Massachusetts Performance Assessment for Leaders (PAL) Technical Report

Summary of Validity and Reliability Studies for 2014-15 Field Trial of PAL

Margaret Terry Orr, Bank Street College

Ray Pecheone, Stanford University

Irena Nayfeld, Stanford University

Ben Shear, Stanford University

Liz Hollingworth, University of Iowa

Barbara Beaudin, Bank Street College

Author Note

Margaret Terry Orr, Educational Leadership Department, Bank Street College; Ray Pecheone, SCALE, Stanford University; Irena Nayfeld, SCALE, Stanford University; Ben Shear, Center for Education Policy Analysis, Stanford University; Liz Hollingworth, Center for Evaluation and Assessment, University of Iowa; and Barbara Beaudin, independent consultant with Bank Street College.

This report was produced with assistance and support from the PAL assessment development team and advisors: Jon Snyder, Stanford University; Joe Murphy, Vanderbilt University; Afton Battle, Bank Street College; and Eric Docter and Ameetha Palenski, ShowEvidence. Prepared for the Massachusetts Department of Elementary and Secondary Education, it provides information pertaining to the Performance Assessment for Leaders (PAL). Wendy Schwartz edited the report.

Correspondence concerning this report should be addressed to Margaret Terry Orr, morr@bankstreet.edu.

**Abstract**

This report provides a summary of the design, development process, and validation results for a new performance assessment for initial school leaders in Massachusetts: Performance Assessment for Leaders (PAL). It provides a measure of leadership candidates’ readiness for initial school leader positions, informing licensure decisions while also supporting candidate learning and preparation program improvement. The assessment consists of four, field-based performance tasks that allow candidates to demonstrate their leadership knowledge and skills in planning for an area of school improvement, facilitating a professional learning group, observing and giving feedback to a teacher, and engaging families and the community in improving student learning. Candidates produce written memos, reports, and video products as evidence of their accomplishment of each task. PAL, which is aligned with state and national leadership standards and indicators, was developed with input from K-12 school and district leaders and higher education faculty. A series of validation studies confirmed PAL’s content validity, job relevance, face validity, feasibility, and construct validity. A pilot study and a field trial both document that PAL is well-aligned to leadership standards, reflects the work of school principals, and reliably measures the ability of initial school leaders to engage in effective practice. Analyses of candidate performance indicate no difference among types of preparation pathways (preparation programs in contrast with others) and some differences by gender (sample sizes were too small for race/ethnicity comparisons).

*Keywords:* performance assessment, principal licensure, validation

**Table of Contents**

Introduction 7

State-Approved Preparation Programs 8

Administrative Apprenticeship/Internship 9

Panel Review 9

PAL Assessment Design 9

Task 1 12

Task 2 13

Task 3 14

Task 4 15

PAL Content Validity 16

Step 1: Standards Alignment 18

Step 2: Design Committee Validation 19

Step 3: Content Committee Validation 19

Step 4: Pilot Study Candidate and Faculty (Face) Validation 22

Step 5: Field Trial Candidate and Faculty (Face) Validation 25

Content Validity, Face Validity, and Implications 31

PAL Bias and Sensitivity 32

Ease of Use and Feasibility 33

Step 1: Pilot Study Candidate and Faculty Ease of Use and Feasibility Assessment 34

Step 2: Field Trial Candidate and Faculty Ease of Use and Feasibility Assessment 37

Ease of Use and Feasibility Analysis and Implications 42

The PAL Scoring System 43

Indicator and Domain Scores 43

Rubric Indicator Proficiency Levels 44

Scorer Recruitment and Training 44

The Scoring Process 47

The Pilot Study (2013-14) 47

Pilot Sample 50

Pilot Study Score Profiles 51

The Field Trial (2014-15) 52

Construct Validity: Evaluations of the Indicators and Domains by Task 54

Task 1 Indicators and Domains 55

Task 2 Indicators and Domains 57

Task 3 Indicators and Domains 59

Task 4 Indicators and Domains 61

Dimensionality Analysis of Domain and Total Scores 63

Total Scores 65

Total Score Comparisons by Candidate Characteristics 70

Scoring Reliability 71

Scorer Agreement 71

Preliminary G Study Analyses 75

Technical Advisory Committee Feedback 78

Setting the Qualification Standards for Readiness for Initial School Leadership Licensure 79

Conclusions 82

References 85

Appendix A: PAL Design Committee Membership 86

Appendix B: PAL Content Validity Committee Membership 87

Appendix C: Degree of Alignment between the Massachusetts Standards for Administrative Leadership by the PAL Assessment Tasks 88

Appendix D: PAL Bias Review Committee Membership 91

Appendix E: Bias Review Topics and Questions 92

Appendix F: PAL Task Domains, Elements, and Indicators 93

Appendix G: PAL Technical Advisory Committee Membership 95

Appendix H: PAL Standards Setting Committee Membership 96

Appendix I: Descriptors of Readiness for the Four PAL Tasks 98

**Table of Tables**

Table 1: Number of State Approved Preparation Programs Whose Candidates Participated in the Field Trial by Type of Program and Candidates’ Estimated to Participate in the Field Trial 9

Table 2: Percentage of Content Validity Committee Members Rating the Task as Important or Very Important (or Well or Frequent), by Task (mean rating scores are included in parentheses) 21

Table 3: Number of Survey Responding Candidates by Current Position and Task Completed, and Number of Faculty by Task Supervised for the Spring Pilot Study 23

Table 4: Percentage of Responding Pilot Study Candidates Who Agree about Selected Qualities of the Assessments, by Task 24

Table 5: Percentage Distribution of Field Trial Candidate Survey Respondents by Selected Demographic Characteristics 26

Table 6: Percentage of Field Trial Candidate Survey Respondents by Current Position 27

Table 7: Percentage of Field Trial Candidate Survey Respondents by Field Work/Internship and Program Completion Status 28

Table 8: Percentage of Field Trial Candidates Who Agree or Strongly Agree with Task Attributes Related to Content Validity (4-point agreement scale) (n=92) 29

Table 9: Number of Respondents to the Field Trial Program Director Feedback Survey by Role  30

Table 10: Percentage of Responding Program Directors Who Agree or Strongly Agree with the Attributes of Each Task 31

Table 11: 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 35

Table 12: Percentage of Responding Field Trial Candidates Who Agree That the Task Was Flexible and Adaptable, by Task (n=92) 38

Table 13: Percentage of Responding 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) 39

Table 14: Percentage of Responding Field Trial Faculty Who Agree or Strongly Agree about Ease of Use, Feasibility, and Program-Related Attributes, by Task 41

Table 15: Number of Domains and Indictors by Task 43

Table 16: Number of Submissions for the Fall and Spring Pilot Studies by PAL Task 50

Table 17: Number of Submissions by Score Level and PAL Task 51

Table 18: Number and Percentage of Candidates Who Participated in the Field Trial by the Number of Tasks Submitted 52

Table 19: Number and Percentage Distribution of Candidates Who Completed All Four Tasks for the PAL Field Trial, by characteristic 53

Table 20: Task 1 Indicators and Domains 55

Table 21: Correlations between all Task 1 Domain Scores 56

Table 22: Task 2 Indicators and Domains 57

Table 23: Correlations between All Task 2 Domain Scores 58

Table 24: Task 3 Indicators and Domains 59

Table 25: Correlations between All Task 3 Domain Scoress 60

Table 26: Task 4 Indicators and Domains 61

Table 27: Correlations between All Task 4 Domain Scores 62

Table 28: Standardized Factor Loadings 63

Table 29: Factor Correlation for Four Factors 64

Table 30: Descriptive Statistics by Task 65

Table 31: Correlations of the Average Total Domain Scores by Task 70

Table 32: Mean and Standard Deviation for Domain Total Score by Demographic Attribute 71

Table 33: Scorer Agreement Rates by Indicator 71

Table 34: Estimated Reliability (G) Coefficients by Task and Number of Scorers 76

Table 35: Reliability of Total Score Based on Number of Scorers for Each Task 78

Table 36: PAL Threshold and Total Passing Scores used as the PAL Performance Standard… 81

**Table of Figures**

Figure 1: The Components of Leadership Action. 12

Figure 2: Distribution of Total Score for Task 1 (average of domain scores). 66

Figure 3: Distribution of Total Score for Task 2 (average of domain scores). 67

Figure 4: Distribution of Total Score for Task 3 (average of domain scores). 68

Figure 5: Distribution of Total Score for Task 4 (average of domain scores). 69

Introduction

The Commonwealth of Massachusetts offers multiple leadership preparation and licensure pathways and is committed to establishing a statewide performance assessment for principal licensure. Five years ago the state’s Department of Elementary and Secondary Education (ESE) allocated a portion of its federal Race to the Top funding to develop, pilot, and field test a performance-based assessment system for principal licensure. To further improve principal quality, 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 the professional standards, and creating a principal evaluation system.

To ensure that all leadership preparation candidates seeking principal/assistant principal licensure in Massachusetts meet state performance assessment requirements, ESE supported the development of the Massachusetts Performance Assessment for Leaders (PAL) by contracting with Bank Street College and its team of national experts on leadership and performance assessments (from Bank Street College, Stanford University, Vanderbilt University, the University of Delaware, and the University of Iowa) to lead the design, implementation, and validation of a performance assessment system for principal licensure. The PAL assessment system consequently became a collaborative project between ESE and its partner contractor, Bank Street College of Education.

The PAL assessment system is designed to provide both clear evidence of a candidate’s readiness for an initial school leadership position and evidence 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.

There are three pathways to principal licensure in Massachusetts, which are described immediately below: completion of a state-approved preparation program, an administrative apprenticeship/internship pathway, and a panel review process.

State-Approved Preparation Programs

Commonwealth-approved preparation programs include those that are university-only, are sponsored by a consortium of organizations that includes regional educational agencies, or are operated by a professional association. All must meet Commonwealth-defined requirements <http://www.doe.mass.edu/edprep/pr.html>). In 2012, the Commonwealth provided new requirements for preparation programs that require redesign and submission to ESE for approval in 2013. Consequently, only a few programs had a sufficient number of candidates to participate in the PAL Pilot Study in Academic Year (AY) 2013-14 and a few programs were still in the early phases of preparing candidates during the Field Trial in AY 2014-15.

The number and types of approved programs whose candidates participated in the Field Trial are shown in Table 1. Most are university-only programs.

Table 1

*Number of State Approved Preparation Programs Whose Candidates Participated in the Field Trial by Type of Program and Candidates’ Estimated to Participate in the Field Trial*

|  |  |  |
| --- | --- | --- |
| Type of Preparation Program | Number of Programs | Number of Candidates Anticipated for the Field Trial |
| University-only | 14 | 338 |
| Consortium-based | 7 | 85 |
| Professional association affiliated | 2 | 46 |
| Total | 23 | 469 |

Administrative Apprenticeship/Internship

This pathway, launched in 2001, was designed to enable districts to support aspiring education leaders by providing seminars and other leadership development learning experiences. Since 2012 candidates in this pathway have been required to complete at least 500 hours of internship experience and demonstrate proficiency in the Professional Standards and Indicators for Administrative Leadership (<http://www.doe.mass.edu/edeval/model/PartIII_AppxB.pdf>).

Panel Review

The panel review option is available to applicants who have completed an accredited leadership or management program and have had at least three years of administrative, leadership, or management experience. Candidates seeking licensure through this option must compile information on their professional education and professional experience and be interviewed by a panel of experienced administrators and educators.

PAL Assessment Design

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. 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.
* Task 4: Leadership for family engagement and community involvement.

The four tasks were identified and developed through a two-year planning and design process in which the Bank Street College team worked closely with a design committee (comprised of representatives from K-12 schools and district leadership and preparation pathways) and ESE staff. During the design process, the Bank Street team, ESE, and the design committee examined the core work of principals, current research, professional standards, and expectations for leadership preparation. From analyzing the themes that arose from this examination, the team distilled the four core tasks described above that would yield actionable and observable candidate performance across multiple standards. Collaborating further with the design committee, the Bank Street team developed work product instructions and rubrics to assess candidate performance on the tasks.

The tasks that comprise the PAL system are aligned to the following:

* 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>).
* 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>).

Each task is divided into four components of leadership action that reflect the cycle of leadership inquiry and learning, as shown on Figure 1.

**

*Figure 1*. The components of leadership action.

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 of 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. Lastly, 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 and identifies the group of teachers that will address it, and that 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 pre-observation template about the teacher and class to be observed with a summary of the pre-observation meeting.

(2) A 15-minute video recording of the observed teacher.

(3) A 15-minute video recording of the post-observation 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 pre-observation meeting, observation, and post-observation 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. Further, 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 implements 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; identification of members of the working group for planning (including family and community member representation). It must also include 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 and how.

PAL Content Validity

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. As noted above, the Massachusetts Department of Elementary and Secondary Education (ESE) worked in partnership with lead project staff from Bank Street College and other research experts to design, pilot, and field test PAL. As part of the process, they brought together representatives from a number of Massachusetts preparation programs and pathways and K-12 education leaders to work with Bank Street College and ESE to develop PAL. These content area experts serve on either a design committee or a content validity committee (see Appendices A and B for members of these committees). Members of each committee reviewed the four draft tasks and the assessment system before they 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 two committees conducted follow-up reviews after the Pilot Study and made revisions before the Field Trial was launched in September 2014, and again after the Field Trial, before the Program Year 2015-16 was launched.

The content validation of the assessment tasks began with the initial design work and continued through each phase of assessment refinement. Determining the content validity requires addressing the question: How well does the content of PAL represent core domains of school leadership knowledge and skills? Pilot Study and Field Trial participants 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 point (not as experts), as complementary to the content validation.[[1]](#footnote-1) In all, Bank Street College and ESE completed three steps to establish content validity and two steps to confirm face validity:

* Standards alignment: PAL task alignment to specific Massachusetts Professional Standards and Indicators for Administrative Leaders, performed by Bank Street College staff and reviewed by ESE staff and design committee members.
* Design committee validation: Design committee review and validation of the PAL tasks, instructions, and rubrics.
* Content validity committee validation: Content validity committee review and validation of the PAL tasks and instructions.
* Pilot study candidate and faculty face validation: Survey of Pilot Study candidates and university faculty who supported Pilot Study candidates whose results were reviewed by the design committee.
* Field Trial candidate and faculty face validation: Survey of Field Trial candidates and university faculty, the results of which were reviewed by the design committee.

Step 1: Standards Alignment

In the initial assessment design phase, completed in 2012-13, the development team engaged in an iterative process of reviewing state and national leadership standards and research; and developing the tasks through collaboration among the project staff, ESE staff, and school and district leaders and leadership faculty members from the Commonwealth who constituted the design committee.

The development team, ESE staff, and design committee members confirmed (through discussion and formal committee member affirmation) alignment of the four PAL assessment tasks to the Massachusetts leadership standards. Appendix C shows the primary and secondary comprehensiveness or proportionality of the assessment tasks to the domains of knowledge or skills that define competent practice by these standards and their indicators.

By design, the PAL assessment tasks were created to sample essential leadership competencies (knowledge and skills). Thus, the four 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 too complex to measure through a performance assessment.

Step 2: Design Committee Validation

Once drafted and revised, but before piloting, the four PAL tasks were reviewed by the ten-member design committee for relevance and feasibility as initial school leader assessments. Committee members provided feedback on the proposed work, work product instructions, and skills assessed by each task. At this design stage, the committee members, through a day-long review and discussion, agreed strongly with focus and nature of the assessment tasks, their relevance for initial school leaders’ work, and their feasibility. They also confirmed the analysis of the tasks and standards alignment. Their primary recommendations for changing the tasks focused on the instructions in order to strengthen the clarity of directions and explore how to gain local school and district support for candidates to complete the work.

Step 3: Content Committee Validation

To formally assess content validity, we convened a group of subject matter experts who were then trained to rate the content validity of each task using a standard rating form and process. This ten-member content validity committee of school and district leaders and leadership preparation faculty members was convened in Spring 2013, prior to the Pilot Study launch.

During the day-long meeting, the committee was trained in content validation. Next, members met in two small groups and individually rated the content validity of each task using a 5-point scale that addressed job relevance, authenticity, and importance by rating each task description according to the following attributes:

* How important the knowledge and skills assessed inthe task are for performing the job of an entry-school principal.
* How important the knowledge and skills assessed inthe task are for improving student learning.
* How well the set of components and products required for the taskreflect the authentic work that an entry-level principal must perform on the job.
* How frequently an entry-level principal demonstrates the knowledge and skills in the work products required for the task while on the job.

Following its review, each small group and then the committee as a whole discussed their ratings, and the committee members submitted their ratings for summary analysis and reporting. To evaluate the ratings, we used Wilson, Pan, and Schumsky (2012) criteria that half or more of the committee members must agree or strongly agree on the task attributes. 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 is 62 percent and for 9 panelists is 78 percent). While the authors (Wilson, Pan, & Schumsky, 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. The PAL content validity ratings were equal to or greater than these recommended values, demonstrating strong content validity.

Table 2 shows that most content validity committee members rated the tasks as important or very important for performing the job of an entry level principal and improving student learning; all mean rating scores were 4.0-4.8 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. The committee members provided suggestions for how the content of the tasks and instructions could be strengthened and focused further on leadership to improve student learning and reduce ambiguity in the meaning of specific terms and processes. The committee members also recommended ways to improve the guidelines for work products to encourage submission of more authentic and relevant forms of communication.

Table 2

Percentage *of Content Validity Committee Members Rating the Task as Important or Very Important (or Well or Frequent), by Task (mean rating scores are included in parentheses)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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.

When asked at the end of the day to evaluate the PAL assessment system as a whole, almost all committee members present (89 percent) rated the four sets of tasks and products as “well” or “very well” in reflecting the authentic work that an entry-level principal must perform on the job. Most (89 percent) also agreed than 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.

Step 4: Pilot Study Candidate and Faculty 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 candidates and supervising 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 (Holden, 2010) p. 637), face 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.” At best, 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. Nonetheless, their feedback is informative and provides additional perspectives. In addition, as Holden (2010) found, face validity is positively associated with other forms of technical validity.

At the end of the fall and spring Pilot Studies, candidates, program directors, and faculty members (spring Pilot Study only) completed an online survey for feedback on their experiences with the tasks, work products, assessment process, and assessment management system. Respondents comprised most candidates who submitted work products and all directors of programs whose candidates participated in the Pilot Study.

Respondent information, shown in Table 3, shows that the majority of candidates who submitted work also completed the survey. Most were currently classroom teachers or non-teaching professional staff. A few were instructional specialists or department chairs. A few candidates completed two tasks (most typically Tasks 1 and 4), and were included in both task counts.

Table 3

*Number of Survey Responding Candidates by Current Position and Task Completed, and Number of Faculty by Task Supervised for the Spring Pilot Study*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Characteristics | Task 3 FallPilot | 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% |
| Current Position |
| Classroom teacher | 8 | 6 | 3 | 4 | 6 |
| Non-teaching support staff (e.g., guidance or social worker) | 0 | 2 | 2 | 0 | 3 |
| Instructional specialist | 1 | 0 | 1 | 2 | 0 |
| Central district nonsupervisory | 0 | 0 | 1 | 0 | 0 |
| Department chair or grade level leader | 3 | 0 | 3 | 1 | 1 |
| Other | 3 | 4 | 0 | 0 | 5 |
| Assistant principal | 1 |  |  |  |  |
| Total | 13 | 12 | 10 | 8 | 15 |
| Total Program Faculty Members (n=4 total) | -- | 1 | 1 | 1 | 1 |

*Note*: Some candidates who responded to the survey reported multiple roles, and some appear to have misidentified themselves as having completed Task 2 or 3 as the total exceeds the number of submissions.

To determine PAL face validity, candidates and faculty members were asked to rate how strongly they agree that the tasks provide authentic experiences and are relevant to their preparation. As shown in Table 4, most candidates agreed that the tasks are aligned to the Massachusetts Standards for Administrative Leadership and complementary to their leadership preparation. Most also agreed that the tasks provided them with authentic job-related experiences and that the tasks are relevant and essential to the work that successful school leaders must be able to do (although somewhat less for Task 2 than for the other three tasks).

Table 4

*Percentage of Responding Pilot Study Candidates Who Agree about Selected Qualities of the Assessments, by Task*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Qualities | Task 3 FallPilot | Task 1 Spring Pilot | Task 2 Spring Pilot | Task 3 Spring Pilot | Task 4 Spring Pilot |
| 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 MA 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 |

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. Most agreed that that they understood the work that candidates must engage in to complete the task and that the process of supporting their candidates in completing one task had been a catalyst for rethinking how they prepare school leaders.

Step 5: Field Trial Candidate and Faculty (Face) Validation

We continued to monitor face validity—specifically task alignment to the Massachusetts Standards for Administrative Leadership and to job relevance—as part of the Field Trial. As in the Pilot Study, participating candidates and supervising faculty were asked to complete an online survey with questions about the tasks’ relevance to an initial school leader job and alignment to the standards. Of the 415 candidates who completed all four tasks, only 58 completed the feedback survey, representing a 14 percent response rate. Table 5 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 we have for the total Field Trial sample (as shown below in the Field Trial results section). 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.

Table 5

*Percentage Distribution of Field Trial Candidate Survey Respondents by Selected Demographic Characteristics*

|  |  |  |
| --- | --- | --- |
| Demographic | Number of Responses | Percentage |
| Gender |
| Male | 14 | 24% |
| Female | 44 | 76 |
| Total | 58 | 100 |
| Race/Ethnicity |
| African American | 3 | 5 |
| American Indian/ Alaskan Native | 0 | 0 |
| Asian | 2 | 4 |
| Hispanic/Latino/a | 1 | 2 |
| White (not Hispanic) | 44 | 77 |
| I would prefer not to answer | 7 | 12 |
| Program/Pathway |
| University-based leadership preparation program | 38 | 67 |
| A consortium or association-based leadership preparation program | 8 | 14 |
| Administrative apprenticeship/internship option | 9 | 16 |
| Panel review option | 2 | 4 |
| Total | 57 | 100% |

Table 6 shows that the majority of Field Trial candidate respondents are currently in a non-supervisory position: classroom teacher, professional support staff, instructional specialist, or other professional nonsupervisory staff.

Table 6

*Percentage of Field Trial Candidate Survey Respondents by Current Position*

|  |  |  |
| --- | --- | --- |
| Current Position | Number of Responses | Percentage |
| Classroom teacher | 46 | 50% |
| Non-teaching support staff (e.g., guidance, social work) | 8 | 9 |
| Instructional specialist | 8 | 9 |
| Central district, non-supervisory staff | 3 | 3 |
| Department chair, team leader, grade level leader | 15 | 16 |
| Principal | 1 | 1 |
| Other  | 19 | 21 |
| Assistant principal | 12 | 13 |
| Total | 92 |  |

*Note*: Respondents could check more than one category so the percentages total more than 100%.

The majority of the respondents had completed their internship and program, but about one third had not, as shown in Table 7.

Table 7

*Percentage of Field Trial Candidate Survey Respondents by Field Work/Internship and Program Completion Status*

|  |  |  |
| --- | --- | --- |
| Candidate Status | Number of Responses | Percentage% |
| Field Trial Status |  |  |
| Have not yet begun a school leadership internship or field experience | 10 | 11% |
| Am currently participating in a school leadership internship/ field experience | 19 | 21 |
| Have completed a school leadership internship or field experience | 58 | 63 |
| Other (Specify) | 5 | 5 |
| Total | 92 | 100 |
| Program Status |
| Have already completed my program | 55 | 60 |
| Will complete this summer | 4 | 4 |
| Will complete this fall | 11 | 12 |
| Will complete next spring | 15 | 16 |
| Will complete after next spring | 7 | 8 |
| Number of Field Trial Candidates | 92 | 100 |

Table 8 shows that most Field Trial candidate survey respondents agreed or strongly agreed that the tasks provided authentic job-related experiences, were relevant and essential to the work of successful school leaders, were aligned to the Massachusetts standards, and were complementary to their leadership preparation. They were most in agreement for Task 3 and least for Task 1. Because of the small sample size and modest survey response rate, these findings should be interpreted cautiously and not be over generalized.

Table 8

*Percentage of Field Trial Candidates Who Agree or Strongly Agree with Task Attributes Related to Content Validity (4-point agreement scale) (n=92)*

|  |  |
| --- | --- |
| Attribute | % Agree or Strongly Agree |
| Task 1 | Task 2 | Task 3 | Task 4 |
| The task provides candidates with authentic job-related experiences | 70.5% | 75.6% | 81.1% | 73.3% |
| The task is relevant and essential to the work that successful school leaders must be able to do | 75.0 | 76.8 | 89.0 | 81.1 |
| The task is aligned to the MA standards | 84.1 | 81.7 | 89.0 | 84.9 |
| The task was complementary to my leadership preparation | 56.3 | 70.7 | 82.2 | 68.5 |

Field Trial candidates were more likely to agree on these attributes for Task 3 than for the other three tasks. 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 over half (56 percent) 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 thus more likely to have somewhat different preparation.

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, we do not know which programs are represented in the feedback provided. Ten faculty provided feedback, representing 43 percent of the 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 9, most were program directors and others were faculty or instructors with their programs.

Table 9

*Number of Respondents to the Field Trial Program Director Feedback Survey by Role*

|  |  |  |
| --- | --- | --- |
| Role | Number of Responses | Percentage |
| Program director | 7 | 70% |
| University faculty member | 1 | 10 |
| Course instructor | 2 | 20 |
| Other (specify) | 0 | 0 |
| Total | 10 | 100 |

As shown in Table 10, 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.

Table 10

*Percentage of Responding Program Directors Who Agree or Strongly Agree with the Attributes of Each Task*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attribute | Task 1 | Task 2 | Task 3 | Task 4 |
| The task that my candidates completed is clearly aligned to the MA standards | \* | 100% | 100% | 100% |
| The task provided candidates with authentic job-related experiences | 90.0 | 87.5 | 87.5 | 88.9 |
| The task is relevant to the work that successful school leaders must be able to do | 90.0 | 100 | 100 | 100 |
| Number of responses | 10 | 9 | 8 | 9 |

\*This question was not asked for Task 1.

Content Validity, Face Validity, and Implications

The PAL tasks have very good content validity, based on the strong agreement from the PAL design and content validity committees, and reinforced by the face validation from Pilot Study and Field Trial surveys of program faculty and candidates. Both committees strongly agreed that the four PAL tasks are aligned to the Massachusetts Standards for Administrative Leadership, provide authentic job-related experiences, and are relevant to the work that successful school leaders must be able to do. The strong agreement among the content validity committee members for all indicators and tasks exceeds professional standards for content validity (Wilson et al., 2012). These results were further confirmed by strongly positive agreement ratings for face validity in the Pilot Study and Field Trial surveys of preparation program faculty and the positive agreement among most Pilot Study and the majority of the Field Trial candidates.

PAL Bias and Sensitivity

To evaluate the assessment tasks for possible bias and sensitivity prior to the Pilot Study, we followed Educational Testing Service assessment review guidelines and Popham’s 2012 booklet on removing assessment bias (Popham, 2012). Working in collaboration with Massachusetts ESE staff, Bank Street College project staff formed a PAL bias review committee of nine experienced educational leaders and program faculty with expertise in detecting varied forms of bias (see Appendix D for members of this committee).

The committee was convened on May 21, 2013. After an orientation to the PAL assessments and their design and an overview of the bias review process, members were trained in the core concepts of bias and sensitivity and means of assessing it using a bias review handbook, developed for PAL, and bias-related evaluation forms for each task. The forms asked five sets of questions covering various types of bias related to content, language, offense, stereotypes, and fairness (see Appendix E for a summary of the questions).

During the bias review meeting, committee members were trained in the types of potential bias and the process for reviewing the tasks and instructions for possible types of bias. Following the training, the committee members evaluated each task independently and then discussed their responses in a group. Finally, they submitted their ratings, which were in the form of qualitative statements about potential bias or insensitivity for each of the five indicators. We compiled these statements by indicator and task to identify areas of agreement.

The committee members’ bias and sensitivity review feedback was generally positive; they found little evidence of either bias or insensitivity. The committee members asked that we replace some educational terminology with more generic language and remove task work examples that suggested possible stereotypes. Most of their discussion focused on how to ensure that all candidates would be able to participate fully and equitably in the completion of assessment tasks and thus demonstrate what they know and can do, given different types of district and support. This points to a general issue of feasibility, which is discussed in the section below.

The results of the committee’s evaluation and feedback were then used to revise the four tasks to reduce potential bias and increase sensitivity. The results were reported in the Psychometric Report to ESE.

Ease of Use and Feasibility

A major question raised by the bias review committee concerned the feasibility of the tasks for candidates to complete in their schools or districts and the ease of use of the instructions for completing the work. The committee members were concerned 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. The committee members were also concerned that the assessment system and its related materials might present technical challenges that could inadvertently limit a candidate’s performance. Similarly, they had been concerned that some candidates might have less access than others to information and support to complete the tasks.

To address this problem, ESE and project staff took several steps to explain PAL and the task requirements to the educational professionals most likely to be involved. We developed communication materials for preparation programs to use with school and district leaders, to explain each task requirements 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.

To investigate this potential assessment issue, candidates and preparation program faculty provided feedback as part of the Pilot Study and Field Trial. Feedback survey questions were designed to determine whether PAL-related guidance materials and information facilitated or hindered task completion and whether there were feasibility issues in candidates’ schools or districts that possibly mediated task completion and candidate performance.

Below is a summary of the findings through the Pilot Study and Field Trial periods.

Step 1: Pilot Study Candidate and Faculty Ease of Use and Feasibility Assessment

As part of Pilot Study, candidates and program directors completed feedback survey questions on the ease of use and feasibility of each task. The results, shown in Table 11, 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 Tasks 2 and 3 candidates thought the *Handbook* was easy to understand, in contrast with almost all of the 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.

 Table 11

*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 Task 4.

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, so agreed to provide little specificity on the format and detail for task submissions in the *Handbook* instructions. This problem was addressed for 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 candidates were also 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 concerned that Task 3 would present challenges because it required video recording a teacher giving instruction, so 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.

Ease of use for the information management system, ShowEvidence, was separately evaluated and generally positive, as shown in Table 11. 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.

The three faculty members who responded to the Pilot Study survey also addressed questions of ease of use and feasibility. They agreed that they understood the task requirements, but were in less accord about how well they understood the differences for score levels for each rubric. Most agreed that they could not explain the scoring criteria and standards used to evaluate the work products. Based on this feedback and the candidate feedback, we concluded that improved instructions and rubric language for the Field Trial should reduce these problems and improve their ease of use for faculty as well as candidates. The primary faculty member suggestion for improving the assessment management system was to make sure the directions were consistent between the *Handbook* and the ShowEvidence directions, particularly on how to bundle various documents for uploading.

Step 2: Field Trial Candidate and Faculty Ease of Use and Feasibility Assessment

As part of the Field Trial, we followed up with candidates and program faculty 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 12, 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.

Table 12

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

|  |  |
| --- | --- |
| Attribute | Percent Agree or Strongly Agree |
| 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) | 65.9% | 47.6% | 67.1% | 48.6% |
| Completing the task required a realistic amount of work | 29.1 | 75.6 | 81.1 | 73.3 |
| The task is flexible and adaptable so candidates in different types of school settings can structure meaningful activities and produce relevant products | 47.1 | 57.3 | 69.9 | 63.5 |

For further consideration of feasibility, candidates were asked to rate the difficulty of completing the steps from each of the four tasks, using a 5-point difficulty scale. We were interested in learning whether any specific step was perceived to be more difficult than others. Such feedback could be useful for preparation programs to highlight areas for further development, and for communicating to school and district leaders about providing candidate support in task completion. We later compared these findings with candidate scores on individual indicators to determine possible relationships.

As shown in Table 13, 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 generally were somewhat less difficult than those in Tasks 2 and 4. The least difficult step appeared to be the candidates’ ability to assess their own 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.

Table 13

*Percentage of Responding 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)*

| Task Requirement | Percent Who Rated the Difficulty as Very Easy to Neutral |
| --- | --- |
| 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 pre-observation conference |  |  | 73.3 |  |
| Document a teacher observation using a district or state guide on effective teaching practices |  |  | 71.6 |  |
| Conduct a post-observation 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 multi-strategy 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% |

Most program directors (63-90 percent, 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 14). The majority (50-75 percent) 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 percent) that the tasks aligned with their programs’ curriculum and the majority (63-75 percent) agreed that supporting their candidates in completing the tasks had been a catalyst for their program’s work on preparation, particularly for Task 3.

Table 14

*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 | 70.0% | 75.0% | 50.0% | 66.7% |
| It is feasible for candidates to complete the task within the structure of a course or internship that my institution offers | 90.0 | 75.0 | 62.5 | 88.9 |
| The task is aligned to the curriculum of the program that my institution offers to prepare new school leaders | 100 | 100 | 100 | 100.0 |
| The process of supporting candidates in completing this task has been a catalyst for rethinking how we prepare school leaders | 70.0 | 62.5 | 75.0 | 66.7 |
| Number of responses | 10 | 9 | 8 | 9 |

Ease of Use and Feasibility Analysis and Implications

Throughout the design, Pilot Study, and Field Trial, the tasks were reviewed for ease of use, feasibility, and adaptability of the task work for a variety of school settings. Feedback from the Pilot Study was used to improve instructions and guidelines for the Field Trial. A comparison of feedback from the two administrations shows that Field Trial candidates were far more likely than Pilot Study candidates to agree that they felt prepared to collect information for Tasks 1 and 4 but were as likely as the Pilot Study candidates to feel prepared for Tasks 2 and 3. The Field Trial candidates were much more likely than the Pilot Study candidates to report that Tasks 2, 3, and 4 were flexible and adaptable to different types of settings and only slightly more likely to report that Task 1 was flexible and adaptable to different types of settings. The perceptions of whether the tasks required a realistic amount of work improved for Tasks 2, 3, and 4 but did not improve for Task 1, when comparing the Field Trial and Pilot Study candidates. The faculty member respondents had different perceptions and continued to perceive Task 3 as the least flexible and feasible and Task 1 as the most or nearly the most flexible. A closer look at candidate feedback on the difficulty of the steps in each task shows that soliciting input or feedback from others was more difficult than analyzing relevant school and community data. Thus, access to student performance data may not be as much of feasibility problem as it had been initially, nor was getting permission to video record teacher observations. Moreover, it appears that changes in the instruction for the Field Trial, primarily provision of more direction on the scope and content of the written documents, were beneficial in making the task work more feasible.

For the implementation year, the instructions were modified further to clarify how the task work could be adapted to different settings; to address this feasibility issue, particularly with regard to the types of data to use to establish an academic priority area; and to indicate more clearly the fact that the teacher observations are not formal teacher observations but, rather, a focus on a few qualities of good teaching. Other changes to the instructions and rubrics were made to improve ease of use.

The PAL Scoring System

The PAL scoring system includes rubrics for indicator and domain scores, scorer recruitment and training, and the scoring process itself. The quality and use of each element, which contributes to the validity and reliability of the assessments, is discussed immediately below.

Indicator and Domain Scores

The four independent tasks, described above, were designed to enable leadership candidates to demonstrate their leadership skills and practices as defined by three or four domains for each task (see Appendix F for a description of the PAL task domains, elements, and indicators). Each domain score is the average of one to three indicator scores. The 13 domains and 26 indicators, as shown in Table 15, 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.

Table 15

*Number of Domains and Indictors by Task*

|  |  |  |
| --- | --- | --- |
| PAL Task | Number of Domains | Number of Indicators |
| 1  | 3 | 6 |
| 2  | 3 | 6 |
| 3  | 4 | 8 |
| 4  | 3 | 6 |
| Total number | 13 | 26 |

Rubric Indicator Proficiency Levels

The PAL assessment system incorporates rubrics that define the initial school leader proficiency levels for each indicator within each domain. In the design phase of the assessment development, the rubric indicators were initially defined and scaled on a 5-point continuum for the Pilot Study. These gradations were reviewed by the design committee to evaluate assumptions of leadership readiness. Following the Pilot Study and a review of the results with the design committee, the development team concluded that a 4-point scale was more applicable. The indicator levels were redefined in which a rubric level 1 represents a performance that is beginning and may not demonstrate competency. Level 2 represents a performance that is developing but does not yet meet the level of performance required for a beginning school leader. Level 3 represents a performance that does meet expectations for a beginning school leader. Level 4 represents a performance that would be exceptional for a beginning school leader. These revised levels were reviewed by the design committee members and applied in the Field Trial.

Scorer Recruitment and Training

Valid and reliable scoring is largely dependent upon the recruitment of experts as scorers and their scorer training and supervision. In the Pilot Study year and the Field Trial, scorers were recruited for their expertise in school leadership; they included school and district leaders and leadership preparation faculty from across the state. To be eligible for a scorer position, recruits had to be an experienced Massachusetts school or district leader or preparation program faculty member, be available for scorer training and scoring throughout the academic year, and also have the time and interest in scoring candidate work for a small fee.

To select the scorer training approach that would yield the best trained and most reliable scorers, two scorer training models were piloted during the Pilot Study and Field Trial. Part of this pilot was to determine the most feasible design that would optimize scorer recruitment, retention, and reliability. For the Pilot Study year and the first half of the Field Trial, scorer training was conducted in person in a convenient location in and around the greater Boston metropolitan area and at ESE offices. Two PAL development staff members designed and facilitated the day-long scorer training. Initial training consisted of provision of an overview of the PAL purpose and development process and a review and discussion of the tasks, rubrics, and assumptions about expectations for candidate proficiency. Prior to training, potential scorers were asked to evaluate and score one sample submission, identifying the evidence found in the submission and the corresponding rubric indicator level. During the training, this sample candidate submission and its scoring were discussed and questions about evidence and scoring indictor levels were answered. Next, potential scorers scored two additional submission samples (representing weaker and stronger examples), and the evidence in the samples was discussed for alignment to the rubrics. At the end of training, potential scorers were assigned a new sample to score as a test. PAL development staff reviewed the potential scorers’ ratings. If they matched the scores determined as accurate by the assessment development staff, the potential scorers would be certified and assigned five submissions to score. If a potential scorer’s ratings did not match the assessment development staff scores for two or more indicators, the assessment development staff provided follow-up training. Follow-up consisted of talking through the evidence and rating with the potential scorer one-on-one and then supplying a new submission sample for score assessment and training follow up.

Due to severe weather conditions during the Field Trial period, the PAL development staff, in consultation with ESE staff, switched to a webinar style training format. Using ZOOM to support video conferencing, potential scorers and PAL development staff logged in for a presentation on PAL, discussion of the submissions and scoring evidence, and organizational and logistical information. The potential scorers were very positive about the webinar-based scoring because it did not require travel time and simplified meeting breaks for lunch and other purposes. There seemed to be little difference in the scorer certification rate between the in-person and web-based training formats. Continuing the scorer certification process, PAL staff reviewed scoring by double scoring up to one in four submissions for each scorer and discussing scoring discrepancies when more than two indicators were scored differently.

Certified scorers were limited to scoring only two of the four tasks (with rare exception), although most only trained and became certified for one task. All scorers rated only one task per candidate, regardless of the number of different tasks they were scoring, thereby ensuring independence in scoring and reducing possible scorer bias. Thus, each candidate is rated by four different scorers, one per task.

During the pilot year, nine scorers were trained and certified for one to three tasks. During the Field Trial, thirty scorers were trained and certified to score one task; of these, eight were trained and certified to score two tasks. At the end of each training session and at the end of the Pilot Study and Field Trial, scorers were surveyed about the training, scoring, and logistics of the scoring process, as well as the tasks, rubrics, and instructions for candidates. Survey results provided useful feedback on how to improve the training, scorer support, rubrics and tasks.

The Scoring Process

Scorers are advised to score one submission at a time. As each submission includes three to four artifacts (consisting of written documents or video recordings), supporting documents, and commentary, scorers were advised to begin by reading through all artifacts and commentary before starting to score the evidence. All scoring is done online, using a confidential assessment management system, ShowEvidence. In this system, scorers rate evidence by rubric indicator and proficiency level, using a tagging process. Scorers highlight text in each artifact and commentary of a candidate submission and apply “tags” that reference a rubric indicator and proficiency level. The documents in a submission are important background for scoring and are to be read but not tagged. As they score, scorers review a candidate’s submission evidence according to the scoring guide and evaluate where the preponderance of evidence is for specific proficiency levels of each indicator. When the tagging is complete, the ShowEvidence system accumulates the score levels by indicator automatically. Scorers review the summary scores by indicator, rubric and total score, and then submit them as final.

To limit potential scoring bias, scorers notify ShowEvidence if they know the candidate or school in the submission (although blinded) so that the submission can be reassigned. To support reliability assessment, 25 percent of the candidate submissions for each task are double scored by other trained scorers or assessment development staff.

The Pilot Study (2013-14)

The Pilot Study, conducted in Academic Year 2013-14, was designed to ascertain the feasibility of the task requirements, the candidates’ interpretation of the tasks and work product instructions, and the form and nature of their submissions.

For each task, candidates submit artifacts, other documents, and commentaries as evidence of planning, implementing, and analyzing leadership practices and of engaging others in improving student learning. The artifacts represent authentic work completed by the candidates and others they involved in task completion. They can be plans, proposals, reports, video recordings of observations and post-observation feedback, memos, and other documents to demonstrate work. As backup to their artifacts and to aid the scorers, candidates also submit additional relevant documents they and their work groups, if applicable, used to complete each task, such as student performance data, school and district demographic information, observation guides and rubrics, meeting agendas and minutes, school vision statements, strategic plans and improvement plans, curriculum guides, and program information. Finally, to examine what they learned from undertaking the tasks, candidates analyzed their experiences in completing them by submitting commentaries describing the artifacts and explaining the leadership practices they employed, their choices, analyses, and reasoning, drawing on the other documents. As part of the Pilot Study, candidates were given broad latitude on the form they could use for their artifacts and their length. Candidates were also given broad latitude on the number and form of the other documents.

During AY 2013-14, Bank Street College lead staff launched fall and spring PAL Pilot Studies to accommodate the disrupted admissions to preparation programs and preparation cycles caused by program redesign and review. To recruit candidates, leadership preparation program directors volunteered to ask their candidates to complete just one task to test its feasibility and usefulness and to generate results to help set scoring benchmarks and exemplars. Candidates were given one semester to complete one task and upload their work products for scoring in the ShowEvidence system. The work product requirements were new to the programs and candidates and programs had little time to align their preparation and faculty support as candidates completed the pilot work products.

Of the thirty-one leadership preparation programs in Massachusetts prior to redesign, only twenty-one (two thirds) were approved by ESE (as part of program redesign) when the Pilot Study was launched in September 2013 and most were approved with conditions. Ultimately, only nine programs volunteered for the fall Pilot Study and only six had candidates who submitted work. Of them, only two programs attempted to support candidates in performing the tasks as designed for the assessment, while the rest had candidates submit coursework that was similar to the PAL tasks in order to test out the system.

Throughout the fall, leadership preparation programs worked toward unconditional program approval or prepared to submit a new or revised program redesign plan. By January 2014, more programs had been approved and started implement their new programs by recruiting and enrolling new cohorts of candidates, expanding the potential recruitment pool for Pilot Study.

We re-launched the Pilot Study in January 2014, making a second effort to recruit programs and candidates. By then, twenty-nine programs had been approved (some with conditions). To add further incentive for program and candidate participation, ESE officials issued a formal announcement that established guidelines requiring all candidates to complete all four tasks when the Field Trial started in fall 2014. This policy action helped us to encourage programs to participate in the Pilot Study as a way of learning more about the assessment tasks and becoming familiar with the system. Many more programs were interested in participating in the Pilot Study, but by the end of the semester, only a few had candidates who submitted completed work products.

In all, six preparation programs had candidates who participated in the fall Pilot Study and five programs had candidates who participated in the spring Pilot Study, with two programs participating in both. In all, nine of the twenty-nine approved leadership preparation programs in the Commonwealth participated in one or both PAL Pilot Studies.

As noted above, at the same time, we recruited potential scorers through requests to the leadership preparation field and the professional associations for principals. Potential scorers were trained in spring 2014. Trained scorers then scored the Pilot Study submissions in March and June 2014. Candidates, scorers, and program faculty provided feedback on the tasks, and assessment support and infrastructure.

Pilot Sample

During 2013-14, 134 candidates submitted work products for scoring and the products were compiled in our online assessment management system, ShowEvidence. Because of the newness of the assessments and the programs’ redesign process, few programs had candidates who were eligible, based on their preparation and interest, to complete even one task during the two Pilot Study phases. To gain at least some feedback on the ShowEvidence system and see examples of candidate work products, we allowed some programs’ candidates to submit only part of the required work products. As shown in Table 16, only half of the candidate work product submissions were complete and storable. Incomplete submissions or submissions that were not germane to the task (such as a course paper) were determined to be not scorable.

Table 16

*Number of Submissions for the Fall and Spring Pilot Studies by PAL Task*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PAL Task | Fall Pilot Study 2013 | Spring Pilot Study 2014 | Total Number of Submissions | Number Scorable |
| 1  | 9  | 24  | 33 | 22 |
| 2  | 7  | 7  | 14 | 6 |
| 3  | 19  | 7  | 26 | 16 |
| 4  | 34  | 27  | 61 | 22 |
| Total number | 69  | 65  | 134 | 66 |

Pilot Study Score Profiles

During spring 2014 we trained nine scorers to assess the submitted work and often used two or more scorers to rate the work products. For the Pilot Study, we used a 5-point scoring rubric, with definitions for three of the five score levels. Table 17 shows that most of the scored submissions were rated between 2 and 4 on the scale, with submissions for Tasks 2 and 3 most frequently rated 3 (meeting expectations), submissions for Task 4 rated just below a 3 rating, and submissions for Task 1 distributed somewhat evenly among levels 1 through 4. These preliminary results show that the complete work products provided sufficient evidence to be rated and that candidate performance could be differentiated based on the evidence and the rating rubrics. The ratings are based only on one task per candidate, however, and will need to be investigated further when candidates complete all four tasks.

Table 17

*Number of Submissions by Score Level and PAL Task*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Score Levels | Task 1 | Task 2 | Task 3 | Task 4 |
| # | % | # | % | # | % | # | % |
| 1-Developing | 2 | 17% | 0 | 0 | 1 | 8% | 4 | 18% |
| 2 | 4 | 33 | 0 | 0 | 3 | 19 | 11 | 50 |
| 3-Meeting | 3 | 25 | 2 | 67 | 12 | 75 | 5 | 23 |
| 4 | 3 | 25 | 1 | 33 | 0 | 0 | 2 | 9 |
| 5-Exceeding | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Based on the Pilot Study scoring experiences, we revised the rubrics along a 4-point scale, and improved the alignment between the instructions and rubrics to improve their use in scoring.

The Field Trial (2014-15)

The PAL Field Trial was conducted from September 1, 2014 to May 15, 2015. During this time, all Massachusetts candidates who intended to seek licensure were required to complete all four PAL tasks, regardless of preparation pathway.

In all, 769 candidates enrolled, but only 477 submitted all four tasks by the May 15, 2015 deadline, as shown in Table 18. Some candidates enrolled during the last week and submitted no tasks (23 percent) and some (15 percent) submitted fewer than four tasks. Only those who submitted all four tasks were included in the Field Trial, and all four of their tasks were scored.

Table 18

*Number and Percentage of Candidates Who Participated in the Field Trial by the Number of Tasks Submitted*

|  |  |  |
| --- | --- | --- |
| Number of Tasks Completed | Number of Candidates | Percent of Candidates |
| 4  | 477 | 62.0% |
| 3  | 9 | 1.2 |
| 2  | 36 | 4.5 |
| 1  | 68 | 9.0 |
| None | 179 | 23.3 |
| Total | 769 | 100% |

We reviewed all task submissions for completion, the minimum expectation for licensure recommendation during the Field Trial. A few submissions did not meet the minimum expectations (as defined by a level 1 on the scoring rubrics) and were returned for resubmission, and a few submissions were removed for other reasons. As a result, 422 candidates (88 percent of the complete submissions) were included in the Field Trial.

Table 19 presents a summary profile of the candidates who submitted all four tasks of the PAL assessment system that were usable for the Field Trial. As shown, most candidates were from a preparation program, female, and White (although 16 percent preferred not to identify their race/ethnicity).

Table 19

*Number and Percentage Distribution of Candidates Who Completed All Four Tasks for the PAL Field Trial, by characteristic*

|  |  |  |
| --- | --- | --- |
| Candidate Characteristics  | Number of Candidates | Percent of Candidates |
| Candidate Preparation Pathway |  |  |
| Preparation program | 347 | 82% |
| Alternative pathway (administrative internship or panel review) | 75 | 18 |
| Gender |  |  |
| Female  | 272 | 65 |
| Male | 150 | 35 |
| Race/Ethnicity and National Origin |  |  |
| White | 314 | 74 |
| African American | 14 | 3 |
| Hispanic | 11 | 3 |
| Asian | 9 | 2 |
| Native Hawaiian, Pacific Islander | 2 | 1 |
| Native American | 1 | \* |
| Multi-race, non-Hispanic | 15 | 1 |
| Preferred not to answer | 66 | 16 |
| Total Number of Candidates | 422 |  |

\*Less than one percent.

Construct Validity: Evaluations of the Indicators and Domains by Task

To evaluate the construct validity of the PAL assessment system, we examined the scoring at four levels: the quality of each indicator as measures of specific aspects of initial school leader knowledge and skills; the relationship among the indicator scores as a coherent measure for their respective domain within a task-defined area of leadership knowledge and skills; the relationship among the domain scores as a coherent measure for their respective task-defined area of leadership knowledge and skills; and 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. The overarching questions of this investigation concern the quality of the measures to distinguish candidate performance levels while capturing candidates’ sufficiently unique dimensions of initial school leadership practice within and among each task.

We began by evaluating the quality of the indicators and their relationship to each other within the domains for each task. We evaluated the quality of each indicator by calculating the indicator scores and examining their descriptive statistics (means and ranges) for score variance. Indicator measures should have sufficient variance to show they are distinguishing candidate performance and should have somewhat similar means to each other to show they are gauging similar levels of candidate performance.

To determine the domains’ measurement quality, we calculated their correlations within each task. Such correlations should show a sufficient degree of association to demonstrate a moderate to strong relationship, while not so strong as to suggest that the different domains are nevertheless measuring the same dimensions of leadership behavior.

Finally, we investigated the relationship among the domains by conducting a factor analysis to determine the extent to which the domains loaded together as distinct factors, demonstrating their relationship to each other within one factor and distinctiveness from other domain measures within other factors.

Task 1 Indicators and Domains

Table 20 shows the mean and standard deviations for the six indicators and three domains in Task 1. The means for the indicators were somewhat similar, ranging between only 2.70 and 2.80 on the 4-point scale, and their standard deviations were comparable. These findings show that the indicator measures are performing similarly.

Table 20

*Task 1 Indicators and Domains*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Task | N | Minimum Score | Maximum Score | Mean | Std. Deviation |
| T1D1 Investigate and prepare a vision | 421 | 1.0 | 4.0 | 2.75 | .626 |
| T1D1a Data collection | 420 | 1.0 | 4.0 | 2.76 | .646 |
| T1D1b Data analysis and priority definition | 415 | 1.0 | 4.0 | 2.74 | .676 |
| T1D2 Design an integrated plan for strategies | 421 | 1.0 | 4.0 | 2.73 | .633 |
| T1D2a Vision and plan focus | 421 | 1.0 | 4.0 | 2.75 | .687 |
| T1D2b Plan details | 417 | 1.0 | 4.0 | 2.70 | .679 |
| T1D3Assess and analyze feedback from candidates | 421 | 1.0 | 4.0 | 2.80 | .647 |
| T1D3a Plan feedback | 418 | 1.0 | 4.0 | 2.80 | .676 |
| T1D3b Assessment of leadership skills and practices | 420 | 1.0 | 4.0 | 2.80 | .735 |
| T1Total (Average of Domain Scores) | 421 | 1.00 | 4.00 | 2.76 | .553 |

*Key*: T=task (1-4), D=domain (1-4), the alphanumerics (a-d) =indicator

Table 21 shows the correlations between the domain scores in Task 1. All correlations were statistically significant and, of them, scores on Domain 2 were most highly correlated with the other two domains. The positive correlations between domains support the intended design of the assessment. No correlations between task measure constructs are above a .9, implying that the constructs measured are related but not identical, also supporting the intended design.

Table 21

*Correlations between all Task 1 Domain Scores*

|  |  |  |  |
| --- | --- | --- | --- |
| Measure | T1D1 Investigate and Prepare a Vision | T1D2 Design an Integrated Plan for Strategies | T1D3Assess and Analyze Feedback from Candidates |
| T1D1 Investigate and Prepare a Vision | 1.000 | -- | -- |
| T1D2 Design an Integrated Plan for Strategies | .681 | 1.000 | -- |
| T1D3Assess and Analyze Feedback from Candidates | .594 | .645 | 1.000 |

*Note*: All correlations significant at the *p*<.01 level.

Task 2 Indicators and Domains

Table 22 shows the mean and standard deviations for the six indicators and three domains in Task 2. The means for the six indicators were fairly similar, ranging slightly between 2.86 and 3.04 on the 4-point scale. The standard deviations for the indicators were fairly similar as well.

Table 22

*Task 2 Indicators and Domains*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Measure | N | Minimum Score | Maximum Score | Mean | Std. Deviation |
| T2D1 Plan to facilitate team learning | 419 | 1.0 | 4.0 | 2.98 | .557 |
| T2D1a Team identification | 417 | 1.0 | 4.0 | 2.98 | .627 |
| T2D1b Team learning plan | 418 | 1.0 | 4.0 | 2.99 | .682 |
| T2D2 Enact a professional learning culture | 420 | 1.0 | 4.0 | 2.95 | .615 |
| T2D2a Team process | 416 | 1.0 | 4.0 | 2.86 | .687 |
| T2D2b Team learning and work | 415 | 1.0 | 4.0 | 3.04 | .676 |
| T2D3 Assess team learning to improve ongoing group learning | 419 | 1.0 | 4.0 | 2.98 | .661 |
| T2D3a Assessment of team process and team work | 417 | 1.0 | 4.0 | 2.93 | .727 |
| T2D3b Assessment of leadership skills and practices | 418 | 1.0 | 4.0 | 3.02 | .746 |
| T2Total (Average of domain scores) | 418 | 1.2 | 4.0 | 2.97 | .529 |

Table 23 shows the correlations between the domain scores for Task 2. Like Task 1, all correlations were statistically significant and, of them, scores on Domain 2 were most highly correlated with the other scores. The positive correlations between domains support the intended design of the assessment. No correlations between domains are above a .9, implying that the constructs measured are related but not identical, also supporting the intended design.

Table 23

*Correlations between All Task 2 Domain Scores*

|  |  |  |  |
| --- | --- | --- | --- |
| Measure | T2D1 Plan to Facilitate Team Learning | T2D2 Enact a Professional Learning Culture | T2D3 Assess Team Learning to Improve Ongoing Group Learning |
| T2D1 Plan to Facilitate Team Learning | 1.000 | -- | -- |
| T2D2 Enact a Professional Learning Culture | .587 | 1.000 | -- |
| T2D3 Assess team learning to improve ongoing group learning | .551 | .729 | 1.000 |

*Note*: All correlations significant at the *p*<.01 level.

Task 3 Indicators and Domains

Table 24 shows the mean and standard deviations for the eight indicators and four domains in Task 3. The means of the eight indicators were fairly similarly, ranging slightly between 2.77 and 3.02 on the 4-point scale, as were their standard deviations.

Table 24

*Task 3 Indicators and Domains*

| Measure | N | Minimum Score | Maximum Score | Mean | Std. Deviation |
| --- | --- | --- | --- | --- | --- |
| T3D1 Plan | 421 | 1.0 | 4.0 | 2.90 | .511 |
| T2D1a Observation focus selection | 420 | 1.0 | 4.0 | 2.92 | .535 |
| T3D1b Pre-observation conference | 416 | 1.0 | 4.0 | 2.89 | .557 |
| T3D2 Conduct the Observation | 421 | 1.0 | 4.0 | 2.85 | .511 |
| T3D2a Use and application of teacher observation rubric | 416 | 1.0 | 4.0 | 2.82 | .608 |
| T3D2b Description of observations | 418 | 1.0 | 4.0 | 2.87 | .545 |
| T3D3 Provide Feedback and Suggest Support | 421 | 1.0 | 4.0 | 2.91 | .493 |
| T3D3a Feedback content | 421 | 1.0 | 4.0 | 2.90 | .548 |
| T3D3b Rapport and teacher engagement | 417 | 1.0 | 4.0 | 3.02 | .530 |
| T3D3c Teacher development | 416 | 1.0 | 4.0 | 2.81 | .625 |
| T3D4 Assess: Analyze and Identify Implications | 416 | 1.0 | 4.0 | 2.77 | .715 |
| T3D4a Assess leadership skills and practices | 416 | 1.0 | 4.0 | 2.77 | .715 |
| T3Total (Average of Domain Scores) | 416 | 1.00 | 4.00 | 2.86 | .472 |

Table 25 shows the correlations between the domain scores for Task 3. Like Tasks 1 and 2, all correlations were statistically significant and, of them, scores on Domain 3 were most highly correlated with the total score for this task. The positive correlations between domains support the intended design of the assessment. No correlations between domains are above a .9, implying that the constructs measured are related but not identical, also supporting the intended design.

Table 25

*Correlations between All Task 3 Domain Scores*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Measure | T3D1 Plan | T3D2 Conduct the Observation | T3D3 Provide Feedback and Suggest Support | T3D4 Assess: Analyze and Identify Implications |
| T3D1 Plan | 1.000 | -- | -- | -- |
| T3D2 Conduct the Observation | .664 | 1.000 | -- | -- |
| T3D3 Provide Feedback and Suggest Support | .675 | .714 | 1.000 | -- |
| T3D4 Assess: Analyze and Identify Implications | .527 | .549 | .612 | 1.000 |
| *Note*: All correlations significant at the *p*<.01 level. |

Task 4 Indicators and Domains

Table 26 shows the mean and standard deviations for the six indicators and three domains in Task 4. The means of the six indicators were fairly similar, ranging slightly between 2.60 and 2.69 on the 4-point scale, as were their standard deviations, although showing the greatest range among the four tasks.

Table 26

*Task 4 Indicators and Domains*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Measure | N | Minimum Score | Maximum Score | Mean | Std. Deviation |
| T4D1 Plan to promote family and community involvement | 420 | 1.0 | 4.0 | 2.67 | .680 |
| T4D1a Investigation of the priority area | 417 | 1.0 | 4.0 | 2.69 | .697 |
| T4D1b Investigation of work group engagement | 420 | 1.0 | 4.0 | 2.65 | .854 |
| T4D1c Preparation of the plan, including strategies | 419 | 1.0 | 4.0 | 2.67 | .747 |
| T4D2 Implement an engagement or involvement strategy | 418 | 1.0 | 4.0 | 2.68 | .801 |
| T4D2a Implementation of the strategy | 418 | 1.0 | 4.0 | 2.67 | .801 |
| T4D3 Analyze feedback from others and candidate assess leadership skills | 420 | 1.0 | 4.0 | 2.64 | .770 |
| T4D3a Assessment and analysis of feedback | 419 | 1.0 | 4.0 | 2.67 | .840 |
| T4D3b Assessment of leadership skills and practices | 416 | 1.0 | 4.0 | 2.60 | .867 |
| T4Total (Average of domain scores) | 418 | 1.0 | 4.0 | 2.66 | .685 |

Table 27 shows the correlations between the domain scores for Task 4. Like the other three tasks, all correlations were statistically significant and of them, scores on Domain 2 were most highly correlated with the other domain scores. The positive correlations between domains support the intended design of the assessment. No correlations between task measure constructs are above a .9, implying that the constructs measured are related but not identical, also supporting the intended design.

Table 27

*Correlations between All Task 4 Domain Scores*

|  |  |  |  |
| --- | --- | --- | --- |
| Measure | T4D1 Plan to Promote Family and Community Involvement | T4D2 Implement an Engagement or Involvement Strategy | T4D3 Analyze Feedback from others; Assess Leadership Skill |
| T4D1 Plan to promote family and community involvement | 1.000 | -- | -- |
| T4D2 Implement an engagement or involvement strategy | .772 | 1.000 | -- |
| T4D3 Analyze feedback from others; assess leadership skills | .726 | .752 | 1.000 |

\*All correlations significant at the *p*<.01 level.

Dimensionality Analysis of Domain and Total Scores

We conducted an exploratory factor analysis on the correlation matrix among all domain scores.[[2]](#footnote-2) These analyses help provide insight into the number and nature of underlying dimensions measured by the submissions. An evaluation of scree plots and parallel analysis pointed towards a 4-factor structure underlying the domain scores. A 4-factor model was fit to the correlation matrix among domain scores, using maximum likelihood (ML) estimation and promax rotation (allowing the factors to be correlated). We were looking for a strong association among the domains within one task and a poor association with the other three tasks to demonstrate the strength of their relationship as measures within each task and their uniqueness, by contributing little as measures of the other tasks.

In our resulting analysis, the four factors corresponded perfectly to the four tasks, with each domain score having a high positive loading on its associated task factor and near zero loadings on other task factors. The results are presented in Table 28.

Table 28

*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 boldfaced to highlight their factor alignment.

*Key*: T=task (1-4), D=domain (1-4).

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. Thus, the four tasks are capturing different dimensions of school leadership practice.

Finally, using the factor loadings, we calculated the correlations between factors, as shown in Table 29. The correlations among factors were moderate and positive, ranging from 0.32 to 0.42 across the four factors. The moderate and consistent positive correlations among task factors suggest a 1-factor model could be fit to the task scores and would fit the data well but would leave considerable variance in task scores unexplained by the single factor. This finding confirms our use of four measures, one for each task.

Table 29

*Factor Correlation for Four Factors*

|  |  |
| --- | --- |
| Factors | Factors |
| F1 | F2 | F3 | F4 |
| F1 | 1.00 | -- | -- | -- |
| F2 | 0.36 | 1.00 | -- | -- |
| F3 | 0.38 | 0.39 | 1.00 | -- |
| F4 | 0.35 | 0.42 | 0.32 | 1 |

Total Scores

Next, we evaluated the quality of the task scores and the total score by looking at the score distributions and the relationship among the task scores.

Table 30 shows the means and standard deviations for the scores of each of the four tasks. While the means were fairly similar, their standard deviations ranged somewhat differently from .472 for Task 3 to .685 for Task 4.

Table 30

*Descriptive Statistics by Task*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Task | Number | MinimumScore | MaximumScore | Mean | Standard Deviation |
| 1 | 421 | 1.00 | 4.00 | 2.76 | .553 |
| 2 | 418 | 1.17 | 4.00 | 2.97 | .529 |
| 3 | 416 | 1.00 | 4.00 | 2.86 | .472 |
| 4 | 418 | 1.00 | 4.00 | 2.67 | .685 |

*Note*: The total score for each task is based on averaging the domain scores.

To evaluate score distribution further, we generated task score histograms and reviewed each for score cluster patterns. We found, as shown below, that the scores center around one level (3) of the rubric for each task. Such results may imply that the rubric levels lead to most submissions fit the description of a level 3, that the training leads scorers to assign a 3 with more frequency (and thus training needs to strengthen how well scorers made fine distinctions between score points) or that the rubrics need to be revised for greater differentiation.

Each task’s histogram is discussed separately below.

Figure 2 is a histogram that illustrates the score distribution for Task 1 within the 4-point range. It shows that scores were clustered around a 3, with some candidates scoring at every level of the scale. In addition, 69 percent of the submissions received a 3 as their total score, 3 percent received a 1, 20 percent received a 2, and 7 percent received a 4.

**

*Figure 2*. Distribution of total score for Task 1 (average of domain scores).

*N*ote: T1OverallcompdbyDomain=total score

Figure 3 is a histogram that illustrates the score distribution for Task 2 within the 4-point range. Like Task 1, the histogram shows scores were clustered around a 3, with some candidates scoring at every level of the scale. Of them, 69.2 percent of the submissions received a 3 as their total score. 1.4 percent received a 1, 12.6 percent received a 2, and 14 percent received a 4.



*Figure 3*. Distribution of total score for Task 2 (average of domain scores).

Figure 4 is a histogram that illustrates the score distribution for Task 3 within the 4-point range. As for Tasks 1 and 2, the histogram shows that scores were clustered around a 3, with some candidates scoring at every level of the scale. Of them, 76.3 percent of the submissions received a 3 as their total score, 1.4 percent received a 1, 10.4 percent received a 2, and 6.4 percent received a 4.



*Figure 4*. Distribution of total score for Task 3 (average of domain scores).

Figure 5 is a histogram that illustrates the score distribution for Task 4 within the 4-point range. Like Tasks 1, 2, and 3, the histogram shows that scores were clustered around a 3, with some candidates scoring at every level of the scale. Of them, 59.7 percent of the submissions received a 3 as their total score, 7.8 percent received a 1, 22.7 percent received a 2, and 7.8 percent received a 4.



*Figure 5*. Distribution of total score for Task 4 (average of domain scores).

Next, we correlated the task scores by averaging the domain scores in each task. Like our correlations of the indicators, we were looking for a modest correlation to support their relationship as an overall measure of leadership, but not so strong as to suggest a lack of independence. Table 31 shows the correlations between the four total task scores (computed by averaging the domain scores within that task). These results are consistent with the factor correlations in Table 29.

Table 31

*Correlations of the Average Total Domain Scores by Task*

|  |  |
| --- | --- |
| Measure | Measure |
| T1Total | T2Total | T3Total | T4Total |
| T1Total | 1.000 | -- | -- | -- |
| T2Total | .350 | 1.000 | -- | -- |
| T3Total | .320 | .315 | 1.000 | -- |
| T4Total | .284 | .340 | .267 | 1.000 |

 N of candidates=422

While the scores on the four tasks were significantly correlated (p<.01), the magnitude of the correlations was weak, ranging from .267 (Task 3 and Task 4) to .35 (Task 1 and Task 2). Given these results, we concluded that the correlations between tasks support the intended design of the assessment, indicating that that the tasks measure related but distinct constructs.

These findings support the results above. They demonstrate moderate correlations among task scores that indicate that a single primary dimension is being measured by the portfolio, but that there are unique aspects captured by each task. When interpreting these results, it is important to note that scorer effects, which could induce an inflated correlation among domain scores from the same task, could drive some part of the factor structure.

Total Score Comparisons by Candidate Characteristics

Next, we investigated possible independent influences on candidate scores by evaluating whether there was gender or race/ethnicity bias in candidate scores or preparation pathway differences in candidate scores. The performance of candidates was compared by subgroups in three demographic categories using the average domain total score: type of preparation pathway (preparation program or alternative certification), gender (male or female), and ethnicity (White, African American, Hispanic, or Asian). The results in Table 32, based on analysis of variance (using an F-test), show statistically significant differences in performance by gender, but not by preparation pathway. Due to missing information and low numbers, the comparisons by race/ethnicity can only be viewed as trends. These results should be monitored in the current program year and investigated further for possible explanations.

Table 32

*Mean and Standard Deviation for Domain Total Score by Demographic Attribute*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Candidates Demographics  | Number | Mean | Standard Deviation | Statistical Significance |
| Preparation Pathway |  |  |  | Not statistically significant, F (1,420) = 1.47, p =.226 |
| Preparation program | 247 | 2.84 | .39 |  |
| Alternative pathway | 75 | 2.78 | .39 |  |
| Gender |  |  |  | statistically significant, F (1,420) = 5.60, p =.018 |
| Female | 272 | 2.89 | .40 |  |
| Male | 150 | 2.79 | .40 |  |
| Race/Ethnicity |  |  |  | Not calculated |
| White | 314 | 2.83 | .38 |  |
| African American | 14 | 2.79 | .33 |  |
| Hispanic | 11 | 2.78 | .58 |  |
| Asian | 9 | 2.82 | .22 |  |

Scoring Reliability

Scorer Agreement

Next, we evaluated scoring reliability using submissions that were scored by two scorers to determine scorer agreement. 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. $κ\_{n}$ compares the observed agreement rate to what we would expect by chance if scorers randomly assigned scores of 1-4 to each submission, showing that scorers are using the full score range in the scoring process. 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.[[3]](#footnote-3)

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

Total agreement rates, as shown in Table 33, indicate that scorers scoring the same submission assigned the same or adjacent scores over 90 percent of the time for all but one rubric. On a 4-point scale, instances of scorers giving discrepant scores (1 and 3, 2 and 4, 1 and 4) should be minimal or, ideally, nonexistent. However, the lack of variability across scorers in which a preponderance of scores is at the 3-point level may inflate the level of scorer agreement because of restrictions in range in scores within and across domains. This finding is consistent with the analyses summarized above that show Task 3 as having very little variability in scores. These analyses suggest review and revision of the design of the rubric levels and the training of scorers for Task 3.

Table 33

*Scorer Agreement Rates by Indicator*

| Task | Indicator | Exact | Kappa (exact) | Exact + Adj | Kappa (Exact+Adj) | 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 |   |

Preliminary G Study Analyses

The next step in the reliability analysis is conducting a generalizability analysis, or G study, which would provide estimates of the variability of measures (Shavelson & Webb, 2005). Although data for double-scored task submissions were not systematically collected to allow a single, comprehensive G study, parts of the data were used to conduct small G studies to estimate the reliability of each task score. The preliminary G studies are used to estimate the reliability of task scores computed as the sum of the average domain scores within each task.[[4]](#footnote-4)

To conduct the G studies, we undertook the following process for each task separately. First, blocks of portfolios that were scored by the same two scorers were identified in the data. Within each of these blocks a completely crossed two-facet G study, with scorer and domain facets, was conducted by estimating the relevant variance components. Variance components were estimated for candidates, scorer, and domain effects as well as all two-way interactions and a combined three-way interaction/error term. A sample-size weighted average of each variance component was computed across the five blocks, and these average variance components were used to estimate a relative G coefficient for each task, treating scorers as random and domains as fixed facets.[[5]](#footnote-5)

The Table 34 presents the estimated relative G coefficients (similar to a reliability coefficient in classical test theory) for each task, when task scores are calculated as the sum of domain scores. Estimated coefficients assuming either 1, 2, or 3 scorers for each task are presented (with 2 scorers, this is the reliability of an average task score across two independent scorers).

Table 34

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

|  |  |
| --- | --- |
| Task | Number of Scorers |
| 1 | 2 | 3 |
| 1 | 0.728 | 0.842 | 0.889 |
| 2 | 0.656 | 0.792 | 0.851 |
| 3 | 0.208 | 0.345 | 0.441 |
| 4 | 0.581 | 0.735 | 0.806 |

As seen here, the reliability looks substantially better for Tasks 1, 2, and 4 and problematically low for Task 3. This is likely due to the very low variance between candidates on Task 3 performance. The reliability coefficients are moderate for all tasks with one scorer, but are all above 0.80 for Tasks 1, 2, and 4 with three scorers.

To further evaluate the scores, we assessed the potential reliability of a total portfolio score computed as the sum of all four task scores. Because each task is scored by an independent scorer, the four task scores provide independent information about the candidate. We can estimate the reliability of the total score using a stratified coefficient alpha, information about the variance of the domain, task, and total scores as well as the estimated reliability coefficients.

Stratified alpha is 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 (Haertel, 2006) ). In the case here, domain scores are drawn from different “categories” (the tasks), but the four tasks overall are intended to measure the same primary construct, and the goal is to estimate the reliability of the summed total score across the four tasks.

Using the relative G coefficients reported in Table 34 and the observed variance in task scores (where task scores are computed as the average of the domain scores), the measurement error for task $i$ was estimated as:

$$SEM\_{i}^{2}=\left(1-Rel\_{i}\right)\*Var\left(X\_{i}\right)$$

where $Rel\_{i}$ is the estimated G coefficient for task $i$ and $Var\left(X\_{i}\right)$ is the observed variance for scores on task $i$. [[6]](#footnote-6) Then, using a version of the formula for stratified coefficient alpha in Haertel (2006, p. 77) we estimated the reliability of the total score as:

$$Rel\_{Total}=1-\left(\frac{\sum\_{i=1}^{4}SEM\_{i}^{2}}{Var\left(Y\right)}\right)=1-\left(\frac{SEM\_{Total}^{2}}{Var\left(Y\right)}\right)$$

where $Var\left(Y\right)$ is the observed variance in total scores, where total scores are computed as the sum of the four task scores. This formula is appropriate because independent raters score each task, and hence we can assume that the measurement errors are independent across tasks. We can repeat this process replacing the estimated reliability of each task, $Rel\_{i}$, with the estimated reliability when either 1, 2, or 3 independent raters score each portfolio. The resulting stratified coefficient alpha estimates are in the Table 35. The results show that there is sufficient reliability with two scorers.

Table 35

*Reliability of Total Score Based on Number of Scorers for Each Task*

|  |  |
| --- | --- |
| Rel. Coef. | Number of Scorers |
| 0.771 | 1 |
| 0.844 | 2 |
| 0.879 | 3 |

Technical Advisory Committee Feedback

As a first step in the process of setting standards for passing scores on the PAL assessment, the PAL technical advisory committee was convened to review the validity and reliability evidence, presented above, and make recommendations for establishing the PAL performance standards. The committee, comprised of national psychometric experts, reviewed the content validity evidence and statistical analyses on the Field Trial scores (see Appendix G for members of this committee). In a meeting in October 2015, committee members discussed the evidence, drew conclusions, and made several recommendations.

Their primary conclusion was to confirm the statistical report findings that indicated that the four tasks were independent measures of leadership readiness and worked well together as one overall leadership readiness measure. Their scoring recommendations were these: a composite score be created for the four task scores based on averaging the indicator scores within each task, and a threshold score for each task defining minimally sufficient task scores for inclusion in the total composite score be determined. The technical advisory committee also recommended charging the standards setting committee with recommending the composite score and threshold score levels. Their other recommendations were to strengthen the scoring by reviewing the rubrics to refine score gradation differentiation and scorer training and oversight, to have two scorers score all submissions, and to have a third scorer review submissions with significant scoring differences.

Setting the Qualification Standards for Readiness for Initial School Leadership Licensure

Acting on the recommendations of the technical advisory committee for establishing qualification standards, we convened a standards setting committee in November 2015 to determine the task score components for the composite score and the threshold score for each task. In all, 15 Massachusetts school and district leaders and leadership preparation program faculty were recruited to serve as part of the standards setting process (see Appendix H for members of this committee). All but one committee member—an individual who is also a member of the design committee—had been trained as a scorer for one or two PAL tasks. For the standards setting process, the committee divided into four groups, each charged with setting standards for one task. To prepare for the standards setting meeting, each committee group was given a set of materials to review: The Field Trial (October 2014) Scoring Guide for their assigned task, the standards setting process description (which outlined the steps for standards setting), and the *Candidates Assessment Handbook* (for the Field Trial). They were asked to review the Scoring Guide, focusing on the difference between the criteria for level 3 performance and level 2 performance for each task indicator, since the likely location of the cut score would be between these two levels.

In addition, prior to the meeting, in an online folder, the committee members were given two sample submissions (for their task assignment) and were asked to summarize the evidence, using an indicator-based chart, that they found in the submission artifacts and commentary to support a conclusion that the indicator could be scored as a level 2 or level 3. The two samples had been scored by the assessment development staff, but these scores were not shared with the committee members until the meeting discussions took place. During the first part of the meeting, each group used the two sample submissions they had scored to discuss how to distinguish among submissions that provide evidence that the candidate is “ready” to assume a school-level leadership role, is approaching readiness, or is not yet ready. Based on this discussion, the group arrived at consensus on a descriptor of what “readiness” looks like for the task. (See Appendix I for the final descriptors for the four tasks.)

Next, the groups were given an ordered packet of five performance profiles submissions scored at various levels between levels 2 and 3. The members first worked independently and then as a group to make a judgment about each profile and answer the question: Has the candidate demonstrated that he /she has developed the necessary leadership skills (minimally met the standards) and is “ready” to do this work as a new school principal? Individually and then collectively, the group members for each task rank-ordered all five submissions on a continuum of readiness to non-readiness and selected the submission whose score reflected what they believed to be the minimum level for “readiness.” After discussion, group members had the opportunity to revise their ratings and then the final ratings were charted. The PAL development team staff presented the groups with impact data that identified the percentage of candidates from the Field Trial who would have been identified as “ready” for the given submission score. Based on these results, each group had a final discussion about the score that distinguished readiness from non-readiness.

Following this discussion, the groups were asked to make a recommendation of what the lowest acceptable score would be for this task (i.e., for a candidate’s task score to be included in calculating a total score). This minimum acceptable score would become the threshold score for each task; candidates’ work would have to achieve at least this minimum score for the results to be incorporated in calculating their total scores.

We shared the standards setting committee recommendations and impact results (based on the Field Trial scores) within one and two standard deviations with Massachusetts ESE staff. The ESE commissioner, based on the recommendations of the PAL technical advisory and standards setting committees, then established the performance standards for PAL, as follows (and as shown in Table 36): first, all completed work must yield a minimum threshold score for each task of 2.1 (on a 4.0 scale). A threshold score is a minimum performance expectation for a score to be included as part of a total score. Second, the performance standard for the four tasks is a candidate’s total average score of all four tasks. Candidates must meet or exceed a passing score of 2.5 (on a 4-point scale) for Program Year 2015-16 and a passing score of 2.75 for Program Year 2016-17. The total PAL score is based on an average of the overall scores for each of four tasks and requires that at least some task-specific scores are higher than the threshold score. Candidates’ work that only meets the threshold scores (2.1) for each of the four tasks will not obtain a passing score of 2.5 or above for the assessment.

Table 36

*PAL Threshold and Total Passing Scores used as the PAL Performance Standard*

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Threshold Score Per Task | Total Passing Score 2015-16 | Total Passing Score 2016-17 |
| 1 | 2.1 |  |  |
| 2 | 2.1 |  |  |
| 3 | 2.1 |  |  |
| 4 | 2.1 |  |  |
| Total Score |  | 2.5 | 2.75 |

Candidates will be required to revise and resubmit work products for each task-specific submission that is scored below the threshold score. Candidates whose total PAL score falls below the total passing score of 2.5 may select which tasks to revise and resubmit to improve their total PAL score performance.

Conclusions

The Massachusetts Performance Assessment for Leaders (PAL) assessment is the first validated performance assessment of candidates for initial school leadership readiness. It consists of four, field-based performance tasks that allow candidates to demonstrate their leadership knowledge and skills in planning for school improvement, facilitating a professional learning group, observing and giving feedback to a teacher, and engaging families and the community in improving student learning.

PAL was created within a technical framework of psychometric practice and principles, guided by *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Its development is the result of a collaborative project between the Massachusetts Department of Elementary and Secondary Education (ESE and its partner contractor, Bank Street College of Education, which engaged the involvement of a team of national experts on leadership and performance assessments including Massachusetts K-12 school and district leader and higher education faculty.

Content validity studies of the tasks conducted by the developers demonstrate that PAL is aligned with state and national leadership standards and indicators and strong reflects the work required of initial school leaders. These results were further confirmed by face validity studies of candidates and higher education faculty feedback, who also confirmed that the assessment was feasible to perform and flexible for a variety of school and district contexts.

A Field Trial was conducted in 2014-2015 to analyze the performance of a selected group of candidates on the tasks and test the evaluation process. Thirty experienced K-12 school and district leaders and higher education faculty were trained and certified in a rigorous 10-15-hour process to score candidates’ PAL task submissions. Scorers applied 13 rubrics of 26 indicators to the four tasks, using a 4-point scale to rate the quality of a candidate’s performance on each task. The score for each task ranged from 1 to 4; the four task scores were averaged for a total score ranging from 1 to 4. The scoring process was coordinated by ShowEvidence, which managed the electronic platform for candidates’ work submissions and the scoring and reporting processes.

Construct validity studies, based on the Field Trial results, confirmed that the task scores measured primary characteristics of effective initial school leader practice and work well as a combined measure of leadership readiness. The Field Trial inter-rater reliability evaluation was .84 (as the percentage of agreement for two scorers). The Field Trial yielded 422 complete submissions that were analyzed by preparation pathway and gender. Candidates scored slightly higher on Task 2, Instructional Leadership for a Professional Learning Culture (2.97), and slightly lower on Task 4, Leadership for Family Engagement and Community Involvement (2.67) than on the other two tasks (Task1, Leadership through a Vision for High Student Achievement, and Task 3, Leadership in Observing, Assessing, and Supporting Individual Teacher Effectiveness). Candidate scores were not statistically significantly different based on preparation pathway, but female candidates scored higher, statistically, than male candidates. Subgroup samples were too small for score comparisons based on race/ethnicity.

Standards setting, conducted in November 2015, used a standards-based consensus process with four separate panels of experts (one per task) from K-12 school and district leadership and higher education. Based on the 2015 Field Trial data, a cut score—representing candidate readiness for initial leadership-- of 2.5 overall (based on an average of the four task scores) was established for 2015-16, and included a threshold score –representing the minimum acceptable score for each task-- 2.1 for each task.

 Based on our findings, we concluded that the PAL’s development process yielded important evidence of the assessment’s validity, reliability, and feasibility for the purposes of determining licensing new school leaders in Massachusetts. Indeed, PAL is highly relevant for demonstrating that candidates entering initial school leader positions are ready to provide effective leadership schools and improve student learning. Finally, integrating PAL into their work with candidates will enable preparation programs to guide and support the candidates as they master the leadership skills and practices necessary for a principalship or assistant principalship.

References

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

Haertel, E. H. (2006). Reliability. In R. Brennan (Ed.), *Educational measurement (4th ed.)* (pp. 65-110). Westport, CT.: American Council on Education and Praeger.

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.

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

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

Shavelson, R. J., & Webb, N. M. (2005). Generalizability Theory. In K. Kempf-Leonard (Ed.), *Encyclopedia of Social Measurement*: Elsevier.

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*(3), 197-210.

Appendix A:

PAL Design Committee Membership

Below are the committee members and their affiliation at the time they served.

**Massachusetts Educational Leaders**

Dr. Matthew Malone\*

Superintendent

Brockton Public Schools

Mary Cjzakowski

Superintendent

Barnstable Public Schools

Dr. Stacy Scott

Superintendent

Framingham Public Schools

Steve Murray

Principal

Andover West Middle School

Andover Public Schools

Mary Driscoll

Principal

Edison K-8 School

Boston Public Schools

**Massachusetts Leadership Preparation Pathways**

Dr. Anna Bradfield

Executive Director for University Initiatives

Bridgewater State University

Dr. Tom Payzant\*

Emeritus Faculty

Harvard University

Richard Pearson

Associate Director/Dir. of Prep Program

MSSAA (Massachusetts Secondary School Administrators Association) & MIAA (Massachusetts Interscholastic Athletic Association)

Dr. Rebecca Woodland

Associate Professor

University of Massachusetts-Amherst

\*Attended only one meeting.

Appendix B:

PAL Content Validity Committee Membership

Below are the committee members and their affiliation at the time they served.

**Massachusetts Educational Leaders**

Jason DiCarlo

Principal

Murkland Elementary School

Lowell Public Schools

Patty Barrett

Principal

Sanborn Elementary School

Andover Public Schools

Jeff Szymaniak

Principal

Whitman-Hanson Regional School

Whitman-Hanson School District

Daniel Gutekanst

Superintendent

Needham Public Schools

Brian Salzer

Superintendent

Northampton Public Schools

Karen McCarthy

Teacher/Literacy Facilitator

Brighton High School

Brighton Public Schools

**Massachusetts Leadership Preparation Programs**

Denise Fronius

Facilitator/Principal

Massachusetts Elementary School Principals Association

Jake Eberwein

Dean of Graduate and Continuing Education

Massachusetts College of Liberal Arts

Lee Tietel

Director, School Leadership Program

Harvard Graduate School of Education

Steve Gould

Assistant Professor

Lesley University

Appendix C:

Degree of Alignment between the Massachusetts Standards for Administrative Leadership by the PAL Assessment Tasks

| MA Leadership Standards/Elements | Task 1Vision/Direction | Task 2Professional School Culture | Task 3Individual Teacher’s Effectiveness | Task 4Family and Community Engagement |
| --- | --- | --- | --- | --- |
| Standard 1InstructionalLeadership | \*\*\*/\*\*/\* | \*\*\*/\*\* | \*\*\*/\*\*/\* | \*\*/\* |
| 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 3Family and Community Engagement | \*\*/\* | \* | \* | \*\*\*/\*\* |
| Family Engagement | \*\* | \* | \* | \*\*\* |
| Effective Communication | \*\* |  |  | \*\*\* |
| Advocacy | \* | \* |  | \*\* |
| Community Connections | \* | \* |  | \*\*\* |
| Cultural Awareness | \*\* | \* |  | \*\* |
| Standard 4Professional 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 | \*\*\* | \*\*\* | \*\* | \*\* |

\*\*\* Primary focus: Task directly assesses performance of indicator.

\*\* Secondary focus: Task requires performance of indicator to successfully complete.

\* Tertiary focus: Task requires knowledge of indicator (performance of the task “bumps into” indicator).

Appendix D:

PAL Bias Review Committee Membership

Below are the committee members and their affiliation at the time they served.

**Massachusetts Educational Leaders**

Ellen Stockdale

Principal

Maquan Elementary School

Whitman-Hanson Regional School District

Gerald Yung

Principal

Martin Luther King School

Cambridge Public Schools

Mark Wood

Principal-Director

Tantasqua Regional High School

Tantasqua School District

Melinda Boone

Superintendent

Worcester Public Schools

Isabelina Rodriguez

Superintendent

Granby Public Schools-Boston

**Massachusetts Leadership Preparation Pathways**

Lauri Johnson

Associate Professor

Boston College

Bennie Agbarha

Associate Professor

Bridgewater State University

Dan French

Executive Director

Center for Collaborative Education

Casel Walker

Supervisor-Leadership Candidates

University of Massachusetts

Appendix E:

Bias Review Topics and Questions

|  |  |
| --- | --- |
| 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? |
| 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? |

Appendix F:

PAL Task Domains, Elements, and Indicators

| Domain | Rubric/Domain | Indicator |
| --- | --- | --- |
| Task 1Leadership through a vision for high student achievement | Plan to create a vision | Data collectionData analysis and priority definition |
| Design an integrated plan to develop and implement improvement in the priority area | Vision and plan focusPlan details |
| Assess and analyze feedback from candidates | Plan feedbackPlanning analysis |
| Task 2Instructional leadership for a professional learning culture | Planning to facilitate team learning | Team identificationTeam learning plan |
| Foster a professional learning culture to support team learning | Team processTeam learning and work |
| Assess team learning to improve on-going group learning | Assessing team process and team workAssessing leadership skills and practices |
| Task 3Leadership in observing, assessing, and supporting individual teacher effectiveness  | Plan | Observation focus selectionPre-observation 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 |
| Assess: analyze and identify implications | Assessment of leadership skills and practices |
| Task 4Leadership for family engagement and community involvement   | Plan to promote family and community involvement | Investigation of a priority areaInvestigation of work group engagementPreparation of the plan, including strategies |
| Implement an engagement or involvement strategy | Implementation of the strategy |
| Analyze feedback from candidates and assess leadership skills | Assessment and analysis of feedback on the family and community engagement plan and strategy Assessment of leadership skills and practices |

Appendix G:

PAL Technical Advisory Committee Membership

Below are the committee members and their affiliation at the time they served.

Dr. Lloyd Bond

Carnegie Foundation for the Advancement of Teaching

Stanford, CA

 Dr. Gregory Cizek

University of North Carolina

Chapel Hill, NC

Dr. Andrew Ho

Harvard University

Cambridge, MA

Dr. Stuart Kahl

Founder, Measured Progress

Dover, NH

Appendix H:

PAL Standards Setting Committee Membership

| Committee Member | Preparation Program Affiliation and Role | K-12 Experience |
| --- | --- | --- |
| Sean Halpin |  | Plymouth Public SchoolsDirector of Student Support Services |
| Rose Ellis |  | Williamstown-Lanesborough/Mt. Greylock Regional School DistrictRetired superintendent |
| Eva Thompson | Educator Leadership Institute/Educational Collaborative (EDCO ELI) Leadership Program, Waltham, MA | Newton Public SchoolsRetired principal |
| Mary Driscoll |  | Boston Public SchoolsPrincipal |
| Marie Galinski | Merrimack College | Beverly Public SchoolsRetired superintendent |
| Janet Sullivan |  | Stoughton Public SchoolsK-12 STEM Curriculum Administrative supervisor |
| Patrick Fraine | Curry College | Plymouth Public SchoolsPrincipal |
| Elinor Freeman |  | Reading Public SchoolsRetired high school principal |
| Joe Russo | Harvard University | Former assistant superintendent, Newton, MA |
| David Thomson |  | Blackstone Millville Regional School DistrictAssistant superintendent |
| David Guglia |  | Stoughton Public SchoolsAssistant principal |
| Phitsamay Uy | UMass Lowell, Assistant professor |  |
| Tom LaLiberte | Massachusetts Secondary School Administrator Association (MSSAA)Program director | William H Galvin Middle School in Canton, MAFormer principal |
| Peg McKay | Curry CollegeProgram director | Plymouth Public SchoolsFormer principal |
| Lena Marie Rockwood |  | Revere Public SchoolsAssistant principal |

Appendix I:

Descriptors of Readiness for the Four PAL Tasks

**Task 1 Performance Descriptor – Readiness**

The Massachusetts Professional Standards for Administrative Leadership Task 1 reflects:

MA Standard 1. Instructional Leadership: Promotes the learning and growth of all students and the success of all staff by cultivating a shared vision that makes effective teaching and learning the central focus of schooling.

MA Standard 2. Management and Operations: Promotes the learning and growth of all students and the success of all staff by ensuring a safe, efficient, and effective learning environment, using resources to implement appropriate curriculum, staffing, and scheduling.

MA Standard 4: Professional Culture: Promotes success for all students by nurturing and sustaining a school culture of reflective practice, high expectations, and continuous learning for staff.

The candidate who is “ready” to enter his/her first school principal position and conduct the work the Task 1 standards require has demonstrated that he/she was able to:

* Collect, review, and analyze relevant quantitative and qualitative data, examining trends in student performance, school culture, and context, and school policies, programs, and practices; the data collection and analysis clearly informed the selection of the priority academic area and targeted student group, and clarified some of the factors that contributed to the group’s learning problems.
* Articulate a vision for increasing the targeted student group’s performance in the priority academic area with clearly defined goals and measurable student outcomes that tie back to the data; propose a plan containing relevant strategies, with sufficient details for implementation, to improve the school’s instructional program that will lead to increased student performance.
* Gather feedback from school leaders and other stakeholders about the relevance and feasibility of the plan to improving the performance of the student group and evaluate and use the feedback to make revisions to the plan; evaluate the strengths and weaknesses of his/her leadership skills relative to conducting this task and identify changes that would be made to improve his/her leadership practice.

**Task 2 Performance Descriptor – Readiness**

The Massachusetts Leadership Standards Task 2 reflects:

MA Standard 4. Professional Culture: Promotes success for all students by nurturing and sustaining a school culture of reflective practice, high expectations, and continuous learning for staff.

MA Standard 1. Instructional Leadership: Promotes the learning and growth of all students and the success of all staff by cultivating a shared vision that makes effective teaching and learning the central focus of schooling.

The candidate who is “ready” to enter his/her first school principal position and conduct the work that Task 2 standards require has demonstrated that he/she was able to:

* Select a purposeful group of teachers with an interest in improving teaching practice for a priority academic area, situating the team in the school’s existing professional culture for teacher teams; develop a team learning plan, which includes team learning structures to foster professional learning, with a focus on improving teaching practice and student performance in the priority academic area.
* Facilitate the group learning process over time, establishing and using shared norms for working as a group, making decisions, and implementing group learning practices and protocols; support teachers’ individual learning and the group’s collective work to improve teaching practice and student learning.
* Gather information about how the group learning process fostered individual teacher and group learning and accomplish the specific tasks the group took on to improve student performance in the priority academic area; evaluate the strengths and weaknesses of his/her leadership skills relative to conducting this task and identify changes that would be made to improve his/her leadership practice.

**Task 3 Performance Descriptor – Readiness**

The Massachusetts Leadership Standards Task 3 reflects:

MA Standard 2. Management and Operations: Promotes the learning and growth of all students and the success of all staff by ensuring a safe, efficient, and effective learning environment, using resources to implement appropriate curriculum, staffing, and scheduling.

MA Standard 1. Instructional Leadership: Promotes the learning and growth of all students and the success of all staff by cultivating a shared vision that makes effective teaching and learning the central focus of schooling.

MA Standard 4: Professional Culture: Promotes success for all students by nurturing and sustaining a school culture of reflective practice, high expectations, and continuous learning for staff.

The candidate who is “ready” to enter his/her first school principal position and conduct the work that Task 3 standards require has demonstrated that he/she was able to:

* Select a teacher to observe and an instructional focus for the observation based an analysis of common observed school practices, and teacher needs and student needs related to the focus; engage the teacher actively in a pre-observation discussion about common observed school practices from other classes related to the focus area, specific student or teacher needs, and/or student performance data.
* Use the district/state teacher observation rubric and conduct a classroom observation identifying the essential qualities of teaching practice concentrating on the instructional focus discussed in the pre-observation meeting; prepare observation notes that were sequential, contained descriptive comments about teacher and student actions, paying attention to the essential qualities of teaching practice, and including a summary judgment about the teacher’s practice.
* Provide the teacher with constructive improvement feedback, linking the feedback to the identified focus for the observation, observation data, the essential qualities of observed teaching practice and teacher and student needs; engage the teacher actively in a two-way conversation about the observed teaching during the post-observation meeting; identify resources to support the teacher in improving his/her practice.
* Evaluate the strengths and weaknesses of his/her leadership skills relative to conducting this task and identify changes that would be made to improve his/her leadership practice.

**Task 4 Performance Descriptor – Readiness**

The Massachusetts Leadership Standards Task 4 reflects:

MA Standard 3. Family and Community Engagement: Promotes the learning and growth of all students and the success of all staff through effective partnerships with families, community organizations, and other stakeholders that support the mission of the school and district.

MA Standard 4: Professional Culture: Promotes success for all students by nurturing and sustaining a school culture of reflective practice, high expectations, and continuous learning for staff.

The candidate who is “ready” to enter his/her first school principal position and conduct the work that Task 4 standards require has demonstrated that he/she was able to:

* Clearly define the role of family and community engagement for improving a student learning priority area and systematically describe existing information and data on school policies, programs and strategies for family engagement and community involvement in relation to the student learning priority; engage an appropriate, representative working group of school leaders, staff and family and community members in planning strategies to improvement family and community engagement; develop a complete, detailed plan that is responsive to family and community engagement in improving the student learning priority.
* Implement one strategy from the plan, identifying the steps, roles, resources, and challenges, and describe how the implementation of the strategy involved and impacted family or community members.
* Systematically gather feedback from work group members and use the feedback to determine the plan’s and strategy’s effectiveness; evaluate the strengths and weaknesses of his/her leadership skills relative to conducting this task and identify changes that would be made to improve his/her leadership practice.
1. As will be discussed in more depth later in the report, 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). [↑](#footnote-ref-1)
2. Analyses in this section are based on the n=361 portfolios with complete scores for all indicators from the primary scorer for the portfolio. 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 4-factor solution and the domain score correlation matrix was more appropriate to analyze. [↑](#footnote-ref-2)
3. Landis & 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). [↑](#footnote-ref-3)
4. Other possibilities include a G study with all indicators for a task, or a multivariate G study for each task, with indicators nested within domains. [↑](#footnote-ref-4)
5. Domains were considered fixed on conceptual grounds (they are not a sample of random equivalent domains, but are carefully selected and important domains of performance). Empirical results also supported treating this facet as fixed, given that the domain facet and the domain-by-candidate and domain-by-scorer facet interactions contributed a very small amount to variance in the scores. [↑](#footnote-ref-5)
6. All task and total score variances are based on the observed variances in the scores for the n=361 portfolios for which there was a valid score from the primary rater on all indicators, hence allowing computation of domain, task and total scores for a stable sample. [↑](#footnote-ref-6)