education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 26 Number 163 December 10, 2018 ISSN 1068-2341

The Validity of a Performance Based Assessment for Aspiring School Leaders

Jack Leonard
University of Massachusetts Boston
United States

Citation: Leonard, J. (2018). The validity of a performance based assessment for aspiring school leaders. *Education Policy Analysis Archives, 26*(163). http://dx.doi.org/10.14507/epaa.26.3914

Abstract: This paper introduces the new Massachusetts Performance Assessment for Leaders (PAL) and uses critical policy analysis to re-examine the validity evidence (using the 2014 Standards for Educational and Psychological Testing and a theory of multicultural validity) for the use and interpretation of the PAL in regards to emerging school leadership. Data sources include two years of PAL test documentation plus candidate surveys and interviews with program directors. The author's role as a test user, faculty instructor, and certified test scorer afforded access to student work, student communications, scorer network training, and state department of education communications and meetings. The paper challenges the content validity, raises questions in regards to evidence based on response processes, internal structure, relation to other variables, consequences, and multicultural validity particularly when the PAL is used as a stand-alone, high-stakes licensure test and offers suggestions to improve the test as a formative assessment.

Keywords: certification; critical policy analysis; educational administration; high stakes tests; performance assessment; principals; validity

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 6/6/2018
Revisions received: 10/9/2018
Accepted: 10/12/2018

La validez de una evaluación basada en el desempeño para aspirantes a líderes escolares

Resumen: Este artículo presenta la Massachusetts Performance Assessment for Leaders (PAL) y utiliza un análisis político crítico para reexaminar la validez de las evidencias utilizando los Estándares de 2014 para las Normas para la Educación y la Psicología de la prueba y una teoría de validez multicultural) el uso e interpretación de PAL en relación al liderazgo escolar emergente. Las fuentes de datos incluyen dos años de documentación de la prueba PAL, además de encuestas con candidatos y entrevistas con directores de programa. El papel del autor como usuario de prueba, instructor del cuerpo docente y artillero de prueba certificado posibilitó el acceso al trabajo del alumno, a las comunicaciones de los alumnos, al entrenamiento de la red de puntuación y al departamento estadual de comunicaciones y reuniones educativas. El artículo discute la validez de contenido, plantea cuestiones en relación a evidencias basadas en procesos de respuesta, estructura interna, relación con otras variables, consecuencias y validez multicultural, particularmente cuando el PAL se utiliza como una prueba autónoma de licencias de "high stakes" y ofrece sugerencias para mejorar la prueba como una evaluación formativa.

Palabras-clave: certificación; análisis crítico de políticas; administración educativa; "high stakes" pruebas; evaluación del desempeño; directores; validez

A validade de uma avaliação baseada no desempenho para aspirantes a líderes escolares

Resumo: Este artigo apresenta a Massachusetts Performance Assessment for Leaders (PAL) e usa análise política crítica para reexaminar a validade das evidências (usando os Padrões de 2014 para Standards for Educational and Psychological Testing e uma teoria de validade multicultural) para o uso e interpretação de PAL em relação à liderança escolar emergente. As fontes de dados incluem dois anos de documentação do teste PAL, além de pesquisas com candidatos e entrevistas com diretores de programa. O papel do autor como usuário de teste, instrutor do corpo docente e artilheiro de teste certificado possibilitou o acesso ao trabalho do aluno, às comunicações dos alunos, ao treinamento da rede de pontuação e ao departamento estadual de comunicações e reuniões educacionais. O artigo contesta a validade de conteúdo, levanta questões em relação a evidências baseadas em processos de resposta, estrutura interna, relação com outras variáveis, conseqüências e validade multicultural, particularmente quando o PAL é usado como um teste autônomo de licenciamento de "high stakes" e oferece sugestões para melhorar o teste como uma avaliação formativa.

Palavras-chave: certificação; análise crítica de políticas; administração educacional; "high stakes" testes; avaliação de desempenho; diretores; validez

The Validity of a Performance Based Assessment for Aspiring School Leaders

Certification is a distinct point in the K12 school leader preparation pipeline, which has garnered attention in recent years as states try to improve school leadership. Until recently, states had only a sit-down test to measure candidate readiness, such as the School Leader Licensure Assessment (SLLA) that is used in 21 states (Educational Testing Service, 2018). Connecticut offers a "more performance-based" assessment for school leadership (Darling-Hammond, Meyerson, LaPointe, & Orr, 2010, p. 163); the test is a one-day, sit-down test, which requires written responses to typical leadership scenarios (Educational Testing Service, 2018). In contrast, the Performance Assessment for Leaders (PAL), developed by Bank Street College and now employed by the Massachusetts Department of Elementary and Secondary Education (DESE), requires students to execute and report on four tasks during their administrative practicum (MA-DESE, 2017). This test is required for all principal licensure[1] in Massachusetts and is under consideration in other states such as California (MA-DESE, 2016c; California Commission on Teacher Credentialing, 2013).

Purpose

The purpose of this paper is to introduce the PAL and to consider the validity evidence for the use and interpretation of this kind of performance assessment in regards to emerging school leadership. According to the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational Psychological Testing, 2014), "Validity is…the most fundamental consideration in developing tests and evaluating tests" (p. 11). The PAL is a suitable "poster-child" for this paper on school leader performance assessments. California piloted aspects of the PAL in 2015–2016 so other versions may emerge (MA-DESE, 2016c). The questions raised here may apply equally well to future iterations of this test as well as other performance assessments for aspiring school leaders. The fact that the PAL is a high-stakes licensure test calls for an even closer consideration of validity, for the *Standards* (2014) state, "Higher stakes may entail higher standards of evidence" (p. 21). The author served as a program director for educational administration and an instructor in courses that addressed the PAL. At this urban university, the administrative candidates represented the racial, ethnic and language diversity of the K12 students. Furthermore, the author was a certified scorer for the PAL during the years of the Massachusetts pilot (2014-2015) and field trial (2015-2016). Validity studies must also pay attention to the interpretation made of test results and the context in which the test is used. For this reason, this paper will also consider the particular use of the PAL in Massachusetts as a high-stakes, "make-it-or-break-it" test for licensure.

Literature Review

Leadership Assessments

In recent years, states have been paying closer attention to the recruitment, preparation, certification, evaluation, and on-going professional development of school principals (Augustine, Gonzalez, Ikemoto, Russell, & Zellman, 2009; Manna, 2015; Shelton, 2011, 2012). On the issue of certification, there are increasing calls for three-tiered licensure systems and state-approved

[1] Although certification and licensure have different meanings in some states, such as Louisiana and New York, both terms refer to state authorization in Massachusetts and in this paper.

alternative licensure programs (Shelton, 2011). Tiered certification systems vary by state but might include graduate coursework, an internship, on-the-job experience, mentoring, and evidence of improved student achievement (Shelton, 2011). The Southern Regional Education Board suggested in 2007 that a new certification system should require "performance-based assessments and tests of knowledge for entry-level licensure" as well as "tools for evaluating the on-the-job performance of practicing principals aligned with state leadership standards" (Fry, Bottoms, O'Neill, & Walker, 2007, p. 21). These multi-pronged approaches acknowledge that school leader effects account for no more than 5-10% of student achievement (Leithwood, Louis, Anderson, & Wahlstrom, 2004) and the pathway to student learning is indirect and complex (Leithwood, 2010).

Many states now include a summative assessment of novice principal competency as a basis for certification decisions (Condon & Clifford, 2012). Current tests of knowledge include the Praxis and the School Leaders Licensure Assessment (SLLA) as well as the Connecticut Administrators Test (CAT). These are sit-down, one-day summative evaluations, which are not publicly available for purchase. However, after 10 years of administration, new research challenges the validity of the SLLA as a screening and/or signaling device for principal effectiveness (Grissom, Mitani, & Blissett, 2017). The PAL addresses the call for a performance-based assessment for entry-level licensure. Similar to the SLLA and CAT, the PAL is a state-administered, summative licensure exam and yet is publicly available because this is not a sit-down examination and can take over 200 hours to complete (MA-DESE, 2016a).

Promise of a PBA

Performance-based assessments (PBAs) promise to provide directness and authenticity beyond the reach of sit-down tests, which will benefit both candidates and preparation programs (MA-DESE, 2017), while avoiding the negative consequences often associated with high-stakes tests (Linn, 1994). PBAs ask students to perform, create, or produce something, which is often challenging and meaningful, requiring problem-solving and higher-order thinking skills in a real-world context where the criteria and standards for performance are known in advance (Aschbacher, 1991). There is not just one right answer. The PBA is a kind of direct measure, which captures "performances that are valued—that people care about" not only in school but also in society (Resnick & Resnick, 2013, p. 25). In this case, there is a belief that, "Human judges are required to certify the quality of what is learned, along with a kind of 'show me' attitude on the part of those judges" (Resnick & Resnick, 2013, p. 26). The actual performance can be submitted as a live format or a secondary product. For example, most states require new drivers to both pass a written exam and a road test. In this case, the written exam has subject-matter validity when it aligns with the state rules of the road while the road test has predictive validity when it measures whether a driver can actually follow those rules (Linn, Baker, & Dunbar, 1991). A written test is not a sufficient assessment of one's readiness to drive a car, to play an instrument, or to compete in a soccer match.

The promise of direct measures, such as the PBA, is that they motivate student choice, collaboration, self-reflection and, for both student and instructor, learning improvements (Darling-Hammond & Adamson, 2014). Nonetheless, PBAs are not without limitations. Aschbacher (1991) found that the concept of PBAs "has seductive face validity" (p. 277) but her survey of K12 PBA employment in 50 states found widespread concerns about cost, logistics, technical requirements and generalizability. The PAL cost $1,955,065 for development (Jennifer Briggs, personal communication, December 4, 2015), including a one-year pilot and one-year field test, followed by an additional year of implementation adjustments. Candidates pay $428 for the test, which covers administrative costs. These explain why PBAs are not more common; less obvious are the validity challenges, which have been raised in the past and are the focus of this paper.

