming to Vanderbilt, she was chair of the Department of Educational Administration, Tel Aviv University, Israel. Her area of expertise focuses on improving schools, with attention to educational leadership, school choice, and parent involvement. Her research on school leadership examines the implementation and effects of professional development, coaching, and performance feedback.

Joseph Murphy, the Frank W. Mayborn Chair of Education and associate dean at Peabody College of Education of Vanderbilt University, works in the area of school improvement, with emphasis on leadership and policy.



COPYRIGHT INFORMATION

TITLE: An Examination of Differential Item Functioning on the
Vanderbilt Assessment of Leadership in Education

SOURCE: J Sch Leadership 19 no6 N 2009

The magazine publisher is the copyright holder of this article and it is reproduced with permission. Further reproduction of this article in violation of the copyright is prohibited. To contact the publisher:
<http://www.rowmaneducation.com/Journals/JSL/>