2023

An Investigation Of The Quality Of Performance Assessments And Implications Of A Grassroots Approach To Accountability Reform

Molly Sandling
College of William and Mary - School of Education, molly.sandling@gmail.com

An Investigation of the Quality of Performance Assessments and Implications of a Grassroots Approach to Accountability Reform

---

A Dissertation

Presented to the

The Faculty of the School of Education

The College of William & Mary in Virginia

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

By

Molly M. Sandling

March 2023

An Investigation of the Quality of Performance Assessments

and Implications of a Grassroots Approach to Accountability Reform

By

Molly M. Sandling

Approved March 2, 2023 by

Dr. Leslie Grant
Committee Member

Dr. Thomas Ward
Committee Member

Dr. Christopher Gareis
Chairperson of Doctoral Committee

Dedication

To my family:

My parents, Derry and Ginny, who never lost faith and kept encouraging me to persevere and Jacob, my light and inspiration, whose patient support and self-reliance allowed me to pursue this dream.

Acknowledgements

I am grateful for my dissertation advisor, Dr. Christopher Gareis, for his advice, guidance, and patience through the design of this study and the long writing process. Also thank you to my committee members, Dr. Leslie Grant and Dr. Tom Ward, for their support. My review team of social studies specialists generously and cheerfully gave of their time and expertise through the summer of 2022, reading and scoring the assessments in this study. Not only did they help fulfill a critical part of this study, but their good humor also made the process joyful. Finally, this study could not have happened without the 12 participants from across Virginia who took time out of their busy schedules to answer my interview questions. I appreciate their willingness to be open and share the work they and their divisions have done on LAAs. I thoroughly enjoyed each interview, and I ended each conversation energized and inspired by their passion for teaching and learning. With leaders like these the future of social studies education in Virginia is bright indeed.

Table of Contents

List of Tables

Abstract

In 2014 the Virginia General Assembly passed legislation replacing end-of-course, multiple-choice assessments with locally developed alternative assessments in five courses, including two middle school social studies courses. This policy allowed Virginia school divisions the autonomy to develop the format, quantity, and focus of their assessments to meet state accountability. Given the grassroots nature of this policy, there has been little oversight of these local alternative assessments (LAAs). Thus, this exploratory study sought to gain insight into how divisions approached the process of preparing for and developing local alternative assessments and Balanced Assessment Plans, as well as the quality of the assessments created. Through a professional referral sampling process 12 divisions were interviewed and submitted two assessments each to be evaluated against the Virginia Quality Criteria Tool for Performance Assessments (VQCT). The divisions in the study responded to the autonomy granted by the state by first engaging in on-going, quality professional development to build teacher capacity. Using a variety of templates and the VQCT, divisions involved teachers in the process of developing the set of LAAs. The division assessments focus on writing prompts centered around tasks authentic to the social studies which require deeper-learning competencies by students, but the structure of the implementation of the assessments are less consistent potentially lessening the quality of the assessments. The work of these divisions suggest that the success of a grassroots performance assessment policy requires quality, on-going professional development and thoughtful analysis of the assessments and their alignment to desired learning goals.

AN INVESTIGATION OF THE QUALITY OF PERFORMANCE ASSESSMENTS AND IMPLICATIONS OF A GRASSROOTS APPROACH TO ACCOUNTABILITY REFORM

CHAPTER 1

INTRODUCTION

The publication of *A Nation at Risk* in 1983 brought the desire for educational reform into the public debate, leading to increased testing for accountability and providing the opportunity to evaluate existing forms of assessment (Gong & Reidy, 1996; Strong & Sexton, 2000; Wren & Gareis, 2019). By the late 1980s, concern from the business community that students were not prepared for the workforce, combined with educator concerns over the effects of multiple-choice, norm-referenced testing on student learning, led to an increased focus on performance assessments (Khattri et al., 1998; Wren & Gareis, 2019). In 1988, Archbald and Newmann expressed the need for student outcomes that were meaningful, significant, and reflective of the practices of successful adults (Cumming & Maxwell, 1999; Frey et al., 2012). The next year, Wiggins (1989) defined the concept of authentic assessment in which students engage in real-world tasks using collaboration and higher-level thinking. Proponents argue that authentic or performance assessments address the shortcomings of traditional standardized testing and when used properly can promote the development of desperately needed higher-order thinking skills and the transfer of knowledge in students (Biemer, 1993; Darling-Hammond & Adamson, 2010; Darling-Hammond et al., 2014). As a result, states and school divisions have implemented a variety of performance assessment reforms to improve teaching and learning.