Design of the PAL

The PAL is presented as an authentic test "of the most essential work of school leaders," which is aligned with the Interstate School Leadership Licensure Consortium (ISLLC) and Educational Leadership Constituent Council (ELCC) standards (Massachusetts Performance Assessment for Leaders, n.d.).[2] The test employs four tasks (to be completed in a school setting in conjunction with faculty and the administrative team) that are summarized as "setting direction, creating a professional learning culture among staff, supporting individual teacher development, and engaging families and community to improving student learning" (Massachusetts Performance Assessment for Leaders,n.d., see also Appendix A for complete definition). Each task includes several dimensions, which describe succinct leadership skills, and each dimension includes one or more indicators (Orr, Pecheone, Hollingsworth, et al., 2017). The test is scored by independent, anonymous, certified scorers (each task by a different pair of scorers) who use a publicly available rubric to evaluate the material on a four-point interval scale (Beginning, Developing, Meeting, and Exceeding).[3] The design team estimated that each task could be scored in 30 to 45 minutes. The scoring software calculates an average for the indicator, dimension, and task scores. For students, each task requires 40 to 80 hours to complete (MA-DESE, 2016a). Candidates describe their performance and upload the 43-page, single-spaced narrative, along with substantiating evidence, to a private web-based platform (Pearson). Most candidates take the test during their administrative practicum, which is commonly conducted as part of a state-authorized principal preparation program.[4]

Purpose of Assessment

Resnick and Resnick suggested four purposes for assessments: certification, monitoring, accountability, and learning improvement (2013). The PAL has a primary certification purpose—"The PAL assessment system is designed to produce clear evidence of a candidate's readiness for an initial school leadership position" (MA-DESE, 2016b, p. 1; Orr et al., 2015, p. 1)—but the other purposes are also affirmed:

b. Monitoring: "Inform preparation pathways' ability to prepare candidates to become leaders who can meet the challenge of educating all students."
c. Accountability: "Advance quality leadership development by integrating licensure assessment with state expectations for leadership preparation, districts' new leader induction, and Principal evaluation…."
d. Learning Improvement: "Educate leadership candidates to be able to lead teachers, children, and schools and prepare students for 21st century skills for college, career, and life." (Massachusetts Performance Assessment for Leaders,n.d.)

---

[2] The original PAL preceded the updated ISLLC standards (Professional Standards for Educational Leaders) and the updated ELCC standards (National Educational Leadership Preparation). Massachusetts, which adopted the test, and California, which piloted portions of the test, also sought to align the test with their own state leadership standards.

[3] During the Massachusetts pilot and field tests, the scorers were primarily active or retired K12 educators plus faculty from state-approved principal preparation programs.

[4] Massachusetts offers three routes to principal certification, including a 500-hour practicum or successful defense before a panel review, in addition to the traditional principal preparation program (MA-DESE, 2016c). Successful completion of the PAL is a requirement for all three routes.

In the future, states that lack "high-leverage policies" that promote the development of effective school principals, such as Florida, Hawaii, Indiana, Michigan, Nebraska, Ohio, Oklahoma, South Dakota, Texas, and Wyoming (Anderson & Reynolds, 2015, p. 42), may find the PAL particularly attractive for its ability to address the four purposes in one package. Resnick and Resnick (2013) argued, "learning improvement has the highest and most urgent priority" (p. 29), confirming the primary purpose of most educational institutions where licensure is an important but secondary outcome.

Field Trial Evaluation

Construct validity. The creators of the PAL assessed the validity of the PAL based upon results from the pilot and field trials. Using work submitted by 416 leadership candidates who completed all four tasks, the investigation team examined construct validity for each indicator, dimension, and task. The team addressed construct validity from four different angles, looking at the measurement quality of each indicator and examining the correlations among indicators, dimensions, and tasks for internal coherence and differentiation. For example, the researchers found strong internal consistency among the leadership dimensions for each task as shown in Table 1, which were balanced by distinctiveness (as none were higher than 0.77):

Table 1
Range of Correlation between Leadership Dimensions for Tasks 1–4

Task	Range of Correlation between Leadership Dimensions
1	.591 to .681
2	.551 to .729
3	.527 to .714
4	.726 to .772

Source: (Orr, Pecheone, Hollingworth, et al., 2017).

Factor analysis showed that the dimensions were "strongly related aspects of school leadership" and made "unique contributions as separate measures within the task, while contributing little to the other tasks" (Orr, Pecheone, Hollingworth, et al., 2017, p. 15). The correlations among the four tasks were smaller, yet moderate and positive (0.208 to 0.269), suggesting that a four-factor model was appropriate for the PAL. The authors concluded that the tasks, dimensions, and indicators "work well as a combined measure…." (Orr, Pecheone et al., 2016, p. 83).

Reliability. During the field trial, approximately 25% of the 416 completed tests were double-scored so that scoring reliability could be measured. (Today, double scoring is standard practice on the Massachusetts PAL). The authors used two forms of assessment to measure scoring reliability. For indicator reliability, the authors determined that exact rates (where scorers with the same portfolio assigned the same score) were above 50% on most rubrics; however, exact agreement was under 50% on five out the six Task 4 indicators, which suggested that "further scorer training for Task 4" might be needed (Orr, Pecheone, Hollingworth, et al., 2017, p. 11). For task scoring reliability, the authors conducted a generalizability analysis (*G* study) and found reliability coefficients above 0.70 for Tasks 1, 2, and 4 with two scorers (Table 2).

Table 2
Estimated Reliability (G) Coefficients for Two Scorers, by Task and Number of Test Submissions

Task	Number of Test Submissions	Reliability (*G*) Coefficient
1	100	0.842
2	99	0.792
3	80	0.345
4	92	0.735

Source: (Orr, Pecheone, Hollingworth, et al., 2017, p. 12)

The coefficient for Task 3 was only 0.345, which they attributed to low variance among candidate Task 3 scores. However, the researchers also admitted that a preponderance of candidates scored at level 3 (Meeting) on all four tasks (Table 3), which they suggested could inflate rather than depress reliability rates (Orr, Pecheone et al., 2016). These seemingly contradictory conclusions—inflate or deflate—for score clustering are analyzed in the results section below.

Table 3
Percentage of Candidates Scoring at Level 3 (Meeting)

Task	% at Level 3
1	69.0
2	69.2
3	76.3
4	59.7

Source: (Orr, Pecheone et al., 2016)

Using a stratified coefficient alpha, the team calculated the potential reliability of a total score (the average of all four task scores) and found an overall reliability coefficient of 0.844 for two scorers (Orr, Pecheone, Hollingworth, et al., 2017).

Content validity, bias review, feasibility, and educative value. The design team measured content validity by three methods: alignment to the MA state leadership standards, a formal validation study completed by a committee of K12 leaders and higher education faculty, and then two rounds of face validation gathered from surveys of leadership candidates and their program faculty. They found content validity to be "consistently strong" (Orr, Pecheone, Snyder, et al., 2017, p. 21). A separate bias review committee found the PAL free from threats to bias and sensitivity for candidates. ANOVA analysis of test scores showed that females averaged 2.89 versus 2.77 for males (on a four-point scale). Candidates in preparation programs outscored their counterparts in alternative pathways by only 0.06 points. Racial disaggregation was unavailable due to insufficient numbers. Surveys of candidates and faculty concluded that the PAL offered ease of use and feasibility for principal preparation programs and flexibility for different types of settings (Orr, Pecheone, Snyder, et al., 2017).

Program quality and effectiveness. More recently, PAL designers surveyed program faculty (n = 12, representing about half the MA preparation programs) and students (n = 53) to assess the influence of the PAL on program quality and effectiveness (Orr & Hollingworth, 2018). Since they gathered responses during difficult adjustment years (the field trial and following implementation year), the authors assumed that results would be likely to over-represent those who were aggravated by the test (Orr & Hollingworth, 2018). Based upon largely favorable responses

from participants, the authors concluded "PAL has the potential to serve as both a high-quality summative assessment tool for licensure as well as a learning opportunity and barometer for programs to judge the quality of their preparation, both in Massachusetts and in other states" (Orr & Hollingworth, 2018, p. 18).

Critical Educational Policy Analysis

PBAs reflect the evolution of assessment theory away from positivistic tests to better offer directness and authenticity while addressing the four purposes listed above. In parallel fashion, the field of policy analysis has evolved in recent decades; scholars have moved away from traditional, positivist, policy analysis research that assumes that educational reform is a deliberate process, that behavior is goal-driven, neutral, and rational, that implementation knowledge is apparent and readily available, and that problems can be easily evaluated and remedied (Diem & Young, 2015). Instead, new approaches are informed by expressions such as feminism or critical race theory, employ new methodologies, such as discourse analysis and policy network analysis, and are more likely to pay attention to policy content, context, development, and unintended consequences. Taylor (1997) described this as a "shift towards exploring the effects of policy rather than on policy intentions" (p. 24). Critical policy analysists examine the roots of policy development, how policy reinforces existing hegemonic social structures, and the differences between policy rhetoric and reality. They "take great care in delineating the perspectives they bring to their work and how those perspectives inform how they do research" (Diem & Young, 2015, p. 844). They are often engaged activists, striving to bridge research and practice. In her 1997 paper, Taylor argued, "methodological issues have been side stepped for too long in education policy analysis. More attention should be given to questions of meaning and interpretation—as well as validity, reliability and subjectivity" (p. 33). In order to expand the analysis and enlarge the conversation, this paper adopts a critical stance, which allows greater attention to content, context, development, meaning, interpretation, culture, and unintended consequences.

Methodology

This study began with the question "What is the validity evidence for the use and interpretation of the PAL in regards to emerging school leadership?" and relies upon multiple sources of qualitative data to offer fine-grained, critical analysis. According to Diem and Young (2015), critical policy analysis (CPA) scholars approach their work with three overriding concerns: context and complexity, concentrated looking, and theory. Each concern is defined and addressed in a separate paragraph in the methodology sections below. First, CPA scholars understand that policy is constructed in complex systems and implemented in widely varying environments. For example, there are currently 23 authorized principal licensure preparation programs in Massachusetts, both degree-granting institutions of higher education and non-degree-granting institutions (MA-DESE, 2018). Students enter with aspirations to lead at the elementary, middle, or high school level in some administrative capacity such as school principal, assistant principal, curriculum director or dean of discipline. Some candidates hope to move into leadership immediately in schools that vary by size, location (urban, suburban or rural), student diversity, public versus private funding, and traditional versus charter status; other candidates choose to delay advancement. The PAL licensure requirement applies equally to all these goals and settings. Some preparation programs also enroll out-of-state or international students who are not seeking licensure. Knowing this, it is important to pay attention to the social, political, and historical context of the PAL policy development and implementation.

Data Sources

As a second concern, CPA scholars engage in "concentrated looking" (Diem & Young, 2015, p. 844) or prolonged consideration of multiple sources of data such as policy texts, observations, interviews, and contextualizing information in order to develop a deep understanding of a phenomenon. The launch of the PAL included a *Candidate Assessment Handbook* (MA-DESE, 2016b), which introduced the test and offered completion strategies as well as scoring rubrics. Other documentation addressed the design process, first-year outcomes (Losee & Orr, 2015), alignment with state standards (MA-DESE, 2016b; Orr, Pecheone et al., 2016), a scorer's manual (MA-DESE, 2014), administrative field guide (MA-DESE, 2016a), leader toolkit (MA-DESE, n.d.), and so on, which were available to candidates and program directors on a secure website.[5] A Freedom of Information Act request yielded further information about test development costs.

For this paper, research on the PAL began with an immersion period of over two years as the author read and reread the guiding state documents (both alone and in the classroom with students), honed scoring skills in the review of student submissions (from various preparation programs), listened to student feedback, and reflected, as a former school leader and preparer of school leaders, on the test. Dating from initial engagement with the pilot project in early 2014, the author (in his role as program administrator of a participating urban university) saved available data in a portfolio to inform planning decisions for the graduate principal preparation program. These data included publicly available information such as website content, memorandums and reports from DESE, and notes from numerous meetings (statewide gatherings, local, and online) with test designers and state officials during the years 2014–2017. In addition, as a course instructor, the author saved student test submissions (hoping to provide exemplars for future cohorts), meeting notes, and email correspondence with over 60 students and several faculty members. After two years, the author decided to conduct formal research on the PAL and gathered letters of informed consent from all students and faculty that included confidential and anonymous use of discussions, email correspondence, student work, survey, and interview data. The author kept a running journal of personal experiences, conversations, reflections, questions, and concerns. As a certified test scorer, the author rated 38 test submissions from students at other institutions; the tests, while not saved, did provide insights for the research journal. As a former public high school principal, the author had practical, first-hand knowledge of school leadership, which also informed data analysis. The focus of this paper is not on the implementation of the test as it occurred in Massachusetts but on the test itself and the validity of the test interpretations.

Data Analysis

For CPA scholars, the choice of theory is the third main concern; theory influences the identification of the research problem, the manner in which the researcher thinks about the problem, the kinds of questions raised, and the lens through which data are analyzed. As questions about the validity of the test began to emerge, the author turned first to the standards and theoretical framework employed by the authors of the technical reports on the PAL pilot and field trial (Orr, Pecheone, Hollingworth, et al., 2017; Orr et al., 2016; Orr, Pecheone, Snyder, et al., 2017): namely, the *Standards for Educational and Psychological Testing* (2014). This resource was developed jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education, and updated periodically since 1974. (This paper also uses language from the 1999 version where appropriate.) The *Standards* (2014) consider validity evidence based on test content, response processes, internal

[5] The websites are updated each year; data for this paper were pulled for school year 2016-2017.

structure, the relations to other variables, and consequences of testing; these five sources became the organizing framework for data analysis. Test developers often concentrate on test content and internal structure (as did the authors of the technical reports) but this paper addresses all five sources in keeping with a holistic CPA approach. "Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use" (AERA et al., 2014, p. 14). Validity has application to the uses and interpretation of a test instrument, not to the test itself. One cannot simply claim, "The test is valid." Furthermore, "When test scores are interpreted in more than one way…each intended interpretation must be validated" (AERA et al., 2014, p. 11).

Using this framework, the author began to sift through the evidence, considering first test content, then response processes, followed by internal structure, relations to other variables, and finally consequences of testing. Each stage began with document analysis of the DESE literature, such as the handbook, field guides, scorer's manual, and standards alignment piece, supplemented by notes from state meetings and memorandums in order to best understand the authors' understanding and intentions. A consideration of the technical reports by the design team (Orr, Pecheone, Hollingworth, et al., 2017; Orr et al., 2016; Orr, Pecheone, Snyder, et al., 2017) often followed. The researcher then compared this information with material gathered from faculty and students plus PAL test scoring, examining everything in light of the literature and personal experience in school leadership. In this way, the research focus moved from policy intentions to policy effects (Taylor, 1997). However, each validity source called for a different mixture of evidence. For example, the consideration of test content validity began with the standards alignment document (Orr et al., 2016), followed by the state leadership standards and the research literature on school leadership; in contrast, the final stage (consequences of testing) began with the design team's analysis of programmatic effects (Orr & Hollingworth, 2018) followed by examination of internal evidence from faculty and students. Each stage of analysis involved extensive reading, reflection, and writing, which often prompted new insights on previous stages already written; a recursive process of consideration, reflection, writing, reexamination, and rewriting ensued.

As an engaged scholar, the author adopted a CPA approach to complement the technical papers offered by the PAL design team. Diem et al. (2015) recommended three particular angles for fine-grained policy analysis, which informed this investigation. First was constant attention to power and voice in PAL policy-making in order to discern who had a voice in the policy process. Next, it was important to unpack assumptions and look for the fundamental ideas that were underpinning the PAL policies (Diem, Young, Welton, Mansfield, & Lee, 2014). Finally, it was important to question: "Is this the way it has to be; what's the value of doing it this way; how are people hurt by this; what are the alternatives?" (Diem et al., 2014, p. 1076).

In this case, the author's experience as an urban principal followed by intense end-of-career work with diverse graduate students who worked in urban settings provoked particular attention to a race-conscious consideration of the data. The evidence called for a theory of multicultural validity (Kirkhart, 2005, 2010) to supplement the *Standards* (2014). Multicultural validity refers to "the accuracy or trustworthiness of understandings and judgments, actions, and consequences, across multiple, intersecting dimensions of cultural diversity" (Kirkhart, 2010, p. 401). Culture is not a neutral variable; instead, "power is attached in varying ways and degrees to different dimensions of culture in different contexts" (p. 402). Kirkhart (2005) noted, "Validity is threatened to the extent that culture is ignored or diversity variables are included as simplistic, atheoretical stereotypes" (p. 24). Policy making is sometimes viewed as "an arena of struggle" between "contenders of competing objectives, where language—or more specifically, discourse—is used tactically" (Taylor, 1997, p. 26, quoting Fulcher 1989). A discourse analytic approach is useful in bringing to light competing

discourses and varying values, which can often identify "slippage between objectives and outcomes" (p. 32). Authorial intentions may not be clear, but policy implementation inevitably involves "reinterpretation and recreation" (p. 27) as the policies take on new meaning for the readers.

Using the lens of multicultural validity, the author repeated the document analysis of the PAL materials and technical reports while looking in particular for vocabulary and/or language specific to race and culture. Starting with the five sources of validity evidence and the attending evidence employed for each stage of analysis, the author compared this with first-hand knowledge and experience with school leadership (from the author and graduate students) as well as the research literature on culturally competent leadership. This final analysis proved fruitful in unpacking assumptions, shedding light on power and voice, and considering alternative policy approaches.

Results

Several papers from the PAL design team offered a preliminary examination of validity evidence and were presented in the literature review (Orr & Hollingworth, 2018; Orr, Pecheone, Hollingworth, et al., 2017; Orr, Pecheone, Snyder, et al., 2017). The results presented here offer a critique of that literature and additional evidence that challenges the validity of the PAL. The results are organized according to the five categories of the *Standards* (2014)—i.e., test content, response processes, internal structure, relation to other variables, and consequences—followed by a consideration of the PAL through the lens of multicultural validity.

Evidence Based on Test Content

Validity evidence based on test content refers to the "themes, wording, and format of the items, tasks, or questions on a test" (AERA et al., 2014, p. 14) as well as administration and scoring. This was a primary concern of the PAL developers who evaluated content validity from three different perspectives, as noted in the literature review. Test developers often begin by defining the domain and constructs that are to be measured by the assessment. In this paper, *domain* refers to the broad concept of readiness for initial school leadership.[6] According to the *Standards* (2014), "a list of the tasks constituting a job domain may be developed from observations of behavior in a job, together with judgments of subject matter experts" (p. 14). Each of the four tasks in the PAL represents a more narrowly defined leadership construct (see Appendix A).

Alignment with the MA leadership standards. The *Standards* (2014) introduced construct under-representation and construct-irrelevant variance. In laymen's terms, does the test measure less than what is intended or more than what is intended (Goodwin & Leech, 2003)? Construct under-representation is addressed here and construct-irrelevant variance is addressed in the section titled Evidence Based on Response Processes. Construct under-representation is judged by "evaluating whether test content appropriately samples the domain set forward in curriculum standards…." (AERA et al., 2014, p. 15). The *Candidate Assessment Handbook* (MA-DESE, 2016b) asserted that the PAL was aligned with the national performance assessment requirements of the Educational Leadership Constituent Council (ELCC), the national educational leadership policy standards of the Interstate School Leadership Licensure Consortium (ISLLC) 2008, and the revised Professional Standards for Administrative Leadership, which were approved by the Massachusetts Board of

[6] The author's use of domain follows general psychometric literature. In contrast, Orr, et al., (2017) used the term domain narrowly to refer to a particular aspect of leadership performance; thus the four tasks of the PAL address 13 different leadership domains (such as data analysis, vision plan and focus, group learning and work, teacher development, etc.); these are termed *dimensions* in this paper.

Education in December 2011. However, only the alignment with the state leadership standards was described in detail (Orr et al., 2016).

The new Massachusetts leadership standards include four overarching standards and 40 specific indicators (Chester, 2012). In addressing alignment, the PAL designers admitted that the test was not comprehensive: "The four PAL tasks reflect three of the four Massachusetts Leadership standards strongly and some indicators of the fourth standard weakly" (Orr, Pecheone, Snyder, et al., 2017, p. 11). In reality, the tasks lack language specifically calling for demonstrated proficiency on each indicator. Alignment studies should include categorical concurrence, depth of knowledge consistency, range of knowledge comparability, and balance of representation (Webb, Horton, & O'Neal, 2002, April). Furthermore, "depth of knowledge consistency is the cornerstone…and is evident if what is elicited from students on the assessment is as demanding cognitively as what students are expected to know and do as stated in the standards" (Webb et al., 2002, April, p. 3). The PAL development team, assisted by DESE staff, school and district leaders, and higher education faculty (the design committee) estimated comparability with the 40 state indicators on a three-point rubric, where each task either "directly assesses performance," "requires performance" or simply requires knowledge ("bumps into") the indicator, (Orr et al., 2016, p. 90). As one example, the technical report stated that Task 2 "directly assesses performance" on a management indicator called *scheduling*, giving this the highest rating possible for this indicator (Orr et al., 2016, p. 89). The state standards use this language for scheduling: "Ensures a comprehensive scheduling system that provides sufficient time for instruction, teacher planning and collaboration" (Chester, 2012, p. 15). One might expect candidates to wrestle with the advantages and disadvantages of block scheduling, looping, rotating schedules, lengthened school days or years, for example. In contrast, Task 2 focused on developing a professional learning culture for teachers and asked candidates to "Schedule…a series of meetings to foster the professional learning of the group over time" (MA-DESE, 2016b, p. 40) and then to report on "the schedule of meetings" and offer "an explanation of how…you secured and scheduled meeting time" (MA-DESE, 2016b, p. 47). There is no attention paid in Task 2 to a comprehensive scheduling system that addresses instruction. Alignment work also calls for range of knowledge comparability and balance of representation (Webb et al., 2002, April). According to the alignment document, the PAL required only some performance on nine indicators (*Technology; English Language Learners; Safe, Orderly, and Caring Environments; Management and Information Systems; Fiscal Systems; Improvement Planning; Advocacy; Cultural Awareness; Managing Conflict*) and only bumped into *Operational Systems* and *Contract Negotiations* (Orr et al., 2016). To summarize, the PAL is not well aligned with the Massachusetts state leadership standards but falls short on depth of knowledge consistency, range of knowledge comparability, and balance of representation (Webb et al., 2002, April). According to the *Standards* (2014), the validity of the PAL is threatened when the test content does not "appropriately sample" the domain of school leadership readiness as set forward in the curriculum standards (p. 15).

Domain coverage. The domain that addresses readiness for school leadership is enormously complex, presenting challenges to standardized assessments such as the PAL. There are two approaches for achieving domain coverage in PBA design: the domain sampling approach and the critical indicator approach (Mislevy & Knowles, 2002). With the former, a large number of tasks are developed representing the knowledge and skills of the domain but a random subset is selected for a particular test administration. The weakness of this approach is the difficulty of creating a large number of tasks and the dubious assumption that all tasks are of equal weight; the strength is that all aspects of the domain are sampled. In contrast, the critical indicator approach (the PAL approach) begins with the assumption that a few tasks are more important than others; if a candidate can successfully address these, then other job requirements should be readily accomplished. This

approach is far more efficient, but the drawback is that it neglects aspects of the domain and begins with the assumption that some leadership tasks are more important than others. "If the assessment tasks are not on target, the results will not provide useful information" (Mislevy & Knowles, 2002, p. 72). Do the four PAL tasks measure the most essential work of school leaders?

Candidate materials clearly stated the construct definitions. For example, Task 1 of the PAL was titled "Leadership through a Vision for High Student Achievement" and introduced as "setting the direction for improved student achievement" (MA-DESE, 2016b, p. 2). Setting direction is widely recognized as a primary leadership function (Bass, 1999; Goleman, 2000; Leithwood et al., 2004; Mendels, 2012; Murphy, Elliott, Goldring, & Porter, 2007; Portin et al., 2009). The instructions for Task 1 asked candidates to develop a vision for one target group of students and for one prioritized academic area. The *Candidate Assessment Handbook* (MA-DESE, 2016b) operationalized Task 1 in this way:

a. Access, collect, and analyze three to five years of quantitative student performance data, qualitative data on school culture and student learning, and overall school context information.
b. Identify a priority academic area where improved student performance is needed, with attention to federally designated priority student groups…
c. Collect additional quantitative and qualitative information about the student group's performance in the priority academic area…including findings from observations and staff and student interviews, focus groups, and/or surveys….
d. Select a target student group.
e. Document existing school programs, services, and practices…and identify the gaps in effectiveness and opportunities for improvement.
f. Solicit input from school leaders, teachers, and other relevant stakeholders… about the student learning needs, priorities, gaps, and opportunities for improvement….
g. Develop a vision, set of action strategies, and a proposed plan to improve the target student group's learning in the priority academic area….
h. Solicit feedback about the need… and the relevance and feasibility of the proposed plan from school leaders and key stakeholder groups….
i. Evaluate the feedback and make appropriate revisions to the plan….
j. Summarize and constructively critique the leadership skills and practices that you used or developed in completing this task. (p. 13)

The *Candidate Assessment Handbook* (MA-DESE, 2016b) recommended that candidates implement the action plan (step g) from Task 1 and use this as the foundation for Tasks 2, 3, and 4; however, this was not always the case. In fact, there was no requirement that Tasks 2, 3 or 4 relate to Task 1 at all, which meant that candidates could conceivably complete four unrelated tasks in order to demonstrate readiness for school leadership. This was not uncommon for urban candidates where many schools were subject to the federal turnaround process and students had to change their practicum setting halfway through the program. The two scorers for Task 1 would never see the products for Tasks 2, 3 or 4 (and vice versa). As a result, the vision carved out for Task 1 might remain on paper only, scored by individuals who would never see an actual performance. This atomization of the domain of readiness for school leadership into four unrelated tasks undermines the purpose and validity of the PAL.

Scoring rubrics. Scoring rubrics can also present a separate validity challenge to any PBA if they "focus on lower levels of thinking rather than on the more complex reasoning and thinking skills that the tasks are intended to measure…" (Darling-Hammond & Adamson, 2014, p. 153). Grain size is an important qualification. Baxter and Glaser (1998) characterized the structure of PBAs along two dimensions. One continuum represented the task demand for content knowledge ranging from lean to rich; the other continuum represented the task demand for cognitive processes ranging from structured/constrained to open. Open processes invite candidates to develop their own strategies and procedures. The greatest cognitive complexity is found in the quadrant that is "process open and content rich" (Darling-Hammond & Adamson, 2014, p. 141). There are three scoring rubrics for PAL Task 1 titled:

> Rubric 1a: "Investigate and Prepare a Vision," which included the dimensions of data collection, data analysis and evaluation of existing policies, practices and programs
>
> Rubric 1b: "Design an Integrated Plan for Strategies to Develop and Implement Improvement in the Priority Academic Area," which included the dimensions of vision and plan focus, solicitation of input, and plan details.
>
> Rubric 1c: "Assess and Analyze Feedback from Participants," which included the dimensions of plan feedback and assessment of leadership skills and practices. (MA-DESE, 2016b, pp. 29–35)

For example, the first step on the to-do list above instructed candidates to "Access, collect, and analyze three to five years of quantitative student performance data, qualitative data on school culture and student learning, and overall school context information" (MA-DESE, 2016b, p. 13). According to Rubric 1a, here is a list of what a candidate would have to do in order to attain a score of proficiency:

a. Collects data on three or more elements.
b. Makes a clear connection between the selection of the priority academic area and the data collected.
c. Collects data for at least two quantitative or at least two qualitative elements.
d. Collects three or more years of data for at least one data element.
e. Collects data for two or more student subgroups and designates a target student group.
f. Collects some relevant data from teachers and/or students about performance and/or student culture that help to clarify some reasons for the target student group's learning problems. (MA-DESE, 2016b, pp. 29–30)

The high specificity predisposes a constrained process. Since the name of the candidate and the identity of the school is hidden, the scorers have no way of knowing what data are accessible (or even real) or which data are most important. The actual school data are submitted online in a window titled "Categories." Scorers are not required to read the data; they must simply confirm the submission of data. From their limited perspective, the scorers must scan the narrative. Did the candidate offer data on three or more elements, describe two kinds of quantitative or two kinds of qualitative data, which address at least two student sub-groups, and one of which extends for at least three years? Did the candidate make a logical connection between the data selected and the priority academic focus? And so on. Understandably, students focus on the same checklist. The process,

which ought to be open becomes more and more constrained. In this way, the initial school leadership task titled "Leadership through a Vision for High Student Achievement" (MA-DESE, 2016b, p. 2), which appears to be cognitively complex, is threatened by the constraints of the scoring rubric and the detachment of the scoring process and becomes a highly constrained and content lean exercise. The fine grain size limits the validity of the use and interpretation of the PAL score on Task 1.

Data analysis for Task 1. In terms of cognitive complexity, data collection pales in comparison to data analysis. Rubric 1a requires that each candidate "Presents a comprehensive analysis of data collected with a clear connection to identify the priority academic area and target student group" (MA-DESE, 2016b, p. 30). However, the candidate narrative, which is capped at 1500 words, offers inadequate space to discuss the analytical steps; most students simply state their findings. Furthermore, scorers are allotted only 30 to 45 minutes to evaluate all of Task 1 (6000 words), which prohibits independent data analysis. In general, the scorers accept the candidate's data selection, which may or may not be accurate, and accept the candidate's analysis at face value. This simple example demonstrates how the PAL is more a product than an actual, live performance.

Task reduction. In some cases, a candidate might submit a product that addresses most of the bullet points on the rubric, but not all. For example, a candidate in a new school might not have access to three years of data. Scorers were instructed to consider the preponderance of evidence as they made a determination (MA-DESE, 2014). In other words, even the bullet points on the scoring rubric could be pared down. This last point is disturbing because the research literature suggests that categorical scoring is prone to its own reductionist tendencies such as the halo effect, leniency, central tendency, and reduction of range (Humphry & Heldsinger, 2014). In short, the rubrics used for PAL scoring were reductive and the instructions allowed this kind of simplification. There are sound psychometric reasons for this kind of reduction in the design of a performance task. Dunbar et al. (1991) explained,

> The contrast between score and rate reliability introduces an inevitable reliability-validity tradeoff into performance assessment. The tradeoff is well known among test developers. Further narrowing and structuring of tasks might be expected to increase score reliability in the same way that writing homogenous objective test items does…. However, doing so narrows the domain to which results generalize…. That narrowing poses an unattractive choice in terms of validity: if inferences are kept broad enough to be important, their validity is undermined; if inferences are narrowed to maintain validity in the face of the restricted definition of the task, they become unimportant. (p. 294)

This will always be a dilemma when one attempts to score a complex domain with a limited number of tasks and a standardized scoring rubric. One can increase the score reliability by increasing the number of tasks (Dunbar et al., 1991; Linn, 1994), but this also increases the time and cost requirements for the test. The other three tasks of the PAL presented similar problems.

Authenticity. In conclusion, Messick (1994) stated that authenticity is the watchword for construct under-representation. During the PAL design process, a content validity committee assessed authenticity (or job relevance). The ten-member committee, composed of Massachusetts K12 school leaders and representatives from state-licensed principal preparation programs, met initially for one day of training and evaluation of the four tasks, under the auspices of state department leaders allied with PAL designers. This arrangement, which lacked independence and

unrestricted time for reflection and discussion, challenged honest reflection, discussion, and deliberation. Preparation program representatives must have experienced a conflict between their desire to be included versus their apprehension that the final test could be used—by the same attending DESE leaders—to evaluate not only candidates but entire programs. The notes revealed that two committee members had to leave early. Nevertheless, under these compromising circumstances, in response to the 5-point Likert scale question, "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" (Orr, Pecheone, Snyder, et al., 2017, p. 13), 20% of the committee members bravely challenged Tasks 1 and 2, arguing that this kind of work would be "performed less frequently by an entry-level school leader" (Orr, Pecheone, Snyder, et al., 2017, p. 12).

Similarly, in the Massachusetts field trial face validity student surveys, "most candidates agreed that Task 3 was complementary to their leadership preparation, two thirds agreed that Tasks 2 and 4 were complementary," but only half (56%) agreed that Task 1 was complementary to their preparation (Orr, Pecheone, Snyder, et al., 2017, p. 13). During a pilot study in California (which employed only a portion of the PAL), the PAL test developers queried 15 candidates on the content validity of the PAL. Four to six participants were unable to access and analyze data on student engagement indicators, school culture indicators, and teacher proficiency and engagement indicators in order to address Task 1. Some candidates reported that their school did not have professional learning groups, which made it difficult to implement the work of Task 2. In this case, the test was not authentic because contextual issues affected the implementation of the tasks and interpretation of the scores.

In Massachusetts, PAL candidates compose 11 narratives (called artifacts) and four self-reflective commentaries totaling 21,500 words (or 43 single-spaced pages). In addition, they gather and assemble scores of school-based documents into 17 separate online submissions (called categories) along with two 30-minute videoclips, which have been edited down to 15 minutes each. This is substantial desk work that is tangential to real school leadership. Real school administrators rarely compose wordy narratives describing their work. In one urban district partnership, for example, educational administration candidates are pushed to work in teams and share their work through Google documents, spreadsheets, and PowerPoints because this is how real work is managed in the district. Solo school leadership is strongly discouraged in favor of collaborative approaches that promote teacher leadership. From this perspective, an assessment that focuses on individual performance and is measured through long written narratives lacks authenticity. The heavy emphasis on writing and preparation of individual submissions, which is not germane to real school leadership, is a good introduction to the next section, which addresses construct-irrelevance. To summarize this section on validity based on test content, it appears that on multiple counts—the selective alignment to the state leadership standards, incoherent atomization of the domain, constraints of the scoring process, the limitations of rating rubrics, plus the contextual threats to authenticity, and heavy emphasis on writing skills—the PAL suffers from construct underrepresentation and lacks validity as a test of readiness for initial school leadership.

Evidence Based on Response Processes

Just as authenticity is the watchword for construct under-representation (Messick, 1994), so directness is the watchword for construct-irrelevant variance. The *Standards* (2014) state the matter this way: "Construct-irrelevance refers to the degree to which test scores are affected by processes that are extraneous to the test's intended purpose…by processes that are not part of the construct" (p. 12). We have already seen how composing 43 single-spaced pages of narrative is different from an actual performance. Online tests such as the PAL can also present other response-process

complications. Serious candidates should read the 106-page *Candidate Assessment Handbook* (MA-DESE, 2016b), sign the Confidentiality and Anonymity Form, save the four Evidence Charts for future reference, download the four model consent forms for use with relevant district, teacher, parent, and child representatives, and view the four task videos (each 9 to 15 minutes long) (MPAL Candidate Resources, n.d.). In addition, the Pearson ePortfolio system offers the Candidate Guide to Using the Pearson ePortfolio System; Frequently Asked Questions about the Pearson ePortfolio System; Tips for Mac Users of the Pearson ePortfolio System; Video compression guides for Macs and PCs; Video exporting guide for iMovie, iPhoto, and Windows Movie Maker; and Recommended Video Formats and Settings (MPAL Candidate Resources, n.d.). A separate webpage describes PAL Policies for Candidate Participation including statements on registration (Assessment Fees and Payment Information, Payment Policy, Changing Registration, Withdrawal/Refund Policy), assessment policies (Rules of Assessment Participation; Confidentiality Guidelines; Video-Recording Permissions; Submission Requirements; Submission Attestations; Retake Policy), and score reporting policies (Reporting of Assessment Results; Retention of Scored Tasks; Voiding of Scores) (MPAL General Policies, n.d.). All these complications recall Aschbacher's research (1991), which uncovered similar technical requirements that compromised the promise of state PBAs.

To summarize, the PAL not only requires a great deal of writing and preparation of submissions, but also the reading, analysis, and interpretation of complicated instructions; technologically challenging exercises (such as learning to use software for video viewing, taping, and editing as well as the Pearson ePortfolio system); distribution and collection of recording permissions; and careful consideration of the registration requirements and legal ramifications of each step. These requirements are incidental to the work of real school leaders. They detract from the administrative practicum and, for candidates, they invite irritation and possibly error on the task submission, which subverts the validity of the PAL scores as a reflection of readiness for school leadership.

Numerical ratings. The *Standards* (2014) state, "Studies of response processes are not limited to the test taker. Assessments often rely on observers or judges to record and/or evaluate test takers' performances or products" (p. 15). This area is concerned with the ways in which scorers avoid relying upon "irrelevant or extraneous factors" (Goodwin & Leech, 2003, p. 184). The Commonwealth employed approximately 30 scorers during the field trial (Orr, Pecheone et al., 2016) comprised of "current or former school and district leaders" complemented by a small number of educational administration faculty (Losee & Orr, 2015, para. 6). The scoring rubric for the PAL forced scorers to look at the submission through a preconceived window, which excluded certain kinds of information and drew attention to others. In their evaluation of a similar scoring practice for the National Board for Professional Teaching Standards, Delandshire and Petrovsky (1998) noted,

> We observed that when assessors were asked to write an interpretive summary of a performance, they took more notes in viewing and reading it and used those notes as the basis for their interpretation. When the focus was on rating, however, they tended to look for features of the performance that were similar to or different from the general description contained in the rubrics, and they decided on the ratings very early on, often before having viewed or read the entire performance…. (p. 21)

The PAL scorers do not write interpretative summaries; they rate the candidate submission for each task on a scale of one to four in 30 to 45 minutes. For example, Task 1 includes about 6000 narrative words plus categorical data. Scorers must work quickly while looking for features stipulated

by the rubrics, which "force fit" performances into general categories (Delandshere & Petrosky, 1998, p. 21). The rating process streamlines the evaluation, but eliminates details.

This is particularly problematic when a candidate lacks sound writing skills. Some candidates organize their writing to match the scoring rubric so that each part of the task submission approximately matches one dimension of the scoring rubric, which makes scoring easier. However, this is not a requirement of the test. For example, for Task 1, candidates are allowed to address three dimensions, eight indicators, and over 20 specific rating points anywhere in the four written submissions. Understandably, a poorly organized submission, especially if the composition is not strong, would frustrate a scorer who has only 45 minutes and could easily overlook important answers or lose patience with the search. This suggests that poor writing skills could detract from the validity of the PAL as a test of initial school leadership.

As noted in the literature review, the PAL research team also noted problems with their own rating rubrics. First, the team found a preponderance of scores at Level 3 (Meeting) on all four tasks during the field trial and offered three possible explanations:

> Such results may imply that the rubric levels lead most submissions to 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. (Orr, Pecheone et al., 2016, p. 66)

This clustering of scores at Level 3 had an effect on inter-rater reliability; the team noted the "problematically low" reliability scores for Task 3 and explained, "This is likely due to the very low variance between candidates on Task 3 performance" (Orr, Pecheone et al., 2016, p. 76). However, in a nod to the problem of halo effects on multi-dimensional scoring rubrics (Humphry & Heldsinger, 2014), the authors also admitted that "scorer effects…could induce an inflated correlation among [dimension] scores from the same task (Orr, Pecheone et al., 2016, p. 70). Therefore, both the rating rubric and scorer effects might lead to higher-than-expected correlations across the dimensions of a single task, as well as lower-than-expected variance on given task scores, in addition to lower-than-expected inter-rater reliability rates for a given task. The lack of precision is obvious. One interpretation for this confusion, which is supported by Dunbar et al. (1991) and Humphry and Heldsinger (2014), is that rating rubrics fall prey to reductionist tendencies and fail to accurately assess complex performances.

In summary, then, we find that the PAL invites construct-irrelevant behaviors on the part of candidates (who wrestle with complicated reading, writing, and preparation requirements) as well as scorers who must rate candidate performance through the preconceived window of the scoring rubric (without ever actually observing any part of the performance). Surely, this calls into question the validity of the use and interpretation of the PAL.

Evidence Based on Internal Structure

In general, if test scoring is simplified so that only a low level of judgment is required then measures of inter-rater reliability may be less meaningful than internal consistency measurements. The *Standards* (2014) state, "Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based" (p. 16). When a test measures several constructs within a domain, it is useful to know how well the various constructs correlate with one another and with the overall domain. Does the test "hang together" well?

The PAL research team examined construct validity at the level of indicators, dimensions, and tasks (Orr, Pecheone, Hollingworth, et al., 2017) and decided that the tasks, elements, and indicators "work well as a combined measure…." (Orr et al., 2016, p. 83). Yet it is difficult to determine whether the results of the technical report are attributable to the task content or the response processes. In other words, candidates may simply be learning to master the logistical requirements of the test. When content validity is challenged, then construct validity is undermined. The author found in his own program that if students were provided a lot of "hand-holding" on Task 1 in terms of attending to anonymity, the stipulations of the scoring rubrics, word limits, formatting, and uploading directions, then they were able to work largely on their own on the remaining tasks. In other words, they appeared to be mastering a complex set of directions rather than learning skills necessary to beginning school leadership, which challenges the stated purpose and the validity of the PAL.

Evidence Based on Relations to Other Variables

The *Standards* (2014) state, "Use of existing evidence from similar tests and contexts can enhance the quality of the validity argument, especially when data for the test and context in question are limited" (p. 13). For many experimental studies, correlational studies, and criterion-group or known-group comparison studies, this is the most common approach to estimate validity (Goodwin & Leech, 2003). The PAL is too new for this kind of comparison but might be usefully compared to the CAT or the SLLA in the future.

This PAL investigation does invite comparison with edTPA, the student teacher performance assessment that is now used by more than 70% of American teacher preparation programs (Greenblatt & O'Hara, 2015).[7] Scholars from the Stanford University Center for Assessment, Learning and Equity (SCALE), who developed edTPA, also contributed to the development and validity testing of the PAL and, similar to the PAL, edTPA requires "a lengthy 'instructional commentary' of 40 to 60 pages" (Greenblatt & O'Hara, 2015, p. 57). Critics have suggested that edTPA relies too heavily on "candidates' reading, writing, and technological skills" (Greenblatt & O'Hara, 2015, p. 59) and shifts candidate's focus to test preparation instead of teaching. Some question whether edTPA privileges financially advantaged candidates and institutions (Au, 2013; Greenblatt & O'Hara, 2015; National Association of Multicultural Education, 2014). In truth, DESE considered and rejected edTPA in 2014 over concerns about cost, the failure to recognize feedback from program supervisors and supervising practitioners, the limitations as a formative assessment, and the lack of alignment with the state performance evaluation document (Pat Paugh, personal communication, April 17, 2017); ironically, the PAL suffers from the same problems. Instead, DESE created a new test of teacher readiness that addresses these shortcomings and does not rely upon paid, anonymous scorers (Ikemoto, Keleman, Tucker, & Young, 2016). Given the similarities between edTPA and the PAL, it is surprising that DESE rejected one and accepted the other. The teacher test invites commercial opportunism: two companies now offer online assistance to teachers facing the edTPA; one company promised, "You send us your videos, lessons, and student work. We do the rest" (Sawchuk, 2015). Of course, the high-stakes PAL could invite the same.

The interpretation of the PAL, which is intended to predict readiness for school leadership, invites comparison to other evaluation measures that have been used in the past. For example, in the process of hiring new school leaders, districts often consider letters of recommendation and interviews with the search committee. New approaches include additional interviews with key

[7] In fact, the PAL design team also made this comparison (Orr & Hollingworth, 2018).

stakeholder groups (teachers, community members); authentic tasks (data reviews, building walk-throughs, teacher observations) that call for initiative, creativity, and teamwork; and even 360 degree performance assessments (Clifford, 2010). On the surface, the PAL has the appearance of a checkpoint on the way to school leadership that will prevent unqualified candidates from stepping into administrative positions but in reality, the checkpoint is quite porous. In Massachusetts, the PAL can be taken repeatedly; candidates can revise and resubmit any task that does not meet the benchmark (at additional cost). The guidelines on independent work are vague; for this reason, candidates seek lots of guidance, especially when they fail a task. Given these extenuating factors, it is difficult to see how the PAL would provide a test of school leadership readiness that is more valid than current measures.

Evidence for Validity and Consequences of Testing

Messick (1994) first proposed consequences as a validity consideration: "if both positive and negative aspects, whether intended or unintended, are not meaningfully addressed in the validation process, then the concept of validity loses its force as a social value" (Messick, 1994, p. 22). The *Standards* (1999) soon added this criterion. There are few guidelines on how to measure this; one suggestion is for focus groups to investigate the consequences. The PAL research team employed a bias review committee and examined test results for disparate impact (Orr, Pecheone, Hollingworth, et al., 2017) as reported in the literature review.

Descriptive vs. prescriptive consequences. Consequences can be divided into *descriptive* and *prescriptive* consequences. For example, a candidate who fails to achieve proficiency on Task 3 is descriptively deemed *not ready* and does not receive an administrator's license. Prescriptive consequences might include recommendations to the candidate as well as the preparation program but DESE denies responsibility for any formative testing purpose and presents the PAL purely as a summative assessment (Jennifer Briggs, personal communication, May 23, 2016). In contrast, California state leaders, in conjunction with the University of San Diego, used a pilot study of Tasks 1 and 2, as a formative assessment (rather than a high-stakes licensure test) (Orr, Hollingworth, & Cook, 2016). The authors concluded, "The assessments could be used formatively as embedded leadership preparation for teacher leaders and aspiring leaders" (Orr, Hollingworth, et al., 2016, p. 15).

Intended vs. unintended consequences. According to the *Standards* (2014), a validity test should address consequences "intended by the test developer" and other consequences "beyond the interpretation or use of scores intended by the test developer" (p. 19). In a follow-up study on the implementation year (2016-2017), PAL researchers explored the impact of the PAL on program effectiveness. They asked program faculty how effective their program was in preparing candidates to complete the requirements for each task and they asked candidates to rate how challenging they found the tasks (Orr & Hollingworth, 2018). See Table 4:

Table 4
Student and Faculty Responses to the Impact of the PAL on Program Effectiveness

Task	% of Students who Responded "Challenging or Very Challenging"	% of Faculty Who Responded "Effective or Highly Effective"
1	53	87
2	54	91
3	38	90
4	63	88

Source: (Orr & Hollingworth, 2018)

There are obvious validity challenges when the inventor of a product also conducts the follow-up impact study. Respondents are influenced by who is gathering the evidence and the investigators are more likely to interpret the evidence positively. In this case, the authors determined that the results, in which the strength of faculty response exceeded the student response, suggested that programs had responded vigorously to the test and students were thereby able to succeed on a difficult test. In short, programs and students were improving because of the test. However, there might be another interpretation. Approximately half the students did not find the PAL challenging or very challenging; possibly it was just a tedious bump in the road. Meanwhile, faculty largely agreed (over 87%) that their program was effective or highly effective in preparing students for the tasks but there was apparently nothing in the questions that directed faculty to focus primarily on PAL-inspired changes in their programs. Faculty are likely to view their programs favorably and their responses would reflect that opinion and not just the adjustments made to address the PAL.

It is quite possible that programs made cosmetic adjustments to address students' concerns while not making substantial changes in the focus, quality, or rigor of programs they highly regarded. The changes were superficial and inconvenient, but not substantial. An independent study examined ten preparation programs that participated in the Massachusetts field trial to capture how the advent of the PAL was affecting program design (Leonard, 2017). Each program had just completed a rigorous re-authorization process in response to new state leadership standards, which required significant re-invention, and achieved DESE approval. According to eight out of ten program directors, the arrival of the standards-light PAL was disruptive and prompted changes in course sequencing, faculty assignments, course syllabi, course assignments, classroom teaching, practicum assignments, and/or graduation standards—despite the fact that most programs already believed they were doing a great job. By their own admission, they made changes to address student anxieties about the test rather than to improve the program quality. For example, some programs changed their course sequence to address the PAL early in the program; some changed major assignments to better align with PAL tasks. The PAL investigators interpreted these changes positively: The PAL assessments "have had had a modest, positive influence on program content through improvements made in alignment and sequencing, the addition of topics related to specific tasks…." (Orr & Hollingworth, 2018, pp. 17–18). The presupposition here, of course, is that the tasks are now the standard for high-quality curriculum (instead of ISLLC as will be discussed below). According to Leonard (2017), program directors decried the narrowing effect of the test and the loss of academic freedom. They found that the test promoted a credentialing function over the educational function of higher education (Leonard, 2017). Some programs wrestled with whether to require successful completion of the PAL as a graduation requirement. Students reported that the PAL consumed an average of 6.67 hours of class time and hundreds of hours of the practicum. Twenty percent of students specifically lamented the boring redundancy of the test. In effect, the high-stakes nature of the PAL tended to push both students and instructors toward test preparation. Of course, similar questions arise in this case of how the identity of the author (Leonard, 2017) might influence responses and interpretations. However, at the very least, the results call into question the educative value of the PAL for both students and faculty. To summarize, the PAL, especially as used in Massachusetts, lacks prescriptive consequences, which might be educative, and imposes a host of other unintended consequences, which serve to narrow the curriculum of preparation programs as well as take important time away from practicum experiences.

Multicultural Validity

In her 2010 article, Kirkhart offered nine strategies for determining the cultural location of evaluation theory, which included an examination of the authors and the process of theory

development. For PAL development, there were five committees (design, bias, standards, content validity, and technical advisory) with 49 members representing K-12 school leadership and preparation programs fairly evenly along with two DESE members. As noted earlier, program leaders in general are beholden to DESE, which grants licensure authority, and were waiting to learn in what ways the test might be used as a high-stakes exam and, possibly, to evaluate programs. The presence of two DESE members suggests an imbalance of power. In addition, the design and content validity committees were 90% White, but the bias review committee of nine members included five non-White members with a balance of both genders. Altogether, the committees did not reflect the demographic characteristics of K12 students in Massachusetts nor the candidates in a typical urban principal preparation programs.

Kirkhart's nine strategies also included attention to language. For example, Davis, Gooden, and Micheaux (2015) used critical document analysis to demonstrate how the ELCC and 2008 ISLLC standards sidestepped issues of race and culture. Using a similar method, this study examined the PAL *Candidate Assessment Handbook* (MA-DESE, 2016b) for race-based terms such as race, racial, ethnic, ethnicity, diverse, diversity, Black, African-American, Hispanic, and Native American. The words race, racial, ethnic and ethnicity were used only once each in the handbook, specifically in a template used by candidates to record the gender, race and ethnicity of the teacher and students in a classroom observation (MA-DESE, 2016b, p. 79). The words African-American, Hispanic and Native American occurred exclusively in parenthetical text or footnotes as an explanation of the term "federally designated priority student groups" (pp. 3, 4, 12, 23, 41, 87), which was a criterion for data analysis. The terms diverse and diversity were used to describe variations in student learning needs (p. 61), community interests (pp. 83, 84, 86), and abstractly with families and community groups (pp. 89, 92, 93). The PAL is silent on the topic of individual or systemic racism and the terms above do not show up at all in the scoring rubrics. Kirkhart (2005) noted, "Validity is threatened to the extent that culture is ignored or diversity variables are included as simplistic, atheoretical stereotypes" (p. 24).

Kirkhart (2010) also asked readers to "Notice the scope of attention to culture" (p. 403). The PAL *Candidate Assessment Handbook* (MA-DESE, 2016b) used culture many times but always to mean one of three things:

> School culture has three components: (1) the professional learning culture (teacher-teacher relationships), (2) the student culture (teacher-student and student-student relationships), and (3) the culture of family and community engagement (school staff, family, and community relationships). (p. 12)

Again, the significance of race, ethnicity, and language diversity and the possibility that students, families, or teachers might manifest multiple cultures was erased. As a result, candidates are not prompted to develop or exercise cultural competence. This is particularly problematic in light of the new ISLLC standards (National Policy Board for Educational Administration, 2015) where cultural competence is embedded in standards 2, 3, 4, 5, and 8. This new emphasis calls into question the relevance and validity of the PAL, as it is currently constructed, as a measure of readiness for modern school leadership.

The requirements of the PAL tasks are not contextually sensitive. For example, a case study of a Massachusetts urban district partnership for principal preparation (Leonard & Daly, 2017), where full-time working teachers were the most common candidates, found that a disproportionate number of minority teachers received their teacher preparation through an alternative route, which left them with lower licensure status and a longer route to the stability of professional status. School authorities often assigned them to schools that were subject to turnaround, placed them in the most

difficult classrooms, and subjected them to transfer and/or termination more often. The rate of unsatisfactory evaluations for Black teachers was five times higher than the White rate while the Hispanic rate was twice as high (Birnbaum, 2013, November 21; Vaznis, 2013, April 24). All these factors intensified the challenges of the PAL.

The PAL design team conducted a face validity survey of school leadership candidates but admitted that "most survey respondents were female, White" (Orr, Pecheone, Snyder, et al., 2017, p. 7). Minority educators faced unique challenges in building professional learning groups. They were more likely to be moved between schools, which interrupted their practicum and broke up the logical sequence of the four tasks. Many minority candidates in turnaround schools and charter schools faced a longer school day, amplified professional development, a top-down bureaucracy, and a pre-determined school vision, which limited opportunities for personal leadership as required by the PAL and hindered their efforts to achieve proficiency. Evidently, the PAL bias committee sensed a problem when they worried that "some candidates might have less access than others to information and support to complete the tasks" (Orr, Pecheone, Snyder, et al., 2017, p. 15). At the same time, however, these minority teachers were receiving valuable experiences in urban education such as how to schedule an extended day, set up inclusive classrooms, teach in a dual-language program, work in an under-resourced school, practice school turnaround, differentiate instruction for diverse students and mid-year arrivals, and exercise cultural competence. Unlike the PAL, which did not measure these things, critical race theory "recognizes that the experiential knowledge of people of color is legitimate and critical to understanding racial subordination" (Parker & Villalpando, 2007, p. 520). Looking through the lens of multicultural validity, the PAL fails to attend to issues of race, ethnicity, and language diversity. The constrained construction of culture does nothing to encourage cultural competence in aspiring leaders and, in fact, would appear to overlook and under-report the actual, unique competencies of urban educators.

Discussion

This paper considered the validity of a performance based assessment, such as the PAL, as a measure of readiness for initial school leadership by using the *Standards* (2014) in addition to Kirkhart's (2005, 2010) concept of multicultural validity. While the test designers paid particular attention to content and construct validity and reliability (Orr, Pecheone, Hollingworth, et al., 2017; Orr, Pecheone et al., 2016), this paper reexamined these areas and also invited attention to response processes, internal structure, relation to other variables, consequences, and issues of race and culture. In particular, this study raised questions in regards to test content (whether the PAL reduces rigorous state leadership standards to low-fidelity alternatives that lacked authenticity and were cognitively constrained) and response processes that distract from the purpose of the PAL as a direct measure of school leadership. Significantly, the PAL falls short of the central feature of a PBA, which is the "show me" aspect (Resnick & Resnick, 2013, p. 26) of the test. Anonymous scorers review a written narrative of the performance, while relevant human judges, such as instructors and supervising practitioners who observe the candidate regularly, are discounted. From a practitioner's standpoint, there are reasons to question whether the 43-page writing requirement is a suitable substitute for other leadership activities in the administrative practicum. The individualistic nature of the exam itself contradicts modern conceptions of school leadership that emphasize collaboration and shared leadership. Other concerns include the unintended consequences (including programmatic changes and student disillusionment), and the multicultural validity of a test that appears to present a color-blind version of leadership. The use of the PAL as a high-stakes test

only magnifies these problems and tends to discourage the educative function of preparation programs while promoting test preparation and certification instead.

Validity is "always a matter of degree rather than an all-or-none judgment" (Linn, 1994, p. 6) so one cannot state conclusively that the PAL is valid or not valid. The PAL is a new test, still subject to revision and experimentation. The findings in this paper are tentative and reflect the early experiences and insights of candidates and program administrators. This validity discussion also serves as a close-up examination of the test for those who are considering adoption of the PAL or a similar PBA. In the end, it is test users who can best determine whether the test seems direct, authentic, informative, and useful as a learning improvement tool. Validity studies are valuable because they can lead to important modifications or improvements. "The validity argument may indicate the need for refining the definition of the construct, may suggest revisions in the test or other aspects of the testing process, and may indicate areas needing further study" (AERA et al., 2014, p. 21). One way to move forward is to consider other possible applications of this test.

The use of the PAL as a formative assessment (instead of a high-stakes licensure exam) would enhance the positive educative value of the PAL for students and the monitoring function for programs. In their evaluation of PBAs for school leadership for the American Institutes for Research, Condon and Clifford (2012) called for transparency in the underlying constructs, the disaggregation of test results, and the transferal of administrative and analytic controls to local educators. The California PAL pilot study honored all three recommendations; even the PAL researchers agreed, "The assessments could be used formatively as embedded leadership preparation for teacher leaders and aspiring leaders" (Orr, Hollingworth, et al., 2016, p. 15). Of course, this would undermine the benchmarking, credentialing value. In addition, the test could be modified to help candidates pay closer attention to issues of race, language, and culture and to develop culturally competent leadership (Furman, 2012; Shields, 2004; Theoharis, 2010).

Quality control concerns suggest that a rigorous test is necessary to limit access to the principal license. Winch (2015) seemed to argue in favor of tests, such as the PAL, by insisting that the assessment process be "sufficiently rigorous to reduce the risk to the public as much as possible" (p. 101). However, Lum (2015) expanded this argument by distinguishing between a *prescriptive* and an *expansive* mode of assessment where the latter focuses broadly on accountability. In the prescriptive mode, predetermined elements are judged in a binary fashion; in contrast, in the expansive mode, the assessors can employ "judgements of significance" and are "at liberty to expand the focus of their attention to take account of any available evidence" (p. 123). The advantage of the prescriptive mode, of which the PAL is an example, is that this allows for summary judgments, regardless of the context, and ready comparison of candidates and their preparation programs. In contrast, the expansive mode draws on the "fullest range of evidence" (Lum, 2015, p. 124)—triangulating evidence, which can be varied in nature and contextually sensitive—to provide a more comprehensive picture of the candidate including strengths, weaknesses, and areas for growth. The possibility of cheating, coaching, and cutting corners is far more difficult with the expansive mode. Lum (2015) asserted, "It is precisely when the stakes are high, when there is an obligation to achieve the very best estimation of knowledge, that it becomes imperative to employ assessment in the expansive mode" (p. 125). This is how we make important decisions outside of K12 education, such as criminal trials or university faculty reviews. If we insist on using a prescriptive approach, then we argue for a low risk and the inconsequential nature of the profession. If we agree that the principal's work is important, then we should argue for an expansive approach. The high-stakes nature of the PAL, exclusive to Massachusetts, should be diminished; the PAL should be one piece among many pieces of evidence. In this case, the validity challenges of the PAL would be balanced by triangulating evidence from other sources.

Conclusion

PBAs were first introduced in an attempt to break free from a positivistic approach to educational assessment. Over 20 years ago, Delandshere and Petrovsky (1994) wrote that "the trend toward new forms of performance assessment… can be thought of as an attempt to develop language, methods, and traditions around a different conception of knowledge and within a different paradigm" (p. 18). The PAL is the latest attempt to employ the strengths of the PBA to measure readiness for school leadership. However, the PAL also demonstrates the inherent limitations of PBAs, particularly when they are used as high-stakes tests for complex performances. In many ways, the PAL is a return to a positivistic approach, particularly in the scoring rubrics, thus defeating one of the primary hopes of the PBA movement.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational Psychological Testing. (1999). *Standards for educational and psychological testing*: American Educational Research Association.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational Psychological Testing. (2014). *Standards for educational and psychological testing*: American Educational Research Association.

Anderson, E., & Reynolds, A. L. (2015). *A policy-maker's guide: Research-based policy for principal preparation program approval and licensure*. Retrieved from Charlottesville VA: http://3fl71l2qoj4l3y6ep2tqpwra.wpengine.netdna-cdn.com/wp-content/uploads/2014/05/UCEA-State-Policy-Report-website-version-Nov2015-v2.pdf

Aschbacher, P. R. (1991). Performance assessment: State activity, interest, and concerns. *Applied measurement in Education, 4*(4), 275-288. http://dx.doi.org/10.1207/s15324818ame0404_2

Au, W. (2013). What's a nice test like you doing in a place like this. *Rethinking Schools, 27*(4), 22–27.

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

Bass, B. M. (1999). Two decades of research and development in transformational leadership. *European journal of work and organizational psychology, 8*(1), 9–32. http://dx.doi.org/10.1080/135943299398410

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complerrity of science assessments. *Educational Measurement: issues and practice, 17*(3), 37–45. http://dx.doi.org/10.1111/j.1745-3992.1998.tb00627.x

Birnbaum, S. (2013, November 21). Boston Teachers Union calls teacher evaluations biased. *WGBH News*. Retrieved from https://news.wgbh.org/post/boston-teachers-union-calls-teacher-evaluations-biased

California Commission on Teacher Credentialing. (2013). *Update on Administrator Performance Assessments*. Retrieved from http://www.ctc.ca.gov/commission/agendas/2013-09/2013-09-4e.pdf.

Chester, M. D. (2012). *Performance Assessment for Leaders Project*. Malden MA: Massachusetts Department of Elementary and Secondary Education. Retrieved from http://www.doe.mass.edu/news/news.aspx?id=7112.

Clifford, M. (2010, February). *Hiring quality school leaders: Challenges and emerging practices*. Naperville, IL: Learning Point Associates.

Condon, C., & Clifford, M. (2012). *Measuring principal performance: How rigorous are commonly used principal performance assessment instruments?* Retrieved from: http://www.air.org/sites/default/files/downloads/report/Measuring_Principal_Performance_0.pdf

Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st century learning*. Hoboken NJ: John Wiley & Sons.

Darling-Hammond, L., Meyerson, D., LaPointe, M., & Orr, M. T. (2010). *Preparing principals for a changing world: Lessons from effective school leadership programs*. San Francisco CA: Jossey-Bass.

Davis, B. W., Gooden, M. A., & Micheaux, D. J. (2015). Color-blind leadership A critical race theory analysis of the ISLLC and ELCC standards. *Educational Administration Quarterly, 51*(3), 335–371. http://dx.doi.org/10.1177/0013161X15587092

Delandshere, G., & Petrosky, A. R. (1994). Capturing teachers' knowledge: Performance assessment a) And post-structuralist epistemology, b) From a post-structuralist perspective, c) And post-structuralism, d) None of the above. *Educational Researcher, 23*(5), 11–18. http://dx.doi.org/10.2307/1177028

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 27*(2), 14–24. http://dx.doi.org/10.3102/0013189X027002014

Diem, S., & Young, M. D. (2015). Considering critical turns in research on educational leadership and policy. *International Journal of Educational Management, 29*(7), 838–850. http://dx.doi.org/10.1108/IJEM-05-2015-0060

Diem, S., Young, M. D., Welton, A. D., Mansfield, K. C., & Lee, P.L. (2014). The intellectual landscape of critical policy analysis. *International Journal of Qualitative Studies in Education, 27*(9), 1068–1090. http://dx.doi.org/10.1080/09518398.2014.916007

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied measurement in Education, 4*(4), 289–303. http://dx.doi.org/10.1207/s15324818ame0404_3

Educational Testing Service. (2018). *SLS State Requirements*. Retrieved from https://www.ets.org/sls/states/

Fry, B., Bottoms, G., O'Neill, K., & Walker, S. (2007). *Schools need good leaders now: State progress in creating a learning-centered school leadership system.* Retrieved from: http://www.wallacefoundation.org/knowledge-center/Documents/Schools-Need-Good-Leaders-Now.pdf

Furman, G. (2012). Social justice leadership as praxis: Developing capacities through preparation programs. *Educational Administration Quarterly, 48*(2), 191–229. http://dx.doi.org/10.1177/0013161X11427394

Goleman, D. (2000). Leadership that gets results. *Harvard Business Review, 78*(2), 78–90.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new Standards for Educational and Psychological Testing: Implications for measurement courses. *Measurement and evaluation in Counseling and Development, 36*(3), 181–192. http://dx.doi.org/10.1080/07481756.2003.11909741

Greenblatt, D., & O'Hara, K. E. (2015). Buyer beware: Lessons learned from EdTPA implementation in New York State. *Teacher Education Quarterly, 42*(2), 57–67.

Grissom, J. A., Mitani, H., & Blissett, R. S. L. (2017). Principal licensure exams and future job performance: Evidence from the School Leaders Licensure Assessment. *Educational Evaluation and Policy Analysis, 25*. http://dx.doi.org/3102/0162373716680293

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher, 43*(5), 253–263. http://dx.doi.org/10.3102/0013189X14542154

Ikemoto, G., Keleman, M., Tucker, P., & Young, M. (2016). *Guide to the state evaluation of principal preparation programs*. Washington, DC: New Leaders, University Council for Educational Administration.

Kirkhart, K. E. (2005). Through a cultural lens: Reflections on validity and theory in evaluation. In S. Hood, R. Hopson, & H. Frierson (Eds.), *The role of culture and cultural context: A mandate for inclusion, the discovery of truth, and understanding in evaluative theory and practice* (pp. 21–38). Greenwich CT: Information Age Publishing.

Kirkhart, K. E. (2010). Eyes on the prize: Multicultural validity and evaluation theory. *American Journal of Evaluation, 31*(3), 400–413. http://dx.doi.org/10.1177/1098214010373645

Leithwood, K. (2010). *How the leading student achievement project improves student learning: An evolving theory of action*. Retrieved from: http://www.curriculum.org/LSA/files/LSATheoryofAction.pdf

Leithwood, K., Louis, K. S., Anderson, S., & Wahlstrom, K. (2004). *How leadership influences student learning*. Retrieved from http://www.wallacefoundation.org/knowledge-center/school-leadership/key-research/Documents/How-Leadership-Influences-Student-Learning.pdf

Leonard, J. (2017). Early effects of the Massachusetts Performance Assessment for Leaders. *Journal of Research on Leadership Education*, under review.

Leonard, J., & Daly, C. (2017). Partnering for a diverse principal preparation pipeline. In R. M. Reardon & J. Leonard (Eds.), *Exploring the community impact of research-practice partnerships in education* (Vol. 1, pp. 1-35). Charlotte NC: Information Age Publishing.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4–14. http://dx.doi.org/10.3102/0013189X023009004

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21. http://dx.doi.org/10.3102/0013189X020008015

Losee, E., & Orr, M. T. (2015). *Performance Assessment for Leaders: Information from the 2014–15 field trial and policies and procedures for program year 2015–16*. Malden MA: Massachusetts Department of Elementary and Secondary Education Retrieved from http://www.doe.mass.edu/news/news.aspx?id=21237.

Lum, G. (2015). Afterword: Can the two positions be reconciled? In G. Lum (Ed.), *Key debates in educational policy: Educational assessment on trial* (pp. 107–134). London: Bloomsbury.

Manna, P. (2015). *Developing excellent school principals to advance teaching and learning: Considerations for state policy*. Retrieved from New York: http://www.wallacefoundation.org/knowledge-center/school-leadership/state-policy/Documents/Developing-Excellent-School-Principals.pdf

Massachusetts Department of Elementary and Secondary Education. (n.d.). *Toolkit for school leader mentors*. Malden MA: Author. Retrieved from http://www.doe.mass.edu/pal/LeaderToolkit.pdf.

Massachusetts Department of Elementary and Secondary Education. (2014). *MA Performance Assessments for Leaders scorer manual*. Malden, MA: Author.

Massachusetts Department of Elementary and Secondary Education. (2016a). *Administrative field guide for leadership preparation programs*. Malden MA: Author.

Massachusetts Department of Elementary and Secondary Education. (2016b). *Candidate assessment handbook*. Malden, MA: Author.

Massachusetts Department of Elementary and Secondary Education. (2016c). *Overview of administrator routes to initial licensure 603 cmr 7.00 & guidelines for the administrative apprenticeship/internship and panel review routes*. Retrieved from http://www.doe.mass.edu/licensure/academic-prek12/panel-review-administrator-routes.docx

Massachusetts Department of Elementary and Secondary Education. (2017). *Massachusetts Performance Assessment for Leaders (PAL)*. (2017). Retrieved from http://www.doe.mass.edu/pal/

Massachusetts Department of Elementary and Secondary Education. (2018). *Preparation Plan Programs: Post-Baccalaureate*.Retrieved from https://gateway.edu.state.ma.us/elar/licensurehelp/ProgramSearchPostControl.ser

Massachusetts Performance Assessment for Leaders. (n.d.). Candidate resources. Retrieved from http://www.ma-pal.nesinc.com/PageView.aspx?f=GEN_CandidatesResources.html

Massachusetts Performance Assessment for Leaders. (n.d.). General policies. Retrieved from http://www.ma-pal.nesinc.com/PageView.aspx?f=GEN_Policies.html

Massachusetts Performance Assessment for Leaders. (n.d.). *What is PAL*. Retrieved from http://www.ma-pal.nesinc.com/PageView.aspx?f=GEN_AboutPAL.html

Mendels, P. (2012). The effective principal. *Journal of Staff Development, 33*(1), 54–58.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23. http://dx.doi.org/10.3102/0013189X023002013

Mislevy, R. J., & Knowles, K. T. (2002). *Performance assessments for adult education: Exploring the measurement Issues:: Report of a workshop*. Washington D.C.: National Academies Press.

Murphy, J., Elliott, S. N., Goldring, E., & Porter, A. C. (2007). Leadership for learning: a research-based model and taxonomy of behaviors. *School Leadership & Management, 27*(2), 179–201. http://dx.doi.org/10.1080/13632430701237420

National Association of Multicultural Education. (2014). *NAME position statement on the edTPA*. Retrieved from https://www.nameorg.org/docs/Statement-rr-edTPA-1-21-14.pdf

National Policy Board for Educational Administration. (2015). *Professional Standards for Educational Leaders 2015*. Reston VA: Author.

Orr, M. T., Beaudin, B., Pecheone, R., Snyder, J., Murphy, J., Buttram, J., & Hollingworth, L. (2015). *Candidate Assessment Handbook*. Malden MA: MA Department of Elementary and Secondary Education.

Orr, M. T., & Hollingworth, L. (2018). How performance assessment for leaders (PAL) influences preparation program quality and effectiveness. *School Leadership & Management*, 1–22. http://dx.doi.org/10.1080/13632434.2018.1439464

Orr, M. T., Hollingworth, L., & Cook, J. (2016, November). *Large-scale Performance Assessment for Leaders in California: A pilot study*. Paper presented at the annual convention of the University Council for Educational Administration, Detroit, MI.

Orr, M. T., Pecheone, R., Hollingworth, L., Beaudin, B., Snyder, J., & Murphy, J. (2017). The performance assessment for leaders: Construct validity and reliability evidence. *Journal of Research on Leadership Education*, 1–23.

Orr, M. T., Pecheone, R., Nayfeld, I., Shear, B., Hollingworth, L., & Beaudin, B. (2016). *Massachusetts Performance Assessment for Leaders (PAL) technical report: Summary of validity and reliability studies for 2014–15 field trial of PAL*. New York: Bank Street College of Education. Retrieved from http://www.doe.mass.edu/pal/TechnicalReport.docx.

Orr, M. T., Pecheone, R., Snyder, J. D., Murphy, J., Palanki, A., Beaudin, B., . . . Buttram, J. L. (2017). Performance assessment for principal licensure: Evidence from content and face validation and bias review. *Journal of Research on Leadership Education*. http://dx.doi.org/10.1177/1942775117701179

Parker, L., & Villalpando, O. (2007). A race (cialized) perspective on education leadership: Critical race theory in educational administration. *Educational Administration Quarterly, 43*(5), 519–524. http://dx.doi.org/10.1177/0013161X07307795

Portin, B. S., Knapp, M. S., Dareff, S., Feldman, S., Russell, F. A., Samuelson, C., & Yeh, T. L. (2009). *Leadership for learning improvement in urban schools*. Retrieved from: http://www.wallacefoundation.org/knowledge-center/Documents/Leadership-for-Learning-Improvement-in-Urban-Schools.pdf

Resnick, D. P., & Resnick, L. B. (2013). Performance assessment and the multiple functions of educational assessment. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: Promises, problems, and challenges* (pp. 46–61). New York, N.Y.: Routledge.

Sawchuk, S. (2015, October 21). Are new teacher tests vulnerable to cheating? *Education Week, 35,* 1, 12.

Shelton, S. (2011). *Strong leaders strong schools: 2010 school leadership laws*. Retrieved from: http://www.wallacefoundation.org/knowledge-center/Documents/2010-Strong-Leaders-Strong-Schools.pdf

Shelton, S. (2012). *Preparing a pipeline of effective principals: A legislative approach* (1580246745). Retrieved from: http://www.ncsl.org/documents/educ/PreparingaPipelineofEffectivePrincipalsFINAL.pdf

Shields, C. M. (2004). Dialogic leadership for social justice: Overcoming pathologies of silence. *Educational Administration Quarterly, 40*(1), 109–132. http://dx.doi.org/10.1177/0013161X03258963

Taylor, S. (1997). Critical policy analysis: Exploring contexts, texts and consequences. *Discourse: Studies in the cultural politics of education, 18*(1), 23–35. http://dx.doi.org/10.1080/0159630970180102

Theoharis, G. (2010). Disrupting injustice: Principals narrate the strategies they use to improve their schools and advance social justice. *Teachers College Record, 112*(1), 331–373.

Vaznis, J. (2013, April 24). Union says teacher evaluation plan has race bias. *Boston Globe*. Retrieved from http://www.bostonglobe.com/metro/2013/04/23/boston-union-officials-black-and-hispanic-teachers-disproportionately-targeted-under-new-evaluation-system/LCghntHAh8zM2R8qPmYrzM/story.html

Webb, N. L., Horton, M., & O'Neal, S. (2002, April). *An analysis of the alignment between language arts standards and assessments for four states*. Paper presented at the annual convention of the American Educational Research Association, New Orleans.

Winch, C. (2015). The nature and purpose of educational assessment – A response to Andrew Davis. In G. Lum (Ed.), *Key debates in educational policy: Educational assessment on trial* (pp. 58–107). London: Bloomsbury.

Appendix

Summary of the Four PAL Assessment Tasks[8]

Task 1 Leadership through a Vision for High Student Achievement

This task asks candidates to focus on two pillars of highly effective schools: the instructional program (curriculum, instruction, and assessment) and school culture (student culture, professional culture, and the culture of family engagement and community involvement). The candidate develops a school vision and improvement plan for one school- based priority area. Specifically, the candidate collects and analyzes quantitative and qualitative data on student performance, student and teacher relationships, and student and school culture; selects a priority area for focus; documents existing school programs, services, and practices; and develops a set of goals, objectives, and action strategies with input from school leaders and key stakeholder groups. The candidate also presents and receives feedback on the plan from relevant stakeholders.

Task 2 Instructional Leadership for a Professional Learning Culture

In this task, the candidate demonstrates 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. The candidate supports teachers in improving an existing curriculum, instructional approach, or assessment strategy. The candidate also documents the process, teachers' teamwork, and changes in practice.

Task 3 Leadership in Observing, Assessing and Supporting Individual Teacher Effectiveness

In this task, the candidate demonstrates instructional leadership skills by planning for a teacher observation, conducting the observation, analyzing the observation and student performance data, providing feedback, and planning support for an individual teacher. The candidate also documents the observation cycle as well as teacher feedback on the quality and use of the feedback and support planning process.

Task 4 Leadership for Family Engagement and Community Involvement

Here, the candidate gathers information related to family engagement and community involvement needs, develops a proposal, and implements one component of it with work group support. The candidate works collaboratively with a work group representing school leadership, staff, families and community members, and students (where appropriate) to select a priority area based on evidence of student strengths, interests, and needs. The candidate, with the work group, develops a comprehensive improvement proposal and implements and monitors the outcomes for one strategy.

[8] This information is quoted verbatim from Orr and Hollingworth (2018, p. 22).

About the Author

Jack Leonard
University of Massachusetts, Boston
jack.leonard@umb.edu
https://orcid.org/0000-0002-1848-9401
Jack Leonard retired in 2017 from the position of associate professor in the Leadership in Education department at UMass Boston. He was the former director of the graduate programs in educational administration and taught courses on leadership, data analysis, and the history of American urban education. Prior to joining the faculty in 2008, he served the Boston Public Schools as a teacher and then principal of an award-winning turnaround high school. He continues to research and write on school leadership and school partnerships as well as educational history.

education policy analysis archives

Volume 26 Number 163 December 10, 2018 ISSN 1068-2341