Original Research

UC UNIVERSITY COUNCIL FOR EA EDUCATIONAL ADMINISTRATION

Journal of Research on Leadership Education
1–23

DOI: 10.1177/1942775117742646
journals.sagepub.com/home/jrl
SAGE

The Performance Assessment for Leaders: Construct Validity and Reliability Evidence

Margaret Terry Orr[1], Ray Pecheone[2], Liz Hollingworth[3], Barbara Beaudin[4], Jon Snyder[2], and Joseph Murphy[5]

Abstract

The Performance Assessment for Leaders (PAL) was developed by a team of nationally recognized experts in response to a Massachusetts requirement to determine and evaluate the leadership abilities of candidates seeking initial school principal licensure. This article describes and evaluates research conducted on all aspects of a 2014-2015 statewide field trial of PAL. Findings suggest that this assessment is a valid and reliable measure of individual candidate competence for granting initial school leader licensure, and is a positive, educative experience for candidates. It concludes with implications for use elsewhere.

Keywords

principal performance assessment, principal licensure, leadership preparation policy

Introduction

In recent years, there has been national criticism of teacher education and leadership preparation programs, along with greater policy demands to improve educator quality

[1]Fordham University, New York, NY, USA
[2]Stanford University, CA, USA
[3]University of Iowa, Iowa City, USA
[4]Avon, CT, USA
[5]Vanderbilt University, Nashville, TN, USA

Corresponding Author:
Margaret Terry Orr, Division of Educational Leadership, Administration and Policy, Graduate School of Education, Fordham University, 113 W. 60th Street, New York, NY 10023, USA.
Email: morr4@fordham.edu

as key means to improve student learning and achievement. Both issues have led states to adopt standardized performance measures of professional skills and qualities for licensure and evaluation of educators, particularly teachers and school leaders (Davis et al., 2011; Shelton, 2012). Among the policy considerations has been a call for valid and reliable performance measures of school leader readiness. Yet, until now, there has been little large-scale research on performance assessment for leadership preparation or principal licensure (Kochan & Locke, 2009). In fact, most existing assessment measures in educational leadership focus on principal practice and lack sufficient validity and reliability (Condon & Clifford, 2010), or they are state exams that focus only on demonstrating knowledge, rather than skills and practices (Davis et al., 2011).

Nonetheless, promising research and development work exists on performance assessment in teacher education and licensure (Pecheone & Chung, 2006; Pecheone, Pigg, Chung, & Souviney, 2005; Sandholtz & Shea, 2012). The most commonly used teacher education performance assessment, edTPA, consists of a series of teaching tasks in planning, instruction, and student assessment (Pecheone, Shear, & Darling Hammond, 2013). Repeated field trial analyses show that edTPA has strong construct validity and reliability in determining candidate readiness for an initial teaching position (Pecheone et al., 2013). Such results represent promising potential for developing similar assessments to determine candidate readiness for principal licensure.

One state, Massachusetts, had required statewide assessment of school leader readiness but lacked the resources to develop an instrument for doing so until 2011. With funds from the federal Department of Education's Race to the Top, Massachusetts engaged a team of national experts on leadership and performance assessments to design, implement, and validate a performance assessment system for principal licensure. The team's work, informed by a representative group of Massachusetts K-12 school and district leaders and higher education faculty, led to the creation and continued development and refinement of the Performance Assessment for Leaders (PAL) system, now incorporated into its state principal licensure system.

PAL was developed as four separate, but interrelated, performance tasks to provide clear evidence of a leadership preparation candidate's readiness for an initial school leadership position. The tasks require use of multiple sources of data (plans, reports, feedback results, video clips, and personal reflections and commentaries) that are organized around four core areas of school leader work. The areas are these: (a) setting direction by developing a plan for an area of school improvement, (b) creating a professional learning culture among school staff, (c) supporting individual teacher development, and (d) engaging families and community in improving student learning. To complete each PAL task, candidates investigate a school or student priority area (based on an academic area where one or more federally designated student subgroups performs less well) and focus for the task, engage in planning for the task, take action to accomplish the task, and solicit feedback and other evidence about the impact of the task, as well as reflect upon the leadership skills used. The PAL assessments are built upon the state's professional standards and indicators for administrative leadership (http://www.doe.mass.edu/lawsregs/603cmr7.html?section=10) and school leader

practices that research shows are effective in promoting school improvement (Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010; Louis, Leithwood, Wahlstrom, & Anderson, 2010; Murphy, Elliot, Goldring, & Porter, 2007).

Completion of all four PAL tasks is required of all candidates seeking initial school leader licensure, regardless of the type or extent of preparation they have undergone. Massachusetts offers three pathways for candidates: completion of a state-approved leadership preparation program (of which 23 are university-based, affiliated with a professional association, or sponsored by a regional educational agency), an administrative apprenticeship/apprenticeship, or a panel review process.

This article presents the results of a field trial of the Massachusetts Performance Assessment, which was completed by 422 candidates in 2014-2015. The results enabled the team to investigate the construct validity and reliability of the assessment system and its component tasks, including an analysis of whether results were associated with the nature of preparation or independent candidate demographic characteristics. The field trial also allowed the team to examine the educative benefits, an assessment attribute, of the performance assessments for candidates. The investigation addressed these questions:

- How valid are the indicators, domains, and total scores for the four performance assessment tasks in evaluating leadership candidate performance?
- What is the relationship among the task scores as independent but related measures of leadership readiness?
- How reliable are the assessment scores in differentiating candidate readiness?
- Given the educative intent of this performance assessment, what are the learning consequences for candidates in completing the four PAL tasks?

Research Background

Large-scale research studies and meta-analyses consistently stress how the leadership abilities of educators' influence student learning; effective leaders develop a vision for educational success and set the direction for attaining it, improve teaching and learning, manage resources and operations, facilitate change, and work effectively with families and communities (Bryk et al., 2010; Leithwood & Jantzi, 2008; Robinson, Lloyd, & Rowe, 2008). Recent research also has drawn attention to the influence of specific leadership practices, particularly teacher supervision and support (May & Supovitz, 2011), the development of professional learning communities (Hayes, Christie, Mills, & Lingard, 2004; Supovitz & Christman, 2005), and family and community engagement (Sebring, Allensworth, Bryk, Easton, & Luppescu, 2006; Weiss & Stephen, 2009). Together, this research demonstrates that certain leadership practices, when performed effectively, can increase school effectiveness and thus student learning. Furthermore, research shows that the way that school leaders are prepared positively influences their leadership ability and thereby improves school practices, as reported by teachers (Orphanos & Orr, 2014; Orr & Orphanos, 2011). Nevertheless, there is little available research on how to reliably assess the effectiveness of

candidates' leadership preparation or determine their readiness for principal licensure (Kochan & Locke, 2009), despite the availability of this research and other evidence.

Still, most states do require completion of an accredited leadership preparation program as part of the licensure process and many require candidates to complete leadership exams as another part of the licensure process (Shelton, 2011). Moreover, state policy analysts stress that candidate assessments tied to professional standards are an important element of exemplary leadership preparation (Shelton, 2012). The analysts recommend that states use new forms of performance assessment for principal licensure that measure "the more complex skills research shows effective school leaders need to have" (Briggs, Cheney, Davis, & Moll, 2013, p. 30). Similarly, federal policy (U.S. Department of Education, 2009) and national accreditation associations (Council for the Accreditation of Educator Preparation, 2015; Educational Leadership Constituent Council, 2011) require states, universities and leadership preparation programs, respectively, provide authentic evidence of school leader effectiveness and candidate readiness for school leadership work.

Among the possible forms of assessment, performance-based assessment appears to be most appropriate to determining candidate readiness for initial leadership work. It is more robust and educational than cognitive-based assessments and is more applicable to assessing complex skills (Gitomer, 1993, 2012). Measurement specialists (Linn, Baker, & Dunbar, 1991) have long advocated for authentic assessments that, as Linquist (1951) explains, "require the examinee to do the *same* things, *however complex*, that he is require to do in the criterion situations" (Linquist, 1951, p. 154). Wiggins (1993) defines performance assessments as authentic assessments that have candidates addressing "Engaging and worthy problems or questions of importance" (p. 228) in which candidates demonstrate readiness by using "knowledge to fashion performances effectively and creatively" (p. 228).

Direct assessments of what candidates are able to perform appear to have "the potential of enhancing validity" (Linn et al., 1991, p. 16). But this appearance provides only face validity, which is insufficient (Linn et al., 1991). Other evidence is needed to support the interpretations of the assessment evidence as measures of performance and demonstrate the assessments' technical adequacy, using available well-established psychometric criteria for evaluating validity and reliability (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Linn et al., 1991).

The design and use of performance assessments to assess educators' professional practice have been limited until recently (Condon & Clifford, 2010; Kennedy, 2010). Yet, performance assessments, including portfolio assessments, have been widely used in medicine to assess clinical skills, teamwork, and leadership practices (Epstein, 2007; Havyer et al., 2014; Kiesewetter et al., 2013). Several psychometric studies of performance assessment in medicine show both promise and challenges. The studies have demonstrated that the assessments have good validity and reliability results, while being costly and time-consuming, and requiring well-qualified scorers to determine candidate proficiency (Gadbury-Amyot, McCracken, Woldt, & Brennan, 2014; Gerhard-Szep et al., 2016; Havyer et al., 2014).

There have also been promising results about assessments covering teacher education. According to Pecheone and Chung (2006), there is evidence that teacher performance assessments better evaluate instructional practice, provide powerful learning experiences for the teacher candidates, and are predictive of teacher effectiveness as measured by student learning gains. In a series of field studies, researchers have found strong validity evidence (Pecheone & Chung, 2006) and strong reliability evidence (Pecheone & Wei, 2007) that meet national psychometric standards (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) for the multiple-dimension measures in Performance Assessment for California Teachers (PACT). More recent research on edTPA, the nationally implemented teacher candidate performance assessment, shows similarly strong validity and reliability evidence (Lalley, 2017; Pecheone, Shear, Whittaker, & Darling Hammond, 2013).

Prior assessment development on educational leadership readiness has been limited to cognitive-based assessments (Grissom, Mitani, & Blissett, 2017) and 360-degree feedback assessments (Condon & Clifford, 2010; Porter et al., 2010), with limited or mixed psychometric results, or significant utilization costs. The broader leadership development field has even more limited assessment resources, relying primarily on self-reports, observation reports, and 360-degree assessments (Black & Earnest, 2009; Hoole & Martineau, 2014; Ozgen, Sanchez-Galofre, Alabart, Medir, & Giralt, 2013; Solansky, 2010). While these assessment tools have strong face validity and reliability measures, they lack the authenticity of performance assessments and are more indirect measures of leadership competence.

Taken together, there is a strong need for valid and reliable performance assessments of educational leader candidate readiness. While few psychometrically sound tools exist to assess leader candidates' skills in performing authentic work, performance assessment experiences from medicine and teacher preparation show promise for the development and use of leadership performance assessments. Psychometric standards for assessment development, particularly for performance assessments, provide critical guidance for validation and reliability expectations.

The Massachusetts PAL

The PAL assessments were created by an assessment development team, that included the authors, through a multistep assessment development process described in a related article (Orr et al, 2017). The PAL assessments are made up of the following four tasks (see Appendix A for a description of each): Task 1: Leadership through a vision for high student achievement; Task 2: Instructional leadership for a professional learning culture; Task 3: Leadership in observing, assessing, and supporting individual teacher effectiveness; and Task 4: Leadership for family engagement and community involvement.

The four independent tasks were designed to enable leadership candidates to demonstrate relevant skills and practices as defined by three or four domains for each task. A domain refers to a general standards-based area of growth and development for a specific component of the performance assessment task. An indicator is what candidates

should know and be able to demonstrate within the domain when completing a task. Specifically, "procedural knowledge that is specific to a knowledge domain or subdomain is referred to as domain-specific based skills" (Tombari & Borich, 1999, p. 103). Altogether, PAL has 13 domains and 26 indicators of these domains, and these are listed in Appendix C.

Candidates produce a portfolio of work that includes three or four artifacts for each task, commentary on leadership skills used to complete each task, and supporting documents that inform scoring (the artifacts are summarized in Appendix B and explained in the companion article; Orr, et al., 2017). Candidates can complete and submit their work products at any time but are only scored when each task portfolio is complete. The portfolio of work is scored by indicator, and the indicator scores are aggregated as domain scores. Scores for each domain consist of the average of one to three indicator scores and each task has three to four domains.

The 13 domains and 26 indicators for the four tasks are scored by rubrics that scale candidate proficiency on a 4-point continuum of beginning, developing, meeting, and exceeding skill and practice expectations reflecting a beginning school leader (see Appendix B for descriptions of the domains and indicators for each task). The domains and indicators are listed in Appendix C.

Valid and reliable scoring depends upon the quality of scorer recruitment, training, and supervision. For the PAL assessment, a scorer must be an experienced Massachusetts school or district leader or preparation program faculty member. Thirty scorers were selected for the field trial and were given 10 to 15 hours of training per task, followed by ongoing supervision and support to maintain scorer reliability. All scorers rated only one task per candidate, ensuring independence in scoring and reducing possible scorer bias. Thus, each candidate was rated by four different scorers. All scoring was done online, using a confidential assessment management system, ShowEvidence.

Data Sources and Methods

To evaluate the construct validity and reliability of PAL, the assessment was field-tested statewide in 2014-2015. All candidates seeking initial principal licensure, regardless of their preparation pathway, were required to complete all four PAL tasks, although no minimum performance standard had been established and no payment of fees was required. In all, 422 candidates submitted all four tasks; of them, 25% were double-scored to evaluate interrater reliability. Six candidates had one or more tasks that were scored as "complete" without a numeric score, so their scores were removed from this analysis. This analysis is based on the remaining 416 candidates (99% of the total).

Table 1 shows that most candidates with four task submissions were from a formal preparation program, female, and White.

Data used in this analysis comprise the candidates' indicator, domain, and total scores on the four tasks, and information about their demographic characteristics. Indicator scores are the ratings of a candidate's performance (1-4) on individual rubric indicators. Domain scores are an average of the indicator scores that are related to

Table 1. Number and Percentage Distribution of Candidates Who Completed All Four Tasks for the PAL Field Trial, by Characteristics.

Candidate characteristics	Number of candidates	% of all candidates
Candidate preparation pathway		
Preparation program	341	82
Alternative pathway (administrative internship or panel review)	75	18
Gender		
Female	266	64
Male	150	36
Race/ethnicity and national origin		
White	309	74
African American	14	3
Hispanic	11	3
Asian	9	2
Native Hawaiian, Pacific Islander	2	1
Native American	1	[a]
Multirace, non-Hispanic	15	1
Preferred not to answer	65	16
Total number of candidates	416	100

Note. PAL = Performance Assessment for Leaders.
[a]Less than 1%.

specific domains of practice. As previously noted, there are one to three indicators per domain. The total task scores are an average of the domain scores within one task. The overall score is an average of the four total task scores (1 = beginning, 2 = developing, 3 = meeting, and 4 = exceeding).

To validate the use of the performance assessment standards–based task scores to evaluate leadership candidates, the assessment development team used the current conception of validity as outlined by the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). To evaluate the construct validity of the PAL assessment system, we examined the scoring at four levels: (a) the quality of each indicator as a measure of specific aspects of initial school leader knowledge and skills, (b) the relationship among the indicator scores as a coherent measure for their respective domain, (c) the relationship among the domain scores as a coherent measure of specific leadership knowledge and skills, and (d) the relationship among the tasks as both unique and complementary within the overall PAL assessment as a combined measure of initial school leadership knowledge and skills. We also evaluated the scoring reliability, doing two forms of reliability assessment described below.

A key principle of performance assessment is that it be educative for the participating candidates in completing the required tasks (Linn et al., 1991). As in the research

on teacher performance assessments (Pecheone & Chung, 2006), we expected that completing the PAL tasks would increase candidates' knowledge about leadership practices and their skills through the authentic nature of the tasks. To evaluate the educative benefit of the PAL assessment, we examined candidates' explanations of the learning consequences in completing the four PAL tasks, as reported in an online feedback survey fielded at the end of the field trial. All candidates were surveyed and asked to identify what they learned in completing each task. About half of the 92 survey respondents (22% of the completers) provided written answers. These responses were analyzed for common themes, using qualitative content analyses techniques (Miles & Huberman, 1994), and the results were summarized by the emergent categories for reporting purposes.

Results

The usefulness of PAL for principal licensure decisions and preparation program improvement depends upon the degree to which it is valid and reliable. Assessment validity here refers to the extent to which candidates can be differentiated between those who meet or exceed performance expectations and those who do not. The focus here, therefore, is on construct validity, independence of measures, reliability, and learning consequences.

Construct Validity

We start with the construct validity of PAL. Construct validity is defined as the degree to which a test measures what it claims, or purports, to be measuring. We began by examining the indicators and domains for each task. While not shown here, the means and standard deviations for the indicators for each task for the field trial sample were somewhat similar within each task, which means that the indicator measures were performing similarly. Table 2 shows the means and standard deviations for the domains and total scores for each task. Like the indicators, the domain scores were fairly similar within each task, suggesting that they were performing similarly.

The domains within each task (not shown here) were moderately to strongly positively correlated: .591 to .681 for Task 1; .551 to .729 for Task 2, .527 to .714 for Task 3, and .726 to .772 for Task 4. All correlations were statistically significant and support the intended design of the assessment. These correlations reflect the strong internal consistency of the domains (as all were above .50) as well as their distinctiveness (as none were higher than .77), suggesting that the constructs measured are related but not identical, also supporting the intended design.

To determine the relationship among the domains, we conducted an exploratory factor analysis on the correlation matrix among all domain scores.[1] An exploratory factor analysis is a multivariate statistical method that explores the underlying structure of a set of variables. The results, presented in Table 3, show that there were four factors and that they corresponded perfectly to the four tasks. Each domain score had a high positive loading on its associated task factor and near-zero loadings on other

Table 2. Means and Standard Deviations for Domain Scores and Total Scores for Each Task (n = 416).

Tasks and domains	*M*	*SD*
Task 1: Leadership through a vision for high student achievement	**2.76**	**0.56**
• T1D1: Investigate and prepare a vision	2.75	0.63
• T1D2: Design an integrated plan for strategies	2.73	0.63
• T1D3: Assess and analyze feedback from participants and self-analysis	2.80	0.65
Task 2: Instructional leadership for a professional learning culture	**2.97**	**0.53**
• T2D1: Plan to facilitate team learning	2.98	0.56
• T2D2: Enact a professional learning culture	2.95	0.61
• T2D3: Assess team learning to improve ongoing group learning and self-analysis	2.98	0.66
Task 3: Leadership in observing, assessing, and supporting individual teacher effectiveness	**2.86**	**0.45**
• T3D1: Plan	2.90	0.51
• T3D2: Conduct the observation	2.85	0.51
• T3D3: Provide feedback and suggest support	2.91	0.49
• T3D4: Assess—Analyze and identify implications	2.77	0.71
Task 4: Leadership for family engagement and community involvement	**2.66**	**0.67**
• T4D1: Plan to promote family and community involvement	2.67	0.68
• T4D2: Implement an engagement or involvement strategy	2.68	0.80
• T4D3: Analyze feedback from participants and assess leadership skills	2.64	0.77

Note. Bold values are to highlight the task scores as distinct from the domain scores.

task factors. By "loading," we mean the relationship of each variable to the underlying factor. These findings confirm that the domains within each task are strongly related and make unique contributions as separate measures within the task, while contributing little to the other tasks, capturing different dimensions of school leadership practice.

We calculated the correlations between task scores, as shown in Table 4, to evaluate the degree of relatedness among the tasks. The correlations among tasks were moderate and positive, ranging from 0.208 to 0.269 across the four tasks. The moderate and consistent positive correlations among tasks suggest a one-factor model could be fit to the task scores but would leave considerable variance in task scores unexplained by the single factor, confirming our use of four measures, one for each task.

Next, we evaluated the quality of the task scores and the total score. Table 2 shows the means and standard deviations for the four tasks. While the means were fairly similar, their standard deviations varied somewhat differently: from a narrower standard deviation of .45 for Task 3 to the broader standard deviation of .67 for Task 4. These differences may be related to differences in preparation program emphasis, as

Table 3. Standardized Factor Loadings.

Domain	F1	F2	F3	F4
T1D1	**0.84**	−0.05	−0.02	−0.04
T1D2	**0.87**	−0.06	−0.03	0
T1D3	**0.73**	0.09	0.01	0.04
T2D1	0.03	**0.62**	0.04	0.02
T2D2	−0.06	**0.92**	−0.03	−0.04
T2D3	0.02	**0.80**	−0.04	0
T3D1	−0.01	−0.03	**0.80**	−0.03
T3D2	−0.02	0.08	**0.82**	−0.06
T3D3	−0.03	−0.09	**0.92**	0.03
T3D4	0.08	0.03	**0.64**	0.05
T4D1	−0.01	0	−0.03	**0.87**
T4D2	−0.03	−0.06	−0.01	**0.94**
T4D3	0.03	0.06	0.05	**0.79**

Note. Strong factor loadings, above .60, are bold-faced to highlight their factor alignment. T = task (1-4); D = domain (1-4).

Table 4. Task Correlation for Four Tasks (n = 416).

	Tasks			
Tasks	T1	T2	T3	T4
Task 1	1.00	—	—	—
Task 2	.257**	1.00	—	—
Task 3	.269**	.208**	1.00	—
Task 4	.229**	.251**	.243**	1.00

**Statistically significant at $p < .01$ (one-tailed).

explained in discussions with preparation program faculty and results from faculty feedback survey: Programs were strong on supervision (which corresponds to Task 3) but weak on leadership for family and community engagement (which corresponds to Task 4) (Orr, Pecheone, Shear, Hollingworth & Beaudin, 2016).

Independence of Measures

Next, as part of a bias review process, we investigated whether candidates performed differently, and received different scores, on the PAL assessments based on their demographic attributes. An ANOVA was used to test possible differences among group means. The results, presented in Table 5, show statistically significant differences in performance by gender, with females performing .12 points higher than males, and by preparation pathway, with preparation program

Table 5. Mean and Standard Deviation for Domain Total Score by Demographic Attribute.

Candidates demographics	*n*	*M*	*SD*	Statistical significance
Preparation pathway				Statistically significant, $F(2, 414) = 6.82$, $p = .001$
Preparation program	341	2.86	.39	
Alternative pathway	75	2.80	.44	
Gender				Statistically significant, $F(2, 414) = 8.034$, $p = .000$
Female	266	2.89	.39	
Male	150	2.77	.40	
Race/ethnicity				Not calculated due to small *N*
White	309	2.86	.40	
African American	14	2.80	.39	
Hispanic	11	2.84	.57	
Asian	9	2.88	.28	
Total	416	2.82	.39	

candidates scoring somewhat higher than alternative pathway candidates. Due to missing information and low numbers, the comparisons by race/ethnicity can only be viewed as trends.

Because standards setting was not conducted before the field trial, we do not know the extent to which these mean score differences might affect the two groups differently. As a result of these findings, the scores will be evaluated again following the 2015-2016 assessment program year to determine whether improvements in the scorer training and supervision, instructions, rubrics, and standards setting reduce any differences or whether modifications are needed to address problem areas.

Reliability

We evaluated scoring reliability using submissions that were scored by two scorers to determine scorer agreement. The interrater reliability, or agreement, is the degree to which there was agreement among raters. Exact agreement rates (scorers assigning the same exact score) and total agreement rates (scorers assigning either the same or adjacent scores) were calculated for each indicator. A version of Cohen's κ referred to as κ_n was computed for each agreement rate. This statistic provides a type of "chance-corrected" agreement, where values near 1 represent higher agreement than values near zero; however, there are no set guidelines for what constitutes an adequate value.[2]

The results in Table 6 show that exact rates are above 50% on most rubrics, indicating the percentage of cases where scorers scoring the same portfolio assigned the same score. Exact agreement is below 50% on five out the six Task 4 rubrics. This lower rate of agreement suggests that further scorer training for Task 4 may be needed to improve scorer reliably interpreting evidence and the rubrics evaluated to determine their use in guiding reliable scoring.

Table 6. Scorer Agreement Rates by Indicator.

Task	Indicator	Exact	Kappa (exact)	Exact + adjacent	Kappa (exact + adjacent)	*n*
1	1	0.73	0.64	0.98	0.95	100
	2	0.65	0.53	0.98	0.95	100
	3	0.64	0.52	0.95	0.87	100
	4	0.72	0.63	0.99	0.97	100
	5	0.71	0.61	0.97	0.92	100
	6	0.69	0.59	0.99	0.97	100
	Average	0.69	0.59	0.98	0.94	
2	7	0.59	0.45	0.98	0.95	102
	8	0.69	0.58	0.97	0.92	99
	9	0.63	0.51	0.98	0.95	100
	10	0.60	0.46	0.95	0.87	102
	11	0.50	0.34	0.97	0.92	101
	12	0.46	0.28	0.97	0.92	102
	Average	0.58	0.44	0.97	0.92	
3	13	0.63	0.50	0.99	0.97	83
	14	0.61	0.48	1.00	1.00	80
	15	0.59	0.46	0.98	0.93	81
	16	0.70	0.60	0.96	0.90	80
	17	0.71	0.61	0.99	0.97	83
	18	0.59	0.46	0.98	0.93	81
	19	0.59	0.46	0.96	0.90	81
	20	0.55	0.40	0.94	0.84	82
	Average	0.62	0.50	0.97	0.93	
4	21	0.46	0.28	0.95	0.86	93
	22	0.42	0.23	0.87	0.66	93
	23	0.38	0.17	0.91	0.77	92
	24	0.46	0.28	0.90	0.74	92
	25	0.53	0.37	0.92	0.80	93
	26	0.49	0.32	0.91	0.77	92
	Average	0.46	0.27	0.91	0.76	

Next, we conducted a generalizability analysis, or *G* study, which is a statistical framework for conceptualizing the reliability or dependability of a set of scores. *G* studies provide estimates of the variability of measures (Shavelson & Webb, 2005). Table 7 presents the estimated relative *G* coefficients (similar to a reliability coefficient in Classical Test Theory[3]) for each task, when task scores are calculated as the average of domain scores. Estimated coefficients assuming either one or two scorers for each task are presented (with two scorers, this is the reliability of an average task score across two independent scorers). The reliability looks substantially better for Tasks 1, 2, and 4 and low for Task 3, likely due to the low variance among candidates'

Table 7. Estimated Reliability (*G*) Coefficients by Task and Number of Scorers.

	Number of scorers	
Task	1	2
1	0.728	0.842
2	0.656	0.792
3	0.208	0.345
4	0.581	0.735

Table 8. Reliability of Total Score Based on the Number of Scorers for Each Task.

Reliability coefficient	Number of scorers
.771	1
.844	2
.879	3

Task 3 scores. The reliability coefficients are moderate for all tasks with one scorer, but are all above 0.70 for Tasks 1, 2, and 4 with two scorers.

To further evaluate the scores, we assessed the potential reliability of a total score computed as the average of all four task scores, using a stratified coefficient alpha,[4] a more appropriate measure of reliability when a test consists of items (or parts) drawn from distinct categories but that are intended to measure the same primary construct. The resulting stratified coefficient alpha estimates, in Table 8, show that there is sufficient reliability with two scorers.

Educative Benefits

To evaluate the educative benefits of completing the PAL assessment, we collected and analyzed the candidates' experiences for each task separately. These results are presented below, showing that almost all candidates who responded reported learning benefits as a result of completing each task.

For Task 1, 50 candidates (54% of the 92 survey respondents) provided written feedback on what, if anything, they learned from Task 1. Forty reported one more areas of learning, and 10 (20% of the written responses) reported learning nothing or finding it to be a waste of time. Of those who reported learning something (80%), most focused on the following:

- Planning in general: "I learned that identifying a priority area often encompasses so many different areas, it can become very complicated and it's important to talk about the priority area with many perspectives and buy-in to the community."

- Use of data: "How to more effectively use data to inform instruction."
- Learning to work with others in the planning process: "How to listen to the input of others and use data to make informed decisions."
- How to put a plan together: "I learned how to put together a plan that was relevant for school improvement."

For Task 2, 45 candidates provided written responses (49% of the 92 survey respondents), and, of these, 87% identified one or more areas of learning, while six reported none. The areas of learning included the following:

- Fostering a professional learning group to improve learning: "Working with teacher team to improve instructional practice."
- Facilitating a group of teachers: "Learning how to work with a team and create and facilitate professional learning communities."
- Facilitating improvements in teacher practice: "Finding ways to support teachers to help students in the priority group. Subsequent to the task, I had the skills and opportunities to provide more strategies for my colleagues."

For Task 3, 40 candidates (44% of the survey respondents) supplied answers about what they learned from completing the task (only three reported that they had not learned anything from doing this task). For most who cited learning something from completing these tasks, their increased knowledge centered around the following components of the task or the combination of task elements:

- Conducting the preobservation conference with the teacher: "spend more time in preobservation conference talking about the standards"
- Observing the teacher's classroom practice: "being goal-oriented," "focus on all parts of an observation"
- Developing rapport with, and engaging, the teacher in learning about his or her practice (reflective questions, listening): "It is important to let the teacher reflect on his or her own practice. As a leader, it is best to give input but allow the teacher to find and correct [his or her] own mistakes to grow . . ."
- Providing good feedback: "How to give effective feedback based on an instructional focus" being constructive, and providing "actionable feedback."
- Using strategies and practices for effective teacher observation and feedback: Being timely in providing feedback after an observation, using the video to provide feedback, and practicing the feedback discussion first.

Finally, for Task 4, 40 candidates (44% of the 92 survey respondents) provided written responses. Only three said they had learned nothing or little from completing this task. Of the remainder, their new knowledge was divided between learning how to overcome the challenges of engaging family and community members in improving student learning and the student health, and learning how to confront the social or emotional issues they were trying to address through family and community engagement:

- Recognizing the importance of family and community engagement: "Family and community engagement is very important to a successful school, especially in a diverse school like mine."
- Fostering collaboration with various stakeholders, particularly in diverse settings, and engaging multiple perspectives in planning: "Understanding the complexity of stakeholder perspectives."
- Creating a family and community engagement plan that supports student learning: "Communities will unite on an initiative if it is meaningful, organized, and student-based."
- Appreciating the usefulness of family feedback surveys as part of planning: "I learned that surveys and parent feedback is valuable."
- Recognizing the importance of principal leadership in fostering family and community engagement: "An understanding of how challenging it is to get meaningful family and community engagement/planning."

The strong, positiveand detailed written responses from half or more of the candidate survey respondents, and few negative ones, provide support for the beneficial learning consequences for them in completing each of the four PAL tasks.

Discussion

Taken together, these findings confirm that the PAL assesses multiple but related dimensions of initial school leader practice. The domains within each task and each task overall have similar means and sufficient variance to support their common use in differentiating candidate readiness for initial school leadership. Moreover, the factor analysis shows that dimensions are strongly related aspects of school leadership and make unique contributions as separate measures within the task, while contributing little to the other tasks. The correlational analysis of the task scores shows that the task measures are modestly related. Candidate feedback on the learning consequences of the four tasks confirmed the development of leadership skills in the four task areas and the educative benefits of the assessment experience for many candidates. The analysis of the measure independence shows possible influence based on gender and preparation pathway, which will require follow-up in subsequent assessment years. Finally, scorer reliability reinforces the use of the assessment in evaluating candidate readiness. The weaker scorer reliability for Task 3 (despite strongly positive results on the factor analyses, correlations, and scorer agreement ratings) suggests further investigation in subsequent assessment years to evaluate whether the lower reliability is due to the lower variance in candidate performance or whether improved scorer training and clarified instructions and rubrics strengthen this reliability.

Conclusion

The Massachusetts PAL is the first validated performance assessment of candidates for initial school leadership readiness in the nation. Based on the research findings, we—as

the team commissioned to design and implement this performance assessment—concluded that the PAL's field testing process yielded important evidence of the assessment's validity and reliability for the purposes of granting licenses to new school leaders in Massachusetts. We also think that PAL has potentially application for other states and localities to measure initial school leader readiness. These findings build on a growing body of research about the development of performance assessments in teacher education that are standards-based and can be used as part of the teacher credentialing process (Pecheone & Chung, 2006; Pecheone et al., 2013). The results also correspond positively to available research on performance assessment in medicine (Gerhard-Szep et al., 2016; Havyer et al., 2014), further encouraging more widespread use.

Comparisons of the field trial candidate scores and subsequent year's[5] candidate scores will yield useful insight into how much the scoring reliability improved, based on changes in the task instructions and rubrics, and in scorer training, and whether candidate performance is influenced by the cut scores used to determine candidate readiness for licensure. Future validity studies that investigate concurrent validation evidence, such as preparation program faculty assessments of candidates, and longitudinal studies of the PAL assessments' predictive validity will strengthen its use in licensure and informing preparation program improvement.

In summary, the PAL assessment provides a valid approach to differentiating leadership candidate readiness in four unique but important and related domains of principal work. The differentiation of candidate performance based on preparation pathway is promising, providing possible evidence of the value of graduate preparation for leadership readiness. The candidates' report of learning consequences of completing the assessments confirms the educative benefits of performance assessments, providing further support for their replication and use in assessing leadership candidate readiness as initial school leaders in other states and localities.

Appendix A

The Four Performance Assessment for Leaders (PAL) Assessment Tasks

Task 1: Leadership through a vision for high student achievement. For Task 1, candidates develop a school vision and improvement plan for one school-based priority area. Specifically, they collect and analyze quantitative and qualitative data on student performance, student and teacher relationships, and student and school culture; select a priority area for focus; document existing school programs, services, and practices; and develop a set of goals, objectives, and action strategies with input from school leaders and key stakeholder groups. After presenting their plan, candidates receive feedback from relevant stakeholders.

Task 2: Instructional leadership for a professional learning culture. For Task 2, candidates demonstrate the capacity to foster a professional learning culture to improve student learning by working with a small group of teachers using structured learning activities to improve the teachers' knowledge and skills. They support teachers in improving an existing curriculum, instructional approach, or assessment strategy.

Task 3: Leadership in observing, assessing, and supporting individual teacher effectiveness. For Task 3, candidates demonstrate instructional leadership skills by planning for a teacher observation, conducting the observation, analyzing the observation and student performance data, providing feedback to the teacher observed, and planning support for that teacher. Candidates also document the observation cycle and teacher feedback on the quality and use of the process.

Task 4: Leadership for family engagement and community involvement. In Task 4, candidates gather information related to family engagement and community involvement needs, develop a proposal, and implement one component of it with work group support. They assemble and work collaboratively with a work group representing school leadership, staff, families, community members, and students (where appropriate) to select a priority area based on the evidence of student strengths, interests, and needs. With the work group, candidates develop a comprehensive improvement proposal, and implement and monitor the outcomes for one strategy.

Appendix B

PAL Assessment Artifacts and Commentary by Task.

Task	Artifacts	Commentary
Task 1	Artifact 1—Priority area and its context Artifact 2—Plan for action strategies Artifact 3—Findings, feedback, and recommendations	Leadership skills developed and learning experienced
Task 2	Artifact 1—Identify priority area, professional group, and professional learning plan Artifact 2—Describe the learning process and work accomplished by the group, emphasizing candidate's role Artifact 3—Present group member feedback on the process, learning, and benefits	Leadership skills developed and learning experienced
Task 3	Artifact 1—Preobservation template Artifact 2—Teacher observation video recording Artifact 3—Postobservation meeting video recording Artifact 4—Analysis of observed teaching Artifact 5—Teacher feedback on observation and postconference	Leadership skills developed and learning experienced
Task 4	Artifact 1—Analysis of data, priority area, and plan Artifact 2—Implementation of one strategy Artifact 3—Feedback on plan and strategy implementation	Leadership skills developed and learning experienced

Note. PAL = Performance Assessment for Leaders.

Appendix C

PAL Task Domains and Indicators.

Task	Domain	Indicator
Task 1: Leadership through a vision for high student achievement	Investigate an academic priority and student needs	Data collection Data analysis and priority definition
	Design an integrated plan to develop and implement improvement in the priority area	Vision and plan focus Plan details
	Assess feedback from school leaders and analyze own skills	Plan feedback Planning analysis
Task 2: Instructional leadership for a professional learning culture	Planning to facilitate team learning	Team identification Team learning plan
	Foster a professional learning culture to support team learning	Team process Team learning and work
	Assess team learning to improve ongoing group learning and leadership skills	Assessing team process and teamwork Assessing leadership skills and practices
Task 3: Leadership in observing, assessing, and supporting individual teacher effectiveness	Plan and prepare to observe	Observation focus selection Preobservation conference
	Conduct the observation	Use and application of a teacher observation rubric Description of teacher observation
	Provide feedback and suggest support	Feedback content Rapport and teacher engagement Teacher development
	Analyze feedback and assess leadership skills	Assessment of leadership skills and practices
Task 4: Leadership for family engagement and community involvement	Plan to promote family and community involvement with others	Investigation of a priority area Investigation of work group engagement Preparation of the plan, including strategies
	Implement an engagement or involvement strategy	Implementation of the strategy
	Analyze feedback from stakeholders and assess leadership skills	Assessment and analysis of feedback on the family and community engagement plan and strategy Assessment of leadership skills and practices

Note. PAL = Performance Assessment for Leaders.

Authors' Note

Copies of the PAL technical report, *Massachusetts Performance Assessment for Leaders (PAL) Technical Report Summary of Validity and Reliability Studies for 2014-15 Field Trial of PAL*, are available from the department or from the authors.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was made possible by contract support from the Massachusetts Department of Elementary and Secondary Education.

Notes

1. Analyses in this section are based on the $n = 361$ submissions with complete scores for all indicators from the primary scorer for the submission (of the 416 completed submissions, some of which lacked all indicator scores). Factor analysis of the indicator score correlation matrix yielded similar results, but the residual correlation matrix suggested the higher correlation among indicators from the same domains was not well modeled by a four-factor solution and the domain score correlation matrix was more appropriate to analyze.
2. Landis and Koch (1977) suggest the following metrics for evaluating agreement: ≤0 = poor, .01–.20 = slight, .21–.40 = fair, .41–.60 = moderate, .61–.80 = substantial, and .81–1 = almost perfect. But the authors suggest that these are suggested and that one has to be careful about making blanket assumptions about their use (Landis & Koch, 1977).
3. Classical Test Theory posits that a person's observed score on a test (X) is the sum of a true, error-free score (T) and an error score (E); $X = T + E$. The reliability of X is defined as the ratio of true score variance to the observed score variance.
4. Feldt and Brennan (1989) develop the stratified coefficient alpha to be used when the items in each stratified layer of test items can be assumed to be unidimensional.
5. The year (AY 2015-2016) following the field trial was the first year of full implementation of Performance Assessment for Leaders (PAL) in Massachusetts, at which time candidates paid a fee for the assessment and had to achieve a defined cut score on all four assessment tasks to be recommended for initial school leader licensure.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Black, A. M., & Earnest, G. W. (2009). Measuring the outcomes of leadership development programs. *Journal of Leadership & Organizational Studies*, *16*, 184-196.

Briggs, K., Cheney, G. R., Davis, J., & Moll, K. (2013). *Operating in the dark: What outdated state policies and data gaps mean for effective school leadership*. Dallas, TX: George W. Bush Institute.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). *Organizing schools for improvement: Lessons from Chicago*. Chicago, IL: University of Chicago Press.

Condon, C., & Clifford, M. (2010). *Measuring principal performance: How rigorous are commonly used principal performance assessment instruments?* Naperville, Il: American Institutes for Research.

Council for the Accreditation of Educator Preparation. (2015). *CAEP evidence guide*. Retrieved from filehttp://caepnet.org/#/media/Files/caep/knoweldge-center/caep-evidence-guide.pdf?la=en

Davis, S., Erickson, D. E., Kinsey, G. W., Moore-Steward, T., Padover, W., Thomas, C., . . . Wise, D. (2011). Reforming the California Public School Administrator licensure system through the alignment of research, policy, and practice: Policy perspectives and recommendations from the California Association of Professors of Educational Administration (CAPEA). *CAPEA Education Leadership and Administration*, *22*, 66-82.

Educational Leadership Constituent Council. (2011). *Program report for the preparation of educational leaders*. Washington, DC: National Policy Board for Educational Administration.

Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine*, *356*, 287-396.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (pp. 105-146). New York, NY: Palgrave Macmillan.

Gadbury-Amyot, C. C., McCracken, M. S., Woldt, J. L., & Brennan, R. L. (2014). Validity and reliability of portfolio assessment of student competence in two dental school populations: A four-year study. *Journal of Dental Education*, *78*, 657-666.

Gerhard-Szep, S., Guntsch, A., Pospiech, P., Sohnel, A., Scheutzel, P., Wassmann, T., & Zahn, T. (2016). Assessment formats in dental medicine: An overview. *GMS Journal for Medical Education*, *33*(4), Doc65.

Gitomer, D. H. (1993). Performance assessment and educational measurement. In W. C. Ward & R. E. Bennett (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing and portfolio assessment* (pp. 241-264). Hillsdale, NJ: Lawrence Erlbaum.

Gitomer, D. H. (Ed.). (2012). *Performance assessment and educational assessment*. New York, NY: Routledge.

Grissom, J. A., Mitani, H., & Blissett, R. S. (2017). Principal licensure exams and future job performance: Evidence from the School Leaders Licensure Assessment. *Educational Evaluation and Policy Analysis*, *20*(10), 1-33.

Havyer, R. D., Wingo, M. T., Comfere, N., Nelson, D. R., Halvorsen, A. J., McDonald, F. S., & Reed, D. A. (2014). Teamwork assessment in internal medicine: A systematic review of validity evidence and outcomes. *Journal of General Internal Medicine*, *29*, 894-910.

Hayes, D., Christie, P., Mills, M., & Lingard, B. (2004). Productive leaders and productive leadership: Schools as learning organisations. *Journal of Educational Administration*, *42*, 520-538.

Hoole, E., & Martineau, J. (2014). Evaluation methods. In D. Day (Ed.), *The Oxford handbook of leadership and organizations* (pp. 167-196). Oxford, UK: Oxford University Press.

Kennedy, M. (Ed.). (2010). *Teacher assessment and the quest for teacher quality*. San Francisco, CA: Jossey-Bass.

Kiesewetter, J., Schmidt-Huber, M., Netzel, J., Krohn, A. C., Angstwurm, M., & Fischer, M. R. (2013). Training in leadership skills in medical education. *GMS Zeitschrift für medizinische Ausbildung*, *30*(4), Doc49.

Kochan, F. K., & Locke, D. L. (2009). Student assessment in educational leadership preparation programs. In M. D. Young, G. Crow, J. Murphy, & R. T. Ogawa (Eds.), *Handbook on the education of school leaders* (pp. 417-456). New York, NY: Routledge.

Lalley, J. (2017). Reliability and validity of edTPA. In J. H. Carter & H. A. Lochte (Eds.), *Teacher performance assessment and accountability reforms* (pp. 47-78). New York, NY: Palgrave Macmillan.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Leithwood, K., & Jantzi, D. (2008). Linking leadership to student learning: The contributions of leader efficacy. *Educational Administration Quarterly*, *44*, 496-528.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), 15-21.

Linquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Linquist (Ed.), *Educational measurement* (pp. 119-184). Washington, DC: American Council on Education.

Louis, K. S., Leithwood, K., Wahlstrom, K. L., & Anderson, S. E. (2010). *Investigating the links to improved student learning* (Final report of research findings). Minneapolis, MN: University of Minnesota.

May, H., & Supovitz, J. A. (2011). The scope of principal efforts to improve instruction. *Educational Administration Quarterly*, *47*, 332-352.

Miles, M., & Huberman, M. (1994). *Qualitative data analysis: A sourcebook of new methods*. Thousand Oaks, CA: SAGE.

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

Orphanos, S., & Orr, M. T. (2014). Learning leadership matters: The influence of innovative school leadership preparation on teachers' experiences and outcomes. *Educational Management, Administration & Leadership*, *42*, 680-700.

Orr, M. T., & Orphanos, S. (2011). How preparation impacts school leaders and their school improvement: Comparing exemplary and conventionally prepared principals. *Educational Administration Quarterly*, *47*, 18-70.

Orr, M. T., Pecheone, R. L., Shear, B., Hollingworth, L., & Beaudin, B. (2016). *Massachusetts Performance Assessment for Leaders (PAL) technical report: Summary of validity and reliability studies for 2014-15 field trial of PAL*. New York: Bank Street College of Education.

Orr, M. T., Pecheone, R. L., Snyder, J., Murphy, J., Palanki, A., Beaudin, B. Q., . . . Buttram, J. (2017, in press). Performance Assessment for Principal Licensure: Evidence from content and face validation and bias review. *Journal of Research on Leadership Education*.

Ozgen, S., Sanchez-Galofre, O., Alabart, J. R., Medir, M., & Giralt, F. (2013). Assessment of engineering students' leadership competencies. *Leadership and Management in Engineering*, *13*(2), 65-75.

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education. The Performance Assessment for California Teachers. *Journal of Teacher Education*, *57*, 22-36.

Pecheone, R. L., Pigg, M. J., Chung, R. R., & Souviney, R. J. (2005). Performance assessment and electronic portfolios: Their effect on teacher learning and education. *The Clearing House*, *78*, 164-176.

Pecheone, R. L., Shear, B., Whittaker, A., & Darling Hammond, L. (2013). *2013 edTPA field test: Summary report*. Palo Alto, CA: The Stanford Center for Assessment, Learning and Equity.

Pecheone, R. L., & Wei, R. C. (2007). *Technical report of the Performance Assessment for California Teachers (PACT). Summary of validity and reliability studies for the 2003-2004 pilot year*. Palo Alto, CA: The Stanford Center for Assessment, Learning and Equity.

Porter, A. C., Polikoff, M., Goldring, E., Murphy, J., Elliott, S. N., & May, H. (2010). Developing a psychometrically sound assessment of school leadership: The VAL-ED as a case study. *Educational Administration Quarterly*, *46*, 135-173.

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

Sandholtz, J. H., & Shea, L. M. (2012). Predicting performance: A comparison of university supervisors' predictions and teacher candidates' scores on a teacher performance assessment. *Journal of Teacher Education*, *63*, 39-50.

Sebring, P. B., Allensworth, E., Bryk, A. S., Easton, J. Q., & Luppescu, S. (2006). *The essential supports for school improvement*. Chicago, IL: University of Chicago Consortium of School Research.

Shavelson, R. J., & Webb, N. M. (2005). Generalizability theory. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 99-105). Cambridge, MA: Elsevier.

Shelton, S. V. (2011). *Strong leaders, strong schools: 2010 school leadership laws*. Denver, CO: National Conference of State Legislatures.

Shelton, S. V. (2012). *Preparing a pipeline of effective principals: A legislative approach*. Denver, CO: National Conference of State Legislatures.

Solansky, S. T. (2010). The evaluation of two key leadership development program components: Leadership skills assessment and leadership mentoring. *The Leadership Quarterly*, *21*, 675-681.

Supovitz, J. A., & Christman, J. B. (2005). Small learning communities that actually learn: Lessons for school leaders. *Phi Delta Kappan*, *86*, 649-651.

Tombari, M. L., & Borich, G. D. (1999). *Authentic assessment in the classroom: Applications and practice*. Upper Saddle River, NJ: Prentice Hall.

U.S. Department of Education. (2009, November). *Race to the top program executive summary*. Washington, DC: Author.

Weiss, H. B., & Stephen, N. (2009). From periphery to center: A new vision for family, school, and community partnerships. In S. Christenson & A. Reschley (Eds.), *Handbook of school-family partnerships* (pp. 448-472). New York, NY: Routledge.

Wiggins, G. P. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, CA: Jossey-Bass.

Author Biographies

Margaret Terry Orr (PhD, Columbia) is associate professor at Fordham University and previously directed a multiyear performance assessment development project for Massachusetts and advised on the national Professional Standards for Educational Leaders. She has conducted numerous regional and national studies over the last 30 years on leadership assessment and

preparation approaches and school and district reform initiatives, and published numerous books and articles.

Ray Pecheone is a professor of Practice, Stanford University. He is the founder and executive director of the Stanford Center for Assessment Learning, and Equity (SCALE), which focuses on the development of innovative performance assessments for students, teachers and administrators at the school, district and state levels.

Liz Hollingworth is the director of the Center for Evaluation and Assessment at the University of Iowa. Her research focuses on issues of leadership, program evaluation, and assessment.

Barbara Beaudin is an independent consultant who works with districts and states on measurement issues and assessment systems. She is the former associate commissioner for the Division of Assessment, Research and Technology in Connecticut.

Jon Snyder is the executive director of the Stanford Center for Opportunity Policy in Education (SCOPE), Stanford University. Snyder works at the intersection of policies, practices, and research that support the connections between educator and student opportunities for learning.

Joseph Murphy is the Frank W. Mayborn chair of Education and associate dean at Peabody College of Education of Vanderbilt University. Murphy works in the area of school improvement, with an emphasis on leadership and policy, including the development of state and national standards and assessments.