Background on Assessment Policy

In the late 1980s and early 1990s, several US states developed statewide performance assessment programs that required students to go beyond factual knowledge to show depth of

mastery and ability to transfer skills to new situations (Conley, 2015). Maryland and Kentucky were the most comprehensive, covering a variety of subjects to include math, science, reading, writing, and social sciences. The Kentucky Instructional Results Information System (KIRIS) evaluated students in Grades 4, 8, and 11 through performance tasks in seven academic areas plus math and writing portfolios of selected student work, while the Maryland School Performance Assessment Program (MSPAP) assessed Grades 3, 5, and 8 through responses to complex, multistage tasks (Stecher, 2010). Covering fewer subjects, the California Learning Assessment System, created in 1991, used performance tasks with Grades 4, 8, and 10 to assess reading, writing, and mathematics, and the Washington Assessment of Student Learning (WASL) added science to the list of subjects measured through a combination of multiple-choice, constructed-response, essay, and problem-solving tasks in Grades 3-8 and 10. The Vermont Portfolio Assessment Program took a different approach, collecting portfolios of writing and math problem-solving work as chosen by teachers and students (Stecher, 2010). While California Learning Assessment System was only administered until 1994, MSPAP, KIRIS, and the Vermont Portfolio Assessment persisted through much of the 1990s and were the subjects of several studies. While these programs drew attention for the ability of assessment programs to shape instructional reform, the experiences in these early programs raised challenges concerning the validity of the tests, the reliability of the scores, and the time and monetary costs of administering the tests (Goldberg & Roswell, 2001; Honig & Alexander, 1996; Wren & Gareis, 2019).

No Child Left Behind

The passage of No Child Left Behind (NCLB) in 2001 and its emphasis on reporting individual student data for Adequate Yearly Progress goals brought changes to state policies and

the decline in such innovative assessment programs as seen in Maryland, Kentucky, and Vermont in the 1990s. Due to the burden of scoring, political challenges, and concerns about the validity of student scores, MSPAP, KIRIS, the California Learning Assessment System, and the Vermont Portfolio Assessment were all replaced by assessments that consisted of a mix of multiple-choice and constructed-response items by the early 2000s (Stecher, 2010). Similarly, the Minnesota Profile of Learning, which allowed locally developed performance assessments for subjects not represented on the state's basic skills tests, was also repealed in the early 2000s (Conley, 2014). To meet NCLB accountability standards, many states transitioned to primarily multiple-choice tests that were less costly and easier to administer and score (Wren & Gareis, 2019).

As assessment systems developed to meet the demands of NCLB, criticisms of the high-stakes, usually multiple-choice assessments implemented under NCLB increased throughout the early 2000s. Critics of the state-mandated tests argued that the tests did not align with standards, as the tests only focus on low-level, fact-based questions and did not test critical thinking skills (DeWitt et al., 2013; van Hover et al., 2010). As a result of the nature of the tests and the pressure to achieve high pass rates, teachers felt pressured to teach to the test, limiting how and what they taught to tested content rather than skills and deeper learning (Stotsky, 2016; van Hover et al., 2010; van Hover et al., 2011). Researchers comparing teachers in courses with state-mandated tests to courses without tests found teachers in courses without the tests had greater freedom in what and how they taught, which resulted in more critical thinking (Gerwin & Visone, 2006; Hong & Hamot, 2015). Existing research argues that state-mandated tests tend to cause teachers to use instructional strategies that contradict their ideas of best practice and limit how and what they teach since they feel pressured to teach to the test, resulting in increased

stress and lower morale (Abrams et al., 2003; Stotsky, 2016; van Hover et al., 2010; van Hover et al., 2011). Such research implies that teachers would implement different learning activities, have greater freedom in what and how they teach, and promote student critical thinking if they were not accountable to state-mandated assessments (Gerwin & Visone, 2006; Hong & Hamot, 2015).

As concerns about the effects of high-stakes multiple-choice testing on teaching and learning increased by the mid-2000s, states began incorporating performance skills alongside multiple-choice assessments for state accountability, similar to the still extant WASL assessments. New Hampshire, Rhode Island, and Vermont collaborated in 2005 and were joined by Maine in 2009 to create the New England Common Assessments Program which assessed reading, writing, math, and science through multiple-choice and constructed-response items including an essay and science inquiry task (Pecheone & Kahl, 2014). Connecticut created a similar, but independent, assessment, the Connecticut Academic Performance Test (Pecheone & Kahl, 2014). All of these focused on reading, writing, math, and science, while WASL provided state performance assessment models for district-determined tests in social studies, health, and arts.

The long-standing New York Regents Exams also consisted of a mix of multiple-choice and constructed-response items but were more comprehensive in the scope of disciplines that are assessed to include social studies and world languages in addition to English, math, and a variety of sciences (Pecheone & Kahl, 2014). The New York Regents Exams were the only ones that remain in use, as the others were replaced in the mid-2010s. The challenges for states implementing performance assessment reforms in the 2000s were similar to the those faced by MSPAP, KIRIS and others in the late 1990s: how to develop quality, valid, and reliable

performance tasks; growing educator capacity to incorporate performance tasks into instruction; and the feasibility of reliably scoring student responses across the state (Marion & Leather, 2015; Pecheone & Kahl, 2014; Stosich et al., 2018).

From NCLB to the Every Student Succeeds Act

The passage of the Every Student Succeeds Act (ESSA) in 2015 provided policy support to change assessment systems and a renewed interest in performance assessments (Foote, 2005; Stosich et al., 2018). In 2009, President Obama summarized the changing mood when he said,

> I'm calling on our nation's governors and state education chiefs to develop standards and assessments that don't simply measure whether students can fill in a bubble on a test, but whether they possess 21st century skills like problem-solving and critical thinking and entrepreneurship and creativity. (Obama, 2009, para. 20)

The subsequent passage of ESSA in 2015 gave states greater autonomy in defining academic standards and programs including allowing states to use multiple interim assessments that can include performance assessments instead of end-of-course multiple choice assessments for federal accountability (Stosich et al., 2018).

Focus on College and Career Readiness

ESSA also required states to submit college and career readiness plans to the U.S. Department of Education and encouraged the integration of academic and technical curriculum in schools, furthering the need for performance assessments (Hackmann et al., 2019). Although the U.S. Department of Education (2020) stresses the need for students to be able to compete in the "global, knowledge-based economy," there is no single, shared definition of college and career readiness, as states describe and operationalize the concept differently. Some states stress

the ability to succeed in post-secondary education settings, others include being career-ready, and still others include language about being engaged citizens (Hackmann et al., 2019).

To measure college and career readiness, 27 states use dual enrollment in any course, 22 use the ACT, 24 use taking Advanced Placement (AP) exams, 23 use taking an International Baccalaureate (IB) exam, 21 use a Career and Technical Education (CTE) sequences, 21 use an Industry-recognized credential, 15 use the SAT, 10 use the Armed Services Vocational Aptitude Battery (ASVAB), 10 a work-based learning experience, and there are even some less frequently used measures (Hackmann et al., 2019). The Virginia Department of Education (VDOE), to ensure students are college and career ready, outlines in the 2016 *Profile of a Virginia Graduate* the "knowledge, skills, experiences and attributes that students must attain to be successful in college and/or the work force and to be 'life ready'" (VDOE, 2022a). The *Profile of a Virginia Graduate* requires the teaching of what the VDOE has defined as the 5 C's: creativity, collaboration, critical thinking, communication, and citizenship. Virginia's 5 C's, like many other state college and career readiness measures, are skills not easily assessed by multiple-choice tests and better measured by performance assessments (VDOE, 2022a). Underlying the different state definitions is a common emphasis on skills over content. The result of the increasing emphasis on college and career readiness is a demand to assess problem-solving, critical-thinking, and other higher-order skills which has contributed to the increased focus on performance assessments in K-12 education.

Developing Performance Assessments

To meet the increased demand for quality performance assessments, states joined consortiums to share resources and expertise in developing quality assessments that include constructed response or performance tasks. Washington, Connecticut, and California have joined

with 11 other states plus the US Virgin Islands and Bureau of Indian Education in the Smarter Balanced Assessment Consortium, which develops computer-adaptive tasks requiring research, writing, and problem-solving that students take as part of state accountability measures as well as for formative use in the classroom (University of California Regents, 2018). Twelve states, including Virginia, have joined the Innovation Lab Network to develop systems of assessments that include performance assessments. A study of the policies of Innovation Lab Network members found that 11 of the 12 use performance assessments in classroom instruction, seven use performance assessments for state accountability, three require performance assessments for high school graduation, and one, New Hampshire, uses performance assessments for federal accountability (Stosich et al., 2018).

At the state level, New Hampshire is currently piloting the Performance Assessment of Competency Education, which uses a system of local and common performance assessments that enable students to demonstrate competency on multiple occasions and in different contexts (Marion & Leather, 2015). The Performance Assessment of Competency Education program is an opt-in program for school divisions and is currently not attempting to meet designated levels of standardization and psychometric specifications, but instead seeks to promote deeper student learning by removing current accountability measures that hinder college and career readiness (Marion & Leather, 2015). The state has created the New Hampshire Test Bank of reviewed, high-quality performance assessments shared across school divisions in the Performance Assessment of Competency Education pilot. Since the passage of ESSA and the greater flexibility for accountability, states have been increasing the implementation of quality performance assessments in K-12 education.

Rationale for Performance Assessments

Performance assessments require students to construct answers or produce products in a context that emulates the conditions of real life and require students to apply knowledge and reasoning (Darling-Hammond, 2017; Darling-Hammond & Adamson, 2010; Stecher, 2010). Performance tasks range from on-demand tasks that are shorter in time, focused on a more limited range of learning outcomes, and provide the student with less choice, such as constructed-response or stand-alone tasks, to extended, long-term performance tasks that cover a greater number of skills, are more integrated into instruction, and give students more control over the assignment such as curriculum-embedded tasks or complex projects (Brookhart, 2015; Khattri et al., 1998; Wren & Gareis, 2019). These categories depict how the performance tasks that make up performance assessments vary depending on the number of intended learning outcomes they are designed to measure, the degree to which the teacher is involved with the student during the task, the prescriptiveness of the response, and the duration of student engagement in the task (Wren & Gareis, 2019). Within these categories, performance assessments can take a variety of forms from writing essays, delivering speeches, and researching projects, to computer simulations and lab investigations (Darling-Hammond, 2017; Reed, 1993; Stecher, 2010).

The emphasis on 21st century skills, as expressed by ESSA and educational reformers and operationalized in Virginia's 5 C's, has led to a demand for better assessments that more effectively reflect student abilities to think critically, problem solve, communicate, and transfer knowledge to new settings (Darling-Hammond & Adamson, 2010; Darling-Hammond et al., 2014). The Hewlett Foundation (2013) characterizes these higher-order learning skills as "deeper learning" which the Foundation defines as "skills and knowledge that students must possess to

succeed in 21st Century jobs and civic life" (p. 1). This includes the six competencies of mastering academic content: thinking critically and solving complex problems, working collaboratively, communicating effectively, learning how to learn, and developing academic mindsets.

The open-ended and real-world nature of performance assessments that allow students to evaluate sources, construct and defend an argument, or design or perform a task are seen as a better means for students to display higher-order understanding and deeper learning than standardized, multiple-choice tests (Baron, 1996; Darling-Hammond & Adamson, 2014; Foote, 2005). Thus, by allowing students to construct responses, performance assessments are seen as better indicators of student learning and reasoning skills than assessments in which students identify a right answer or respond with memorized information.

Proponents of performance assessments claim that performance tasks not only better promote and measure deeper learning, but also "transform" traditional assessment and instruction (Van Duinen, 2006, p. 142). Scholars argue that teaching students to be prepared for the demands of performance assessments helps teachers to develop more effective instructional techniques and leads to a shift in classrooms from lower-level content-based teaching to more time spent problem-solving, writing, and developing transferable skills (Darling-Hammond & Adamson, 2010; Khattri et al., 1995; Parke & Lane, 2008). The embedded use of performance tasks throughout the school year allows teachers to assess students' abilities in an ongoing manner as part of instruction and to further student learning by providing students with specific feedback to support deeper learning (Darling-Hammond & Adamson, 2010; Darling-Hammond & Ancess, 1996; Khattri et al., 1995; Stosich et al., 2018). Thus, the implementation of an

effective performance assessment system can reform schools and classroom organization as well as student outcomes (Darling-Hammond & Ancess, 1996).

Preparing Teachers for Performance Assessments

The shift to quality performance assessments requires quality and sufficient professional development (O'Brien, 1997; Stosich et al., 2018). The broad definition of performance assessments encompassing a wide array of student activities makes developing shared meanings and understandings of performance assessments challenging in practice. Identifying the constructs to be measured based on state and national standards and then creating functional understandings of what constitutes acceptable demonstration of conceptual understanding or effective communication is overwhelming to teachers who lack experience and training in this work (Darling-Hammond & Ancess, 1996; Goldberg & Roswell, 2000; Khattri et al., 1995). The struggle teachers face in constructing meaningful performance assessments is not just due to the difficulty of understanding how to construct the test but also having the time and resources to devise and field test complex performance tasks (Baron, 1996; Goldberg & Roswell, 2000). As a result of the lack of time and understanding, many teacher-created assessments end up being hands-on activities that lack the depth of true performance tasks and often are an assortment of activities that have been inserted into existing instruction (Firestone et al., 1998; Goldberg & Roswell, 2000; Gong & Reichy, 1996; Messick, 1994).

To address these challenges and best enable the positive learning outcomes of performance assessments, researchers argue that states and school districts need to clearly delineate their purpose for and definition of performance assessments and the feasibility of authenticity of tasks when implementing performance-based initiatives to ensure that the desired goals and student outcomes are achieved (Messick, 1994). Then schools and districts need to

provide teachers with training, models of quality performance assessments, and time to plan and collaborate in order for performance assessments to bring about the instructional reform advocated by researchers and proponents of these assessments (Goldberg & Roswell, 2000; Khattri et al., 1995).

Balanced Assessment Systems

Performance assessments have a purpose and can measure particular outcomes, but they should be one of a system of assessments (Haney & Madaus, 1989). No single type of assessment can provide all the data that students and schools need to improve teaching and learning; therefore, balanced assessment systems that use a variety of assessment measures in a purposeful approach are needed (Chappuis et al., 2017; Wren & Gareis, 2019). Even the strongest proponents of performance assessments, such as Darling-Hammond, see them as a system of ongoing assessment and learning on the part of both teacher and student to best, and most equitably, capture and represent the extent of student learning (Conley, 2015; Reed, 1993). Policy makers, administrators, and teachers need to be clear as to the goal of each assessment and use the appropriate assessments that measure the desired outcomes, recognizing the limitations of any type of assessment. Then they can plan out a system of assessments, including performance assessments, that best provide evidence of student learning.

Assessment Policy in Virginia

To address the concerns created by the heavy use of multiple-choice assessments and to meet the college and career readiness needs of the *Profile of a Virginia Graduate*, the Virginia Department of Education (VDOE) is changing educational policy to incorporate and encourage greater usage of performance assessments. Since 1998, public school students in Virginia have taken state-mandated end-of-course standardized tests in the core subjects of Math, English,

Social Studies, and Science known as Standards of Learning (SOL) tests in Grades 3, 5, 8, and high school. These tests later increased to include other elementary and middle school grades. Designed to measure the extent to which students had mastered the learning objectives outlined by the VDOE in the SOLs, these tests originally consisted solely of multiple-choice questions with the exception of the Writing test that required a writing sample. Starting in 2011, "technology-enhanced items" were introduced to the English and Science SOL tests and subsequently expanded to include the Math and Social Studies tests (VDOE, 2023a). The technology-enhanced items were not multiple choice but were still within the same family of select-response items, and were purported to require students to engage in more complex tasks to demonstrate student critical thinking and problem-solving skills, such as through drag-and-drop and multiple-answer response formats (VDOE, 2023a).

The shift to performance assessments in Virginia began in 2014 when the Virginia General Assembly passed of House Bill 930 and Senate Bill 306 removing five state end-of-course SOL assessments including Grade 3 science, Grade 5 writing, Grade 3 social studies, US History to 1865 (USI), and US History 1865 to the Present (USII), both usually taught in late elementary or middle school (VDOE, 2014). When the Washington state legislature initiated a similar reform in 1992, they created a commission designated to create the performance assessments schools would use and established an accountability system to monitor each school's progress in a top-down mandate (Stecher et al., 2000). The VDOE chose a more grassroots implementation approach putting decisions about the assessments in the hands of local school divisions as the eliminated SOL tests were to be replaced with locally developed alternative assessments (LAAs) that provided evidence of proficiency in content and skills identified in the SOLs. The VDOE Guidelines for Local Alternative Assessments specifically

stated that “the development and/or selection of the local assessments are left to the discretion of the school division” (VDOE, 2014, p. 3) Given the goal of demonstrating skills, the VDOE recommended that the LAAs include some performance assessments to promote increased usage of performance assessments and deeper learning (Abbott, 2016; Stosich et al., 2018; VDOE, 2014). In 2019, the VDOE required school divisions to prepare Balanced Assessment Plans for each of the five courses in which SOL tests had been removed (VDOE, 2019c). The detailed assessment plans for each course would indicate the types of assessments being used to demonstrate student mastery of each of the standards in the course, and the plan must include some performance assessments (VDOE, 2019c). While the evolution of the policy has led to greater guidelines from the VDOE, divisions still have flexibility in determining the set of assessments within their balanced assessment plan, the ability to develop their own performance assessments that best fit their local context, and the responsibility to score and set accountability standards on the assessment in this grassroots approach to implementation.

Both of these trends, the replacement of SOL tests with LAAs and the focus on the 5 C’s, required changes in both teaching and assessment, most immediately in the requirement for Virginia school districts to create locally developed, performance-based assessments that would award students credit for meeting state standards. The VDOE has given divisions considerable flexibility in how divisions develop and select LAAs, including the number and type of assessments as well as whether the division will use division-wide assessments or school-based assessments (VDOE, 2014). Given the autonomy granted by Virginia’s grassroots approach to policy implementation, in this study I explored the types, role, and quality of performance assessments developed by local divisions to meet the VDOE mandate in order to investigate the

efficacy of a grass-roots approach to developing LAAs as performance-based accountability measures.

Conceptual Framework

With an increased emphasis on assessments promoting higher quality learning and 21st Century skills, school divisions and teachers will have to alter instructional and assessment strategies to prepare students for the demands of performance assessments (Abbott, 2016). To guide teachers and divisions in the development of performance assessments, the VDOE created the Virginia Quality Criteria Tool (VQCT) for Performance Assessments to "support comparability in rigor and quality across the state" (VDOE, 2019d, p. 1). Because this study will focus on the LAAs developed by Virginia school divisions to meet state accountability, this study will use the conceptual framework for quality performance assessments as specified in the VQCT (Appendix A).

The Virginia Quality Criteria Tool

The VQCT consists of seven criteria: Standards/Intended Learning Outcomes, Authenticity, Language Use for Expressing Reasoning, Success Criteria for Students, Student Directions Prompt and Resources/Materials, Accessibility, and Feasibility. Each criterion has from one to three subcategories that define the expectations for a total of 17 sub-criteria that are each rated on a scale from 0-3, with 0 indicating no evidence of the criterion, 1 indicating limited, 2 indicating partial evidence, and 3 indicating full evidence (VDOE, 2019d).

Criterion 1: Standards/Intended Learning Outcomes

Criterion 1, Standards/Intended Learning Outcomes, focuses on the alignment of the assessment with the standards, both content and skills, to be measured and the need for the assessments to require higher order thinking and/or 21st Century skills. Criterion 1A requires the

Virginia SOLs measured in the assessment to be identified in a task template to ensure that the standards being measured are appropriate for the grade-level and that the assessment is anchored in the standards, which is critical to quality assessments (Brookhart, 2015; Gareis & Grant, 2015; Shiel, 2017; Wren & Gareis, 2019). The criterion then evaluates the alignment of all elements of the assessment to those stated SOLs and the developmental level of the students, including the task, resources used to complete the task, and the products required of students (VDOE, 2019d). Quality performance assessments, and the resources students use to complete them, should be complex enough to assess student abilities to use a variety of skills and knowledge to complete the task, while being developmentally feasible and appropriate, which is the focus of Criterion 1A (Wiggins, 1998; Wren & Gareis, 2019).

Criterion 1B evaluates whether the assessment goes beyond simple recall to require complex thinking such as applying concepts or using skills (VDOE, 2019d). Performance tasks, such as evaluating sources, constructing and defending an argument, or designing or performing a task, are seen as a better means for students to display higher-order understandings and deeper learning than standardized, multiple-choice tests; thus, Criterion 1B evaluates whether that goal is being met (Baron, 1996; Darling-Hammond & Adamson, 2014; Foote, 2005). Criterion 1C evaluates whether the assessment, beyond measuring specific intended learning outcomes, allows students to demonstrate deeper learning skills such as critical thinking, working collaboratively, problem solving, and effective communication (Hewlett Foundation, 2013; VDOE, 2019d). The criterion also measures opportunities for the student to demonstrate "Life-Ready competencies" such as workplace skills, civic responsibility, use of technology and cross-disciplinary connections, thus measuring the assessments' ability to promote 21st Century learning skills that transcend the classroom and prepare the student for success beyond school (Darling-Hammond

& Adamson, 2014: VDOE, 2019d; Wren & Gareis, 2019). In sum, Criterion 1 examines the degree to which the assessment requires students to demonstrate and utilize skills, knowledge, and deeper learning that are representative of a cogent set of intended learning outcomes articulated in the state standards and in broader educational aims of the 5 C's.

Criterion 2: Authenticity

Criterion 2, Authenticity, evaluates performance assessments on their authenticity to both the purpose and audience of the task, as well as to the discipline. While many researchers and educators use authentic and performance assessments interchangeably, authenticity focuses on a real-world context, making learning relevant beyond the classroom. Wiggins (1989) defined authentic assessment as tasks that replicate real-world challenges and performances of professional adults that require the posing of questions, solving of problems, and explanation of responses. The VDOE defines authentic as tasks that mirror real-life situations and/or are authentic to the academic discipline (VDOE, 2019c). Since the 5 C's emphasize college and career readiness, the first part of Criterion 2 evaluates whether the topic, resources, purpose, and products in the tasks are relevant to the real-world, and the second part evaluates whether the actual work students are required to do in the task corresponds to what practitioners of the discipline do (VDOE, 2019d). Quality performance tasks should engage students in meaningful tasks that resemble professional practice and prepare students for the real-world, which is measured by Criterion 2 (Darling-Hammond, 2014; Gulikers et al., 2004; Wiggins, 1998; Wren & Gareis, 2019).

Criterion 3: Language Use for Expressing Reasoning

Criterion 3, Language Use for Expressing Reasoning, focuses on student language use and ability to communicate. Criterion 3A measures the extent to which the assessment promotes

language development by having students communicate their reasoning through academic language. Criterion 3B states that the assessment "should" require students to use or at least practice using different forms of language or language media in their expression, such as text, video, audio, or oral (VDOE, 2019d). The emphasis on developing academic language, effective communication and being able to communicate through a variety of mediums corresponds with Virginia's emphasis on the 5 C's, including communication, and the *Profile of a Virginia Graduate* as well as the U.S. Department of Education's (2020) emphasis on preparing students to be life-ready in a global, knowledge-based economy (VDOE, 2022a).

Criterion 4: Success Criteria for Students

Criterion 4, Success Criteria for Students, emphasizes the need for a clear rubric or success criteria for students. Criterion 4A evaluates the existence of a clear scoring tool or rubric that aligns to the intended learning outcomes of the assessment. Criterion 4B measures the extent to which the scoring tool clearly describes expectations for students, allowing the rubric to provide feedback to students to improve their work or performance. Criterion 4C states that the rubric or feedback should be used repeatedly throughout the course. Besides being a way to communicate student proficiency, the goal of the repeated use, and thus familiarity with, the rubric is to improve both teaching and learning by providing students with consistent expectations and demonstrating growth over time as well as informing instructional decisions (VDOE, 2019d; Wiggins, 1998). Establishing clