University Council for Educational Administration

Journal of Research on Leadership Education
1–30
© The University Council for Educational Administration 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1942775117701179
journals.sagepub.com/home/jrl
SAGE

Performance Assessment for Principal Licensure: Evidence From Content and Face Validation and Bias Review

Margaret Terry Orr[1], Ray Pecheone[2], Jon D. Snyder[2], Joe Murphy[3], Ameetha Palanki[4], Barbara Beaudin[5], Liz Hollingworth[6], and Joan L. Buttram[7]

Abstract

This article presents the validity bias review feedback and outcomes of new performance-based assessments to evaluate candidates seeking principal licensure. Until now, there has been little empirical work on performance assessment for principal licensure. One state recently developed a multi-task performance assessment for leaders and has implemented it for statewide use in initial principal licensure decisions; this development process is described here, focusing on content validity and bias review, and incorporates candidate and program faculty validation as well. The results demonstrate the content validity, relevance, and feasibility of this new performance assessment for leaders, and yield implications for leader assessment generally.

Keywords

certification requirements, policies, standards, certification, leadership preparation research, leadership preparation programs, policy environment for leadership preparation, principal preparation, pipeline

[1]Bank Street College of Education, New York, NY, USA
[2]Stanford University, CA, USA
[3]Vanderbilt University, Nashville, TN, USA
[4]Educopia, Santa Clara, CA, USA
[5]Independent consultant, Santa Clara, CA, USA
[6]University of Iowa, Iowa City, USA
[7]University of Delaware, Newark, USA

Corresponding Author:
Margaret Terry Orr, Bank Street College of Education, 610 W. 112th Street, New York, NY 10025, USA.
Email: morr@bankstreet.edu

Introduction

Having schools lead by high-quality school leaders begins with the licensure of qualified aspiring leaders. Determining candidate readiness for initial school leadership positions is typically the responsibility of state education licensure agencies. In recent years, many states and local districts implemented policies to strengthen the quality and effectiveness of school leaders (Augustine et al., 2009; Shelton, 2012). Among these policies is attention to the nature, form, and criteria for determining candidate readiness for initial school leader roles and responsibilities (Briggs, Cheney, Davis, & Moll, 2013; Davis et al., 2011).

Current educational leadership policies are based in part on strong research agreement that principals are second only to teachers among school effects in influencing student learning and achievement outcomes (Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010; Louis, Leithwood, Wahlstrom, & Anderson, 2010; Robinson, Lloyd, & Rowe, 2008). Large-scale research studies and meta-analyses consistently stress how leadership influences student learning by developing a vision and setting direction, improving teaching and learning, managing resources and operations, facilitating change, and engaging families and community (Bryk et al., 2010; Leithwood & Jantzi, 2008; Robinson et al., 2008). More focused research demonstrates the influence of specific leadership practices on student learning, 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). Given these findings, the challenge becomes how to determine aspiring leaders' readiness to enact these practices as new school principals.

During the early 2000s, federal and foundation initiatives (such as the federal Race to the Top grants (U.S. Department of Education, 2009), and The Wallace Foundation pipeline initiative (Turnbull, Riley, Arcaira, Anderson, & MacFarlane, 2013)) encouraged states and districts to create multiple policy levers to improve leadership preparation, selection, development, and evaluation of school leaders (Sun, 2011; Wallace Foundation, 2006). Among these policy developments have been efforts to identify better evidence of qualifications for principal licensure and readiness for initial leadership roles (Shelton, 2012).

Until recently, most states determine eligibility for an initial principal's license based on educator experience, such as a minimum number of years of teaching or pupil personnel work, and advanced graduate preparation, typically a graduate degree in school leadership (Kaye, 2016; Shelton, 2012). Many states also required candidates to complete a leadership exam, either a nationally available one, such as Praxis or School Leaders Licensure Assessment (SLLA) from the Educational Testing Service, or a state-designed one (Shelton, 2011). As a result of recent policy attention to educator quality generally, some states adopted the use of performance assessment for teacher licensure (12 states currently require EdTPA, and three others are taking steps toward this according to the American Association of Colleges for Teacher Education [AACTE] website, http://edtpa.aacte.org/state-policy), and a few have explored the means of performance assessment for principal licensure (Davis et al., 2011).

A first step in assessing evidence to determine candidate readiness for principal licensure is having valid evidence that is aligned to professional standards, relevant to a school leader's job, feasible for candidates to produce and does not incur bias in its production. One state's investment in developing valid performance assessments for initial principal licensure shows significant promise in meeting these expectations. This article describes the assessment development process, presents validity and bias-review evidence, and candidate and program director feedback on the assessments' relevance, feasibility and ease of use. The implications of the assessments for principal licensure decisions are then considered.

Background

Since 2001, the Commonwealth of Massachusetts has offered multiple leadership preparation and licensure pathways, and has committed to establishing a statewide performance assessment for principal licensure (http://www.mass.gov/edu/docs/ese/educator-effectiveness/licensing/panel-review-administrator-routes.pdf). Beginning in 2011, Massachusetts made several policy changes to leadership licensure by revising its leadership standards and indicators, requiring preparation program redesign, strengthening alternative pathways, requiring all pathways to prepare candidates according to its professional standards, and creating a principal evaluation system. All were undertaken as a means of improving principal quality and effectiveness.

Massachusetts created three pathways to principal licensure: completion of a state-approved preparation program (university-only, those sponsored by a consortium of organizations that includes regional educational agencies, or those operated by a professional association), an administrative apprenticeship/internship pathway, and a panel review process. All pathways must meet Commonwealth-defined requirements (http://www.doe.mass.edu/edprep/pr.html). In 2012, the Commonwealth required the redesign of preparation programs to align with new guidelines, and then be approved by Massachusetts Department of Elementary and Secondary Education (ESE) in 2013. The impact of this policy reduced the number and types of approved programs from 29 to 23 as shown in Table 1.

Beginning in 2012, Massachusetts added other requirements. Preparation program and apprenticeship/internship candidates were then required to complete at least 500 hours of internship experience, and demonstrate proficiency in their Professional Standards and Indicators for Administrative Leadership (http://www.doe.mass.edu/edeval/model/PartIII_AppxB.pdf). The state also allocated federal Race to the Top funding to develop, pilot, and field test a performance-based assessment system for principal licensure to ensure that all candidates seeking principal/assistant principal licensure in Massachusetts meet these state performance assessment requirements. To undertake this assessment development work, ESE staff contracted with Bank Street College faculty to form a team of leadership and psychometric experts (as an

Table 1. Number of State-Approved Preparation Programs Whose Candidates Participated in the Field Trial by Type of Program, 2013 and 2015.

Type of preparation program	Number of programs 2013	Number of programs 2015
University only	19	14
Consortium based	8	7
Professional association affiliated	2	2
Total	29	23

assessment development team) to design, implement, and validate a performance assessment system for principal licensure.

Conceptual Background

Until now, there has been little large-scale research on performance assessment for principal licensure. Most existing performance assessment measures in educational leadership focus on assessing principal practice, and lack sufficient validity and reliability (Condon & Clifford, 2010). In reviewing new state efforts to evaluate principal performance, Condon and Clifford (2010) concluded that few tools exist, indicating that "states and school districts are often challenged to determine how to measure principal performance in ways that are fair, systematic, and useful" (p. 1).

Until recently, the most commonly used assessments have been state-required licensure exams, such as the SLLA, designed and administered by the Educational Testing Service. It is a 4-hour, computer-based standardized exam that was aligned with the 2008 national educational leadership standards and, as of 2016, is required in 18 states (Educational Testing Service, 2009, n.d.). Published research on its psychometric properties is limited, and related research suggests validity and bias problems. Grissom, Mitani, and Blissett (2017), using 10 years of data for Tennessee, show that SLLA cut scores may be biased against non-White candidates and that SLLA score results do not predict subsequent principal job performance quality (using statewide evaluation, teacher feedback, and administrative data; Grissom et al., 2017).

Performance assessments based on a portfolio of authentic work show promise, and have been found in prior work to be preferable to constructed-response and multiple-choice assessments or simulated performance assessments for several reasons (Gitomer, 1993; Messick, 1994). First, performance assessments build on research of applied learning theory, whereby candidates demonstrate their knowledge and skills in performing authentic work (Kolb, 1984; Mezirow, 2000). As described by Khattri and Sweet (1996), performance assessment requires that students "perform, demonstrate, construct, and develop a product or a solution under defined conditions and standards" (p. 3). Given that participants apply new knowledge and skills to construct work products or solutions for evaluation, performance assessment has an educative function. At the same time, as an assessment, it has an evaluative function—in this case, determining professional readiness.

Second, according to Messick (1994), performance assessments have claims of being authentic and direct, and address two major threats to assessment construct validity—construct underrepresentation of the skills being assessed and construct-irrelevant variance, respectively. To achieve this, however, performance assessment design and use must address four questions: (a) Is the target of assessment the performance or the product? (b) What are the theoretical assumptions that underlie what is being tested? (c) How generalizable the findings are, based on the breadth and depth of domain coverage in the task performance samples? and (d) How transparent is what is being assessed, the criteria and standards of good performance, scoring, and the assessment's relevance to improve performance? As well, as Messick (1994) acknowledged, it is challenging to disentangle component skills used in complex tasks.

Various psychometric experts (Linn, Baker, & Dunbar, 1991; Messick, 1994, 2005; Stiggins, 1987) and professional research associations (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) provide guidance to address key assessment validity and reliability considerations and challenges. Important among these is that assessments are to use recognized standards to provide transparency on what is being assessed and clarify the conceptual foundation. Thus, they are dependent upon the quality of the standards used and must demonstrate that the assessment content is valid reflection of the work being performed.

Designing and implementing valid performance assessments is a complex endeavor (Martineau, 2004; Stiggins, 1987; Turnbull et al., 2013; Wei & Pecheone, 2010). It begins with a design process that yields assessment content with strong content validity. Typically, the assessment content is created through a multistep psychometric process, which includes a review of standards and research, job and field analysis, and an iterative content and design generation process. Next, the task description and work product instructions, related materials, and rubrics are developed; the tasks and assessment system are pilot tested for feasibility and use; and the assessment system is field tested to ensure that it is valid, reliable, and bias free. Third, the assessment system itself is developed and evaluated, including the assessment participation instructions, scoring guides, scorer training, assessment management system, and supporting documents. Finally, a strategic set of communications is created to prepare the field—in this case, higher education and K-12 education fields—to support implementation and use of the new assessments with fidelity (Lane, 2010).

The first consideration is determining the leadership skills to be addressed in the assessments. Designing conceptually valid performance assessments for leadership is challenged by the complexity of leadership work and the multiple dimensions of leadership practice that must be evaluated within a body of evidence. Duckor and others (2014) addressed this challenge in teacher performance assessments, noting that a conceptual framework for a licensure assessment may be focused on a single dimension of behavior or several distinct components that are thought to be interrelated and homogeneous. They investigated the validity assumptions about a multiscore-based teacher licensure performance assessment—Performance Assessment for California Teachers (PACT)—to determine whether there is "empirical evidence for increasing task difficulty across the scales" (p. 403) and whether subsets of assessment tasks "function differently for groups of similarly scoring examinees" (p. 403). They confirmed its content validity, by concluding that they "found a sufficient degree of internal structure validity evidence to support the continued use of the PACT instrument as intended to measure California teacher candidates' 'skills and abilities' in accordance with the state's professional standards in teaching" (p. 413). They did not confirm the content validity argument for a five-domain-based factor structure, and point to the need for more construct definition work and further instrument development, demonstrating challenges in performance assessment development.

Early implementation experiences with teacher education performance assessments show that the operational aspects of the assessment system (its design and ease of use) and candidate support (through assessment materials and resources and preparation program guidance) could potentially mediate candidate performance and scores (Lit & Lotan, 2013; Pecheone & Chung, 2006; Young, Cruess, Cruess, & Steinert, 2014). These experiences underscore the importance of establishing feasibility and ease of use, before using assessments evaluatively. The validation process of a performance assessment system helps uncover whether and how the assessment system itself might independently affect candidate performance.

Method

Given these considerations, the Performance Assessment for Leaders (PAL) assessment design and implementation process repeatedly evaluated the content validity ("How well does the content of PAL represent core domains of school leadership knowledge and skills?" "Are the right skills being assessed?" and "Is job-appropriate content gathered to evaluate these skills?"), potential bias ("Do the tasks and their use create any bias among candidates?"), and the feasibility of the assessments ("Are the tasks feasible to complete?" "Do the tasks provide appropriate challenge, particularly with supports?"). The content validity was determined by having a group of experts (as a design committee) assess the PAL assessment tasks alignment to the state leadership standards, having their job relevance assessed by a group of experts (as a content validity committee), and having candidates and preparation program faculty provide face validity (through feedback surveys). The potential bias was evaluated by an expert bias-review committee. The assessments' feasibility was evaluated through candidate and preparation program faculty survey feedback about feasibility, ease of use, and challenge in completing task steps.

Sample

Three types of participant samples provided input into the PAL validation and feasibility assessment. Representatives from several Massachusetts preparation programs and pathways, and K-12 school and district leaders served on one of three committees (10 members each): design, content, or bias review. Participating candidates and preparation programs completed online surveys following the pilot study and field trials.

Members of the design committee and content validity committee reviewed the four draft tasks and the assessment system before these were piloted to determine their importance and relevance in relationship to state and national leadership standards, the research literature on effective school leadership, and the committee members' knowledge of the job of new leaders. The design committee also reviewed and confirmed the findings on the tasks and standards alignment. The two committees reviewed the PAL tasks and assessment system after the pilot study, and made revisions before the field trial was launched in September 2014, and again after the field trial, before the first full implementation of the PAL assessment was launched in Program Year 2015-2016. Their feedback was documented in meeting notes and used to inform revisions. The bias-review committee met before the pilot study and after the field trial to evaluate the potential bias of the tasks' content and the performance results of the field trial assessment. Each member completed a bias-review rating form to evaluate each task, and their feedback was documented in meeting notes.

Follow-Up Surveys on Face Validity and Feasibility

Pilot study and field trial participating candidates and program faculty members provided face validation to answer the same content validation question about task alignment with the standards and the appropriateness and relevance of the assessments to initial school leader work, from their vantage points (not as experts), as complementary to the content validation.[1] At the end of the pilot study (May and June 2014) and field trial (May and June 2015), participating candidates and program faculty were asked to complete an online survey (with three or more follow-up requests).

Pilot study respondents comprised of most candidates who submitted work products and all directors of programs whose candidates participated in the pilot study. Respondent information, as shown in Table 3, shows that 35 candidates provided feedback (to calculate the response rate, we used the 58 ratings of individual tasks on the feedback survey, which is 43% of the 134 task submissions). Based on these results, it was concluded that the feedback survey results are representative of typical candidates. Most survey respondents were currently classroom teachers or nonteaching professional staff. A few were instructional specialists or department chairs. Some candidates completed two tasks (most typically Tasks 1 and 4) and were included in both task counts.

Field trial completers (those who had completed all four tasks during the field trial period) were asked to complete a similar online feedback survey with questions about the tasks' relevance to an initial school leader job and alignment to the standards. Of the 416 candidates who completed all four tasks, 92 completed the feedback survey, representing 22% of the completers. Table 3 shows that most survey respondents were female, White, and from a preparation program (although only some chose to provide this information). These percentages are similar to the demographic information for the total field trial sample (data not shown). Table 2 shows that the majority of field trial candidate survey respondents are currently in a nonsupervisory position: classroom teacher, professional support staff, instructional specialist, or other professional

Table 2. Percentage Distribution of Pilot Study and Field Trial Candidate Survey Respondents by Selected Demographic Characteristics.

Demographic	% pilot study respondents (*n* = 28-31)	% field trial respondents (*n* = 92)
Gender		
Male	29	24
Female	71	76
Total		100
Race/ethnicity		
African American		5
American Indian/Alaskan Native		0
Asian		4
Hispanic/Latino/a	10	2
White (not Hispanic)	90	77
I would prefer not to answer		12
Program/pathway		
University-based leadership preparation program	100	67
A consortium or association-based leadership preparation program		14
Administrative apprenticeship/ internship option		16
Panel review option		4

nonsupervisory staff. These results are similar to the pilot study candidate response sample. Table 3 shows that field trial candidates were more likely to have completed their program at the time of the assessments than were pilot study candidates. No information exists on aspiring candidates in Massachusetts from the various programs and pathways to determine how well this response sample reflects the aspiring candidates in Massachusetts.

At the end of the pilot study, preparation program directors with candidates who completed one or more tasks were asked to complete a feedback survey. Four responded, representing 45% of the nine programs that participated in the pilot study. At the end of the field trial, all program directors and faculty who were the primary contacts for information about PAL were surveyed. Due to the anonymity of the survey, program representation could not be assessed. Fifteen faculty responded and 10 to 11 (depending upon the question) provided feedback, representing 48% of the 23 programs that had field trial candidates. Some responding faculty members were only familiar with some tasks, as noted by the response patterns to questions about each task (respondents were asked first if they were familiar with a task, and only those who reported yes were given the follow-up questions). As shown in Table 4, most were program directors, and others were faculty or instructors with their programs.

Table 3. Percentage of Pilot Study and Field Trial Candidate Survey Respondents by Field Work/Internship and Program Completion Status.

Candidate status	% pilot (*n* = 35)	% (*n* = 92)
Field work status		
Have not yet begun a school leadership internship or field experience	14	11
I am currently participating in a school leadership internship/field experience	83	21
Have completed a school leadership internship or field experience	3	63
Other		5
Total	100	100
Have already completed my program		60
Will complete this summer		4
Will complete this fall		12
Will complete next spring		16
Will complete after next spring		8

PAL Assessment Design

The PAL assessment system was developed and refined through a standards-based design process to ensure content validity and alignment to the state standards and expectations for beginning school leaders, similar to the one used for PACT and edTPA (formerly the Teacher Performance Assessment) (Pecheone, Shear, & Darling Hammond, 2013; Pecheone & Wei, 2007). Through a 2-year planning and design process, the assessment development team worked closely with a design committee (comprised of 10 representatives from K-12 schools and district leadership and preparation pathways) and ESE staff. During the design process, the assessment development team, ESE, and the design committee examined the core work of principals, current research, professional standards, and expectations for leadership preparation.[2] Upon analyzing the themes that arose from this examination, the team distilled four core tasks that would yield actionable and observable candidate performance across multiple standards. Collaborating further with the design committee, the assessment development team created work product instructions and rubrics to assess candidate performance on the tasks. This work became the PAL assessment.

PAL consists of four performance assessment tasks of leadership knowledge and skills. The tasks ask licensure candidates to set direction by developing a plan for an area of school improvement, creating a professional learning culture among school staff, supporting individual teacher development through observation and feedback, and engaging families and community in improving student learning. Specifically, the four tasks comprise the following: 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;

Table 4. Number of Respondents to the Pilot Study and Field Trial Program Director Feedback Surveys by Role.

Role	Number of responses	%
Number of faculty members completing the pilot survey	4	
Number of faculty members completing the field trial survey	10	
Program director	7	70
University faculty member	1	10
Course instructor	2	20
Other (specify)	0	0

and Task 4: Leadership for family engagement and community involvement. Each task (as outlined in Appendix A) is divided into four components of leadership action that reflect the cycle of leadership inquiry and learning, as shown in Figure 1.

The PAL assessment system is designed to provide both clear evidence of a candidate's readiness for an initial school leadership position and data for preparation programs on a candidates' performance. It builds on the new Commonwealth regulations for preparation program approval, and is aligned with other Commonwealth leadership development efforts to support and evaluate principals and assistant principals. Candidates' performance on this new assessment should inform licensure decisions, while also serving an educative purpose for candidates, preparation programs, and policy makers. PAL is designed as a summative assessment of a candidate's key leadership knowledge and skills. It is standardized across all programs and pathways to school leader licensure.

Findings

The findings are presented in three parts: content validation, bias review, and feasibility.

Content Validation

The content validation results are presented in three parts: standards alignment, content validation by the design and content validation committees, and face validation by participating candidates and preparation program faculty.

Standards alignment. In the initial assessment design phase, completed in 2012-2013, the assessment development team reviewed state and national leadership standards and research on leadership effectiveness, and developed the tasks through collaboration among the project staff, ESE staff, and the design committee. Together, these groups confirmed (through discussion and formal committee member affirmation) the

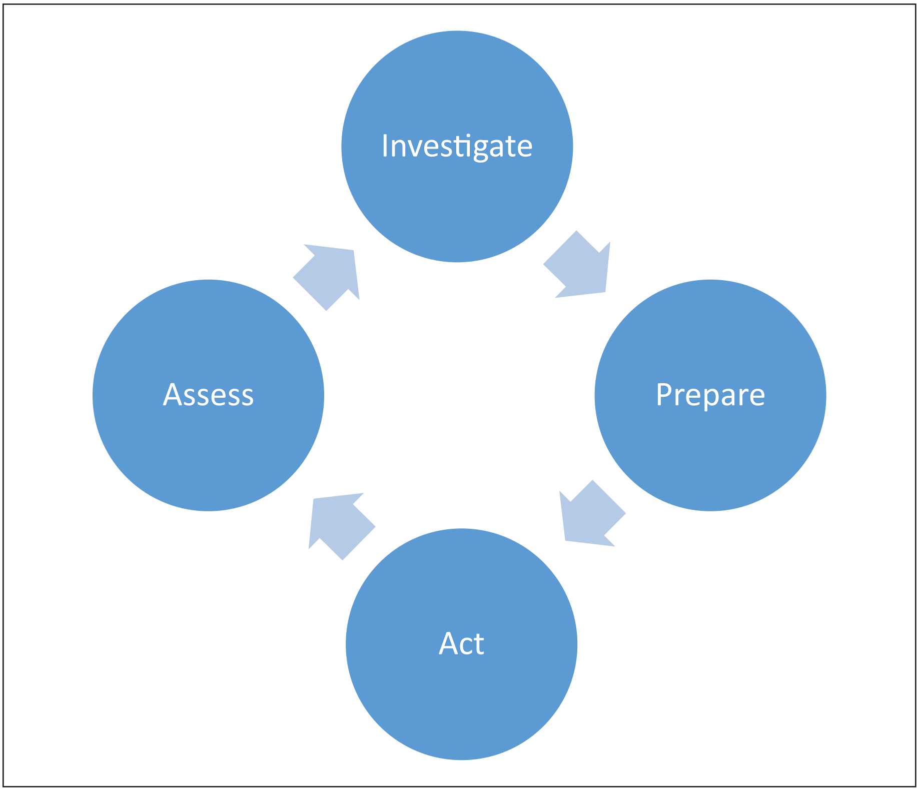

Figure 1. The components of leadership action.

extent of alignment of the four PAL assessment tasks to the Massachusetts leadership standards, as shown in Appendix B. The four PAL tasks reflect three of the four Massachusetts leadership standards strongly and some indicators of the fourth standard weakly. Many indicators from the four standards are reflected in one or more tasks. The other indicators were beyond the construct measure in the four tasks, and were determined by the design committee to be too complex to measure through a performance assessment.

Design committee validation. Once drafted and revised, but before piloting, the four PAL tasks were reviewed by the 10-member design committee for relevance and feasibility as initial school leader assessments. Through a day-long review and discussion, committee members agreed strongly with the focus and nature of the assessment tasks, their relevance for initial school leaders' work, and their feasibility. Their primary recommendations for changing the tasks focused on the instructions to strengthen the clarity of directions, and explore how to gain local school and district support for candidates to complete the work.

Content committee validation. A 10-member content validity committee met in spring 2013, prior to the pilot study launch for a day-long meeting. The committee was trained in content validation, and then individually and collectively rated the content validity of each task on a 5-point scale for job relevance, authenticity, and importance. They answered questions on how important the knowledge and skills assessed in the task are for performing the job of an entry-school principal and for improving student learning; how well the set of components and products required for the task reflect the authentic work that an entry-level principal must perform on the job; and how frequently an entry-level principal demonstrates the knowledge and skills in the work products required for the task while on the job.

To evaluate the committee's ratings, we used Wilson, Pan, and Schumsky's (2012) criteria that half or more of the committee members must agree or strongly agree on the task attributes for the content to be valid. Their recommended critical values of content validation scores (which they calculate as an average of the ratio of the number of panelists agreeing that each item is essential among all panelists) for 10 panelists and nine panelists are 62% and 78%, respectively. While the authors (Wilson et al., 2012) applied the critical values criteria to the combined ratings for a given test, we applied the criteria to the percentage of committee members rating each task and item on importance and relevance.

Table 5 shows that most content validity committee members rated the tasks important or very important for performing the job of an entry-level principal and improving student learning; all mean rating scores were 4.0 to 4.8 on a 5-point scale. Most committee members also agreed that the components and artifacts required for the task reflected authentic work that beginning principals need to do on the job. Those who rated Tasks 1 and 2 as being performed less frequently by an entry-level school leader than did the majority of others and rated Task 1-related knowledge and skills as less important for improving student learning explained that they did not expect new leaders to be able to take on this type of work as intensively in their first year.

In evaluating the PAL assessment system as a whole, almost all committee members present (89%) rated the four tasks and sets of products as "well" or "very well" in reflecting the authentic work that an entry-level principal must perform on the job. Most (89%) also agreed that an entry-level principal would need to frequently demonstrate the knowledge and skills in the work products required for the combination of four tasks and work products of PAL assessment system while on the job. Taken these results together, we concluded that because the PAL content validity ratings were equal to or greater than the Wilson and others (2012) recommended values, the PAL assessments demonstrated strong content validity.

Face validation. To complement the content validity, we collected face validity information from a broader audience by including questions about the tasks' job relevance and standards alignment in pilot study and field trial feedback surveys from participating candidates and supervising preparation program faculty. Face validity is distinct from content validity, in that it is more subjective and is reported by those who participate in the assessment (or support those who do). According to Holden (2010), face

Table 5. Percentage of Content Validity Committee Members Rating the Task as Important or Very Important (or Well or Frequent), by Task.

Criteria	Task 1	Task 2	Task 3	Task 4
Number of committee members	10	10	8	9
How important the knowledge and skills assessed in the task are for performing the job of an entry-school principal	100	90	100	100
How important the knowledge and skills assessed in the task are for improving student learning	70	100	100	89
How well the set of components and products required for the task reflect the authentic work that an entry-level principal must perform on the job	80	70	88	100
How frequently an entry-level principal demonstrates the knowledge and skills in the work products required for the task while on the job	80	80	88	100

Note. The number of committee members varied by task, as not everyone was able to participate throughout the whole day.

validation, unlike content validation, assesses the "appropriateness, sensibility or relevance of the test and its items as they appear to the persons answering the test" (p. 637) and is positively associated with other forms of technical validity (Holden, 2010). Candidates and program faculty can confirm face validity because they have some knowledge of the leadership skills and knowledge being assessed and the nature of initial school leader work but are not trained in content validation. To determine PAL face validity, surveyed pilot study candidates rated how strongly they agree that the tasks provide authentic experiences and are relevant to their preparation. As shown in Table 6, most responding candidates agreed that the tasks are aligned to the Massachusetts Standards for Administrative Leadership, were complementary to their leadership preparation, provided authentic job-related experiences, and were relevant and essential to the work that successful school leaders must be able to do.

The same questions were repeated in the field trial feedback survey and yielded similar responses, as shown in Table 7, with some differences by task. They were more likely than the pilot study candidates to agree on these attributes for Task 2, and were similar in their agreement or somewhat better in their agreement for Task 4 for two of the four attributes. Of the four attributes, most candidates agreed that Task 3 was complementary to their leadership preparation, two thirds agreed that Tasks 2 and 4 were complementary, and just more than half (56%) agreed that Task 1 was complementary to their preparation. These percentages were somewhat lower than those reported by pilot study candidates who were reporting on only one task and were primarily from one program (which had selected the task their candidates piloted). Thus, the pilot study candidates' programs might have been more closely aligned to the task they completed than were the field trial candidates who were from multiple preparation programs and pathways, and thus more likely to have somewhat different

Table 6. Percentage of Responding Pilot Study Candidates Who Agree About Selected Qualities of the Assessments, by Task.

Qualities	Task 3 Fall Pilot	Task 1 Spring Pilot	Task 2 Spring Pilot	Task 3 Spring Pilot	Task 4 Spring Pilot
Number of respondents as % of work product submissions	68	50	100	100	55
The tasks provide candidates with authentic job-related experiences	69	100	71	100	92
The tasks are relevant and essential to the work that successful school leaders must be able to do	85	89	57	100	82
The tasks are aligned to the Massachusetts standards	77	90	86	100	92
The tasks were complementary to my leadership preparation	77	90	71	100	91
Number of candidates	13	12	10	8	15

Note. A few candidates completed two tasks.

preparations. Because of the small sample size and modest survey response rate, these findings should be interpreted cautiously and not be overgeneralized.

The pilot study and field trial yielded similarly positive face validity feedback from program faculty. In the pilot study (data not shown), three faculty members agreed that the tasks (Tasks 2 and 3, on which they provided feedback) completed by their candidates were clearly aligned to the Massachusetts Standards for Administrative Leadership, provided candidates with authentic job-related experiences, and were relevant to the work that successful school leaders must be able to do. They also all agreed that the tasks were aligned to their programs' curriculum. The responses were even more positive after the field trial. As shown in Table 8, most responding program directors and other faculty confirmed the face validity of the tasks and rated the tasks highly—agreeing or strongly agreeing—that the tasks are aligned to the state standards, provide authentic job-related experiences, and are relevant to the work of school leaders.

PAL Bias and Sensitivity

To evaluate the assessment tasks for possible bias and sensitivity, we followed Educational Testing Service assessment review guidelines and Popham's (2012) booklet on removing assessment bias (Popham, 2012). Working in collaboration with ESE staff, the assessment development team convened a nine-member PAL bias-review committee (of higher education and educational leaders with expertise on bias review) in spring 2013 before the pilot study. The committee was trained in the core

Table 7. Percentage of Field Trial Candidates Who Agree or Strongly Agree With Task Attributes Related to Face Validity (4-Point Agreement Scale; *n* = 92).

	% agree or strongly agree			
Attribute	Task 1	Task 2	Task 3	Task 4
The task provides candidates with authentic job-related experiences	71	76	81	73
The task is relevant and essential to the work that successful school leaders must be able to do	75	77	89	81
The task is aligned to the Massachusetts standards	84	82	89	85
The task was complementary to my leadership preparation	56	71	82	69

concepts of bias and sensitivity, and means of assessing it using a bias-related evaluation forms for each task that addresses types of bias related to content, language, offense, stereotypes, and fairness (see Appendix C for a summary of the questions). Following the training, the committee members evaluated each task independently and collectively, and submitted their ratings and written feedback about potential bias or insensitivity for each task on the five indicators. An analysis of the results showed that committee members' bias and sensitivity review found little evidence of either bias or insensitivity, except to identify terms that could be misleading and were edited for the pilot study.

Following the field trial, the committee was reconvened to review the score performance of participating candidates, based on race/ethnicity, gender, and preparation pathway. The results were either inconclusive because of insufficient information (many candidates did not voluntarily identify their race/ethnicity), or no discernible pattern existed among task scores based on gender and pathway. The committee reviewed the tasks and instructions again for possible bias issues and found none. The committee, while finding no identifiable instances of bias, recommended that subgroup differences continue to be monitored in the future.

Ease of Use and Feasibility

The bias-review committee raised questions about feasibility and ease of use, and these issues were monitored through candidate and preparation program feedback after the pilot study and field trial. The bias committee members' first concern was that there might be some challenges that would unfairly hamper the ability of some candidates to complete the work, based on school or district policies and conditions. Their second concern was that the assessment system and its related materials might present technical challenges that could inadvertently limit a candidate's performance, and that some candidates might have less access than others to information and support to complete the tasks.

Table 8. Percentage of Responding Program Directors Who Agree or Strongly Agree With the Attributes of Each Task, Field Trial.

Attribute	Task 1	Task 2	Task 3	Task 4
The task that my candidates completed is clearly aligned to the Massachusetts standards	[a]	100	100	100
The task provided candidates with authentic job-related experiences	90	88	88	89
The task is relevant to the work that successful school leaders must be able to do	90	100	100	100
Number of responses	10	9	8	9

[a] This question was not asked for Task 1.

To address these concerns, ESE and the assessment development staff took several steps to explain PAL and the task requirements to the educational professionals most likely to be involved, by developing communication materials for preparation programs to use with school and district leaders, to explain each task requirement and the PAL policy generally. ESE staff disseminated information on the PAL policy and task requirements generally to school and district leaders, and discussed communication about PAL with the state professional associations.

As part of pilot study, candidates and program directors provided feedback on the ease of use and feasibility of each task. The results, as shown in Table 9, were mixed about how easy it was to use the resources and technology. While responding candidates agreed that the website was easy to use and the majority found the *Candidates Assessment Handbook* and rubrics easy to understand, some did not, and this finding varied by task. Only a few Spring Pilot Task 2 and 3 candidates thought that the *Handbook* was easy to understand, in contrast with almost all Fall Pilot Task 3 candidates. Less than half the candidates agreed that the instructions (in the *Handbook*) were clearly written, including none of the Task 3 Spring Pilot candidates.

To some extent, this problem over the instructions was purposeful. The design committee wanted to see variation in work production as part of the pilot study, hence agreed to provide little specificity on the format and detail for task submissions in the *Handbook* instructions. This problem was addressed in the field trial, when the task descriptions, *Handbook* instructions, and rubrics were revised to add more clarity and direction, and provide more specifics about work product expectations, including format and length requirements for work products and outlines for plans, reports, and feedback attributes.

In terms of feasibility, the pilot candidates were mixed in their agreement about whether the tasks were flexible and adaptable to different settings, and required a realistic amount of work. Here, they were most strongly positive about Task 3 and least strongly positive about Task 2. The various advisory committees had been

Table 9. Percentage of Responding Pilot Study Candidates Who Agree That the PAL System and Resources Were Easy to Use and That Completing the Task Was Feasible.

Attributes	Task 3 Fall Pilot	Task 1 Spring Pilot	Task 2, Spring Pilot	Task 3 Spring Pilot	Task 4 Spring Pilot
Ease of use					
The *Candidates Assessment Handbook* was easy to understand	85	60	28	14	50
The PAL website was easy to use	77	80	86	57	83
The rubrics helped me to understand the scoring criteria and standards used to evaluate the work products	69	50	57	71	58
The instructions for the tasks and work products were clearly written	23	40	43	0	33
Feasibility					
The tasks are flexible and adaptable so candidates in different types of school settings can structure meaningful activities and produce relevant products	62	40	43	57	42
I felt prepared to collect information on student and school culture	69	33	43	86	88
Completing the tasks required a realistic amount of work	69	67	29	43	45
Ease of use of ShowEvidence Information Management System Elements					
Instructions	62	70	57	71	75
Enrollment	69	90	86	86	92
Uploading documents	85	90	86	71	92
Number of pilot study candidates	13	12	10	8	15

Note. The responses of candidates who completed both Tasks 1 and 4 were included in the reports for Tasks 1 and 4.

concerned that Task 3 would present challenges because it required video recording a teacher giving instruction, but this feedback proved otherwise. The assessment development team and design committee agreed that improved work product instructions would make the tasks easier to perform, and thus improve perceptions of task feasibility.

Table 10. Percentage of Responding Field Trial Candidates Who Agree That the Task Was Flexible and Adaptable, by Task (*n* = 92).

	% agree or strongly agree			
Attribute	Task 1	Task 2	Task 3	Task 4
I felt prepared to collect information on student and school culture (student, teacher, and other stakeholder culture; and climate surveys, focus groups, and interviews)	66	48	67	49
Completing the task required a realistic amount of work	29	76	81	73
The task is flexible and adaptable, so candidates in different types of school settings can structure meaningful activities and produce relevant products	47	57	70	64

Ease of use for the information management system, ShowEvidence, was separately evaluated and generally positive, as shown in Table 9. In written feedback, a few candidates stated that they would like more clarity on how to combine and upload documents, and complained about the two-step system enrollment process (which was not required for the field trial), the time required for uploading videos (which may be related to videos that exceeded the time requirement), and the desire for an electronic notification that all their materials had been successfully submitted (which was done). These challenges did not appear to hinder candidate performance on the tasks and were addressed in the improvements for the field trial.

As part of the field trial, we followed up with candidates to determine whether the changes in task descriptions and instructions improved perceptions of PAL's ease of use and feasibility, particularly with respect to completion in different settings. As shown in Table 10, the majority of the candidates agreed that the tasks were flexible and adaptable for different school settings. Their agreement was lowest for Task 1 and highest for Task 3. The most common reason given for the lower Task 1 ratings was that this task was not adaptable to setting-related differences, such as those where there was limited availability of state assessment data. The video-recording requirement of Task 3 did not present significant challenges, a concern that had been raised by the committee as a possible feasibility problem if districts did not permit candidates to video record a teacher observation. With the exception of Task 1, most candidates agreed that the tasks required a realistic amount of work. They varied in rating how well prepared they were to collect task-related information on student and school culture, but this variation did not appear to be systemically related to different settings or candidate demographic attribute.

For further consideration of feasibility, field trial candidates were asked to rate the difficulty of completing the steps from each of the four tasks, using a 5-point Difficulty scale. As shown in Table 11, the task demands on candidates appear to be appropriately challenging. About two thirds of candidates reported that the tasks were not difficult, rating them as neither difficult nor easy, or rating them as easy or very easy. The steps in Tasks 1 and 3 were generally somewhat less difficult than those in Tasks 2 and 4. The least difficult step appeared to be the candidates' ability to assess their own

Table 11. Percentage of Responding Field Trial Candidates Who Reported That Selected Task-Specific Requirements Were Not Difficult (Rating Them Neutral to Very Easy) by Requirement and Task (5-Point Scale, *Very Difficult* to *Very Easy; n* = 92).

	% who rated the difficulty as very easy to neutral			
Task requirement	Task 1	Task 2	Task 3	Task 4
Task 1				
Solicit input from students, teachers, families, and other stakeholders	59.8			
Analyze relevant school and community data	71.3			
Identify a priority area	81.6			
Plan for improving school or teacher practice in a priority academic area	63.2			
Develop improvement strategies	70.1			
Task 2				
Consistently facilitate a teacher group's learning in a focus area over time		62.5		
Support individual teachers and a teacher group on improving curriculum, instruction, or assessment as a professional learning group		67.5		
Collecting and analyzing teacher feedback on group facilitation and group learning		65.0		
Task 3				
Conduct a preobservation conference			73.3	
Document a teacher observation using a district or state guide on effective teaching practices			71.6	
Conduct a postobservation conference that facilitates teacher rapport and learning			68.9	
Supports an observed teacher by providing constructive feedback and strategies for improvement			74.3	
Collects and analyzes teacher feedback on the effectiveness of the observation, feedback, and support			71.6	
Task 4				
Identifies a priority area for improving family and community engagement that would directly or indirectly enhance student learning in a priority area				68.0
Creates a multistrategy plan on how to improve family and community engagement in support of student learning priority area				62.7
Implements one planned strategy				60.0
Gathers and analyzes feedback and other evidence on the plan and strategy's effectiveness for improving family and community engagement				57.3
Assess your leadership skills in completing task	80.2	77.5	75.1	74.7

Table 12. Percentage of Responding Field Trial Faculty Who Agree or Strongly Agree About Ease of Use, Feasibility, and Program-Related Attributes, by Task.

Attribute	Task 1	Task 2	Task 3	Task 4
The task is flexible and adaptable enough, so that candidates in different types of school settings are able to structure meaningful activities and produce relevant products	70	75	50	67
It is feasible for candidates to complete the task within the structure of a course or internship that my institution offers	90	75	63	89
The task is aligned to the curriculum of the program that my institution offers to prepare new school leaders	100	100	100	100
The process of supporting candidates in completing this task has been a catalyst for rethinking how we prepare school leaders	70	63	75	67
Number of responses	10	9	8	9

leadership skills. According to candidates, the most difficult requirements were those related to direct action: facilitating a group of teachers, providing feedback to teachers, and implementing a family engagement strategy. Collecting input and feedback from others was also more challenging than developing plans. These differences by type of action seem consistent with the demands of the task requirements: Preparing and inviting critique was generally less challenging than doing.

The issues related to feasibility and ease of use were also explored with preparation program faculty who provided candidate support. The three faculty members who responded to the pilot study survey agreed that they understood the task requirements, but were in less accord about how well they understood the scoring criteria and standards used to evaluate the work products. The primary faculty member suggestion for improving the assessment management system was to make sure that the directions were consistent, particularly on how to bundle various documents for uploading. This feedback was used to improve instructions and rubric language for the field trial.

We followed up on the same questions after the field trial. More preparation program faculty responded, and the majority were positive about the tasks' feasibility and the assessment system's ease of use, underscoring the relationship between the assessments and their preparation programs. Most program directors (63%-90%, depending on the task), in reflecting on their candidates' field trial experience, agreed that the tasks were feasible for candidates to complete within the structure of a course or internship; nearly all agreed for Tasks 1 and 4, and the majority agreed for Tasks 2 and 3 (see Table 12). The majority (50%-75%) agreed that the tasks were flexible and adaptable enough for candidates in different school settings. They were least strong in their agreement when rating Task 3—citing concern about district or school staff cooperation with the video-recording requirement of the task. They all agreed (100%) that the tasks aligned with their programs' curriculum, and the majority (63%-75%) agreed that supporting their candidates in completing the tasks had been a catalyst for their program's work on preparation, particularly for Task 3.

Discussion

The content validity of the PAL assessments was scrutinized in three ways: through an evaluation of leadership standards alignment, a formal content validation study completed by K-12 school and district leaders and higher education faculty, and two rounds of face validation by participating leadership candidates and preparation program faculty. Through each process, the results were consistently strong (including Wilson and others', 2012, content validity strength measure), and in agreement with the fact that the PAL assessment tasks are well aligned to the state leadership standards, provide authentic job-related experiences, and are relevant to the work of school leaders. The modest survey response rates provide some caution to the face validation, but it is likely that the survey would have captured candidate and faculty criticism, to the extent that any existed, rather than systematically excluded it.

Based on available evidence from the bias-review committee, the PAL assessments do not, in their design and implementation, present potential threats to bias and sensitivity for the candidates. Candidate and faculty assessments of feasibility and ease of use were generally positive and improved between the pilot study and field trial, with the addition of clarifications for the instructions and rubrics. Importantly, the candidates reported that the task work, generally, flexible and adaptable to different types of settings, and the video-recording requirement for Task 3 was not problematic. Questions about feasibility and ease of use were balanced against the degree of assessment challenge. The results are quite positive, showing that while the assessments have generally positive candidate and faculty ratings for feasibility and ease of use, the candidates were well distributed on ratings on assessment challenge, showing these assessments to be somewhat, but not overly, challenging, with some differences by task.

Conclusions and Implications

The content validation results for the PAL assessments are quite positive, demonstrating the feasibility of creating performance assessments for principal licensure that adhere to professional guidelines for psychometric assessment development (American Educational Research Association et al., 2014). These assessments meet the first three assessment development criteria: (a) they are well aligned with state leadership standards and reflect the job of school leaders, particularly for improving student learning; (b) the tasks address multiple, interrelated skill domains and are somewhat challenging for candidates to perform; and (c) the assessment system has positive ratings on ease of use and the tasks are generally feasible to complete, with some differences in skill complexity by task component. In addition, the assessment system itself was strengthened sufficiently to minimize its independent influence on candidate performance, an important attribute for assessment validation.

While the assessments were designed to meet Massachusetts standards and leadership assessment needs, the results show that the assessments have promise for other states and localities. Specifically, the candidates agreed that the tasks were applicable to a wide variety of schools and settings, which, for Massachusetts, range from rural to urban school districts, as they do in other states. Thus, with the growing need for better assessments to determine leadership candidate readiness, and the lack of valid evidence around other existing assessments such as the SLLA (Grissom et al., 2017), these assessments are worth replicating elsewhere.

Appendix A

Task 1

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.

Candidates prepare three artifacts to demonstrate their work:

1. A three-page memo that describes the priority area and context, and presents qualitative and quantitative data analyses, a rationale for the priority area selected, an analysis of existing programs and services, and input from others.
2. A four-page plan that outlines set of goals, objectives, and action strategies to improve learning in the priority area for the targeted student group, and a theory of action that describes how these strategies will lead to improved student performance.
3. A three-page report that describes how they collected feedback from school leaders, the leadership team, and other stakeholders about the proposed plan, and synthesized and interpreted the feedback.

Candidates must also write a two-page commentary that evaluates the leadership skills used in developing the plan and in soliciting and using feedback to revise it. They must assess how they would improve their leadership skills and practices. Finally, candidates must provide a series of documents with summary quantitative and qualitative data, data collection forms, and school mission and vision statements.

Task 2

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.

To demonstrate their work in performing this task, candidates prepare three artifacts:

1. A two-page memo that explains the academic priority focus area, identifies the group of teachers that will address it, presents a plan for how they will work together as a professional learning group.
2. A five-page report that summarizes what the group did over the course of its meetings; the role the candidate had in fostering the teachers' learning individually and collectively; and the challenges of, and strategies for, working together.
3. A three-page memo that provides an analysis of the group members' feedback on group learning, group task accomplishment(s), the candidate's facilitation role, and evidence of the benefits of the work for improving teaching practice and student learning.

In a two-page commentary, candidates evaluate their learning and leadership development through this experience by drawing on the activities and feedback received from group members about how they also influenced their professional learning. Candidates provide documentation on the group membership, group norms, agendas and minutes, and feedback forms used.

Task 3

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.

To complete this task, candidates prepare and submit five artifacts:

1. A preobservation template about the teacher and class to be observed with a summary of the preobservation meeting.
2. A 15-min video recording of the observed teacher.
3. A 15-min video recording of the postobservation meeting between the candidate and the observed teacher.
4. A two-page memo to the observed teacher providing summary documentation and analysis of the observed teaching using the district's effective teaching rubric or protocol.

5. A two-page memo to the teacher analyzing the teacher's feedback about the preobservation meeting, observation, and postobservation meeting, and the implications of the feedback received for the teacher's work and student learning.

Candidates submit several documents for this task, including the district's observation rubric, evidence about the lesson under observation, and relevant student and teacher information. Furthermore, they provide a two-page personal analysis as a commentary that evaluates the leadership skills used in the task, how these benefited the observed teacher, and implications from completing the task for improving leadership skills.

Task 4

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 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.

Candidates prepare three artifacts to demonstrate their work:

1. A five-page proposed plan to improve or increase family and community involvement that will directly or indirectly improve student learning. The plan must include a description of the priority area and existing policies, practices, and programs to engage family and community to address this area; members of the working group for planning and a well-defined plan that lays out goals and objectives; a theory of action; two or more improvement strategies; the resources, roles, and responsibilities for the strategies; a timeline; and a proposed evaluation process.
2. A three-page memo describing the implementation of one proposed strategy, with detailed steps; a description of participation and involvement; an analysis of strengths and weaknesses; and identification of benefits.
3. A three-page report analyzing feedback from family and community members, school leaders, and staff about the plan and implemented strategy, and their implications for improving family and community engagement and addressing the priority area.

Candidates must include supporting documents with data about the priority area and family and community engagement, and conclude with a two-page commentary about the leadership skills developed in completing the task, what was most effective, and what could be improved.

Appendix B

Massachusetts leadership standards/elements	Task 1 vision/ direction	Task 2 professional school culture	Task 3 individual teacher's effectiveness	Task 4 family and community engagement
Standard 1				
Instructional leadership	***/**/*	***/**	***/**/*	**/*
Goals	***	**		*
Aligned curriculum	***	**		
Instruction	***	**	***	
Assessment	***	***	***	
Evaluation	*	**	***	
Data-informed decision making	***	**	***	**
Equity and excellence	***	***	**	**
Accountability	*	***	***	
Closing proficiency gaps	***	***	**	*
Intervention strategy	***	**	**	*
Professional development		***	***	
Program evaluation	***	**	*	*
Technology	**	**	**	*
English-language learners	**	**	**	
Standard 2				
Management and operations	***/**/*	***/**/*	***/*	***/**/*
Safe, orderly, and caring environments	**	*	*	
Operational systems	*	*	*	*
Human resources management and development	*	***	***	
Scheduling	*	***	*	
Management information systems	**	**	*	*
Laws, ethics, and policies	***	**	***	**
Fiscal systems	**	*	*	
Improvement planning	**	*	*	
School committee relations	**	*	*	***
Contract negotiations	*	*	*	
Standard 3				
Family and community engagement	**/*	*	*	***/**

(continued)

Appendix B (continued)

Massachusetts leadership standards/elements	Task 1 vision/ direction	Task 2 professional school culture	Task 3 individual teacher's effectiveness	Task 4 family and community engagement
Family engagement	**	*	*	***
Effective communication	**			***
Advocacy	*	*		**
Community connections	*	*		***
Cultural awareness	**	*		**
Standard 4				
Professional culture	***/**	***/**/*	***/**/*	***/**/*
Mission and core values	***	*	**	*
Shared vision	***	**		*
Personal vision	***	**	**	**
Transformational and collaborative leadership	***	***	***	**
Cultural proficiency	***	*	*	***
Ethical behavior	***	*	**	**
Continuous learner	**	***	***	*
Communications	***	***	**	***
Managing conflict	**	**	*	**
Team building	***	***	**	*
Time management	***	***	**	**

*Tertiary focus: Task requires knowledge of indicators (performance of the task "bumps into" indicator).
**Secondary focus: Task requires performance of indicators to successfully complete.
***Primary focus: Task directly assesses performance of indicators.

Appendix C

Topic	Question
Content	Does any element of the tasks and work products contain content that unfairly disadvantages a candidate because of gender, race, ethnicity, sexual orientation, national origin, religion, age, disability, or cultural, economic, or geographic background?
Language	Does any element of the tasks and work products contain language that unfairly disadvantages a candidate because of gender, race, ethnicity, sexual orientation, national origin, religion, age, disability, or cultural, economic, or geographic background?
Offense	Is any element of the tasks and work products presented in such a way as to offend a candidate because of gender, race, ethnicity, sexual orientation, national origin, religion, age, disability, or cultural, economic, or geographic background?

(continued)

Appendix C (continued)

Topic	Question
Stereotypes	Does any element of the tasks and work products reflect a stereotypical view of a group based on gender, race, ethnicity, sexual orientation, national origin, religion, age, disability, or cultural, economic, or geographic background?
Fairness	Taken as a whole, are the tasks and work products fair to all candidates regardless of gender, race, ethnicity, sexual orientation, national origin, religion, age, disability, or cultural, economic, or geographic background?

Acknowledgments

The authors thank the anonymous reviewers for their comments on earlier drafts, and the Massachusetts Department of Elementary and Secondary Education staff for their feedback on reports that are the basis for this article.

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) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Content validation uses experts to answer the question about how well the assessments reflect the core domains of knowledge and skills being assessed. Face validity answers the same question but is more subjective, reported by those who participate in the assessment (or support those who do).
2. The tasks that comprise the PAL system are aligned to the following standards and policies: The revised Professional Standards for Administrative Leadership, approved by the Massachusetts Board of Elementary and Secondary Education in December 2011; Educator Licensure and Preparation Program Approval regulations (603 CMR 7.00), which were amended and approved by the Board on June 26, 2012 (http://www.doe.mass.edu/boe/docs/2012-06/item4.html); the national performance assessment requirements of the Educational Leadership Constituents Council (ELCC), as enumerated in its national accreditation program standards (http://npbea.org/wp-content/uploads/2012/06/ELCC-Building-Level-Standards-2011.pdf); and National educational leadership policy standards, the Interstate School Leadership Licensure Consortium (ISLLC) 2008 standards (http://www.ccsso.org/Documents/2008/Educational_Leadership_Policy_Standards_2008.pdf).

References

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.

Augustine, C. H., Gonzalez, G., Ikemoto, G. S., Russell, J., Zellman, G. L., Constant, L., . . . Dembosky, J. W. (2009). *Improving school leadership: The promise of cohesive leadership systems*. Santa Monica, CA: Rand Corporation.

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.

Clifford, M., Menon, R., Gangi, T., Condon, C., & Horung, K. (2012). *Measuring school climate for gauging principal performance: A review of the validity and reliability of publicly accessible measures*. Washington, DC: American Institutes for Research.

Condon, C., & Clifford, M. (2010). *Measuring principal performance: How rigorous are commonly used principal performance assessment instruments?* Naperville, IL: Learning Points Associates.

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). *Education Leadership and Administration*, *22*, 66-82.

Duckor, B., Castellano, K. E., Telle, K., Wihardini, D., & Wilson, M. S. (2014). Examining the internal structure evidence for the Performance Assessemnt for California Teachers: A validation study of the elementary literacy teaching event for Tier I teacher license. *Journal of teacher education*, *65*(5), 402-420.

Educational Testing Service. (2009). *Multi-state standard setting report: School Leaders Licensure Assessment (SLLA)*. Princeton, NJ: Author.

Educational Testing Service. (n.d.). *School leadership series: State requirements*. Retrieved from https://www.ets.org/sls/states/

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.

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.

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.

Holden, R. B. (2010). Face validity. In I. B. Weiner & W. E. Craighead (Eds.), *The Corsini encyclopedia of psychology* (4th ed., pp. 637-638). Hoboken, NJ: Wiley.

Kaye, L. S. (2016). *Requirements for certification of teachers, counselors, librarians, administrators for elementary and secondary schools, eighty-first edition, 2016-2017*. Chicago, IL: University of Chicago Press.

Khattri, N., & Sweet, D. (1996). Assessment reform: Promises and challenges. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises, problems, and challenges* (pp. 1-22). Mahwah, NJ: Lawrence Erlbaum Associates.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice-Hall.

Lane, S. (2010). *Performance assessment: The state of the art*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

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.

Lit, I., & Lotan, R. (2013). A balancing act: Dilemmas of implementing a high-stakes performance assessment. *The New Educator*, *9*, 54-76.

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.

Martineau, J. (2004). Laying the groundwork: First steps in evaluating leadership development. *Leadership in Action*, *23*(6), 3-8.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Messick, S. (2005). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5-9.

Mezirow, J. (2000). *Learning as transformation: Critical perspectives on a theory in progress*. San Francisco, CA: Jossey-Bass.

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., Shear, B. W. A., & Darling Hammond, L. (2013). *2013 edTPA Field Test: Summary report*. Palo Alto, CA: 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 :Stanford Center for Assessment, Learning and Equity.

Popham, W. J. (2012). *Assessment bias: How to banish it*. Boston, MA: Pearson.

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.

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

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.

Stiggins, R. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practice*, *6*(3), 33-42.

Sun, C. (2011). *School leadership: Improving state systems for leader development*. Arlington, VA: National Association of State Boards of Education.

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

Turnbull, B. J., Riley, D. L., Arcaira, E. R., Anderson, L. M., & MacFarlane, J. R. (2013). *Six districts begin the principal pipeline initiative*. Washington, DC: Policy Studies Associates.

The Wallace Foundation. (2006). *Leadership for learning: Making the connections among state, district, and school policies and practices*. New York, NY: Author.

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

Wei, R. C., & Pecheone, R. L. (2010). Performance-based assessments as high-stakes events and tools for learning. In M. M. Kennedy (Ed.), *Handbook of teacher assessment and teacher quality* (pp. 69-132). San Francisco, CA: Jossey-Bass.

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.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, *45*, 197-210.

Young, M. E., Cruess, S. R., Cruess, R. L., & Steinert, Y. (2014). The Professionalism Assessment of Clinical Teachers (PACT): The reliability and validity of a novel tool to evaluate professional and clinical teaching behaviors. *Advances in Health Sciences Education*, *19*, 99-113.

Author Biographies

Margaret Terry Orr is a faculty member of Bank Street College of Education where she directs the Future School Leadership Academy. She was a project director for designing and field testing the Massachusetts Performance Assessment of Leaders. Her current research focuses on designing effective models for leadership preparation and assessment.

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.

Jon D. 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.

Ameetha Palanki is an executive director for Implementation and Content at Educopia. Palanki oversees the implementation strategy and instructional design for K-12 performance-based assessments in licensure, coaching, educator effectiveness, and leadership.

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.

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.

Joan L. Buttram serves as the director of the Delaware Education Research & Development Center at the University of Delaware. She also teaches courses in the education leadership doctoral program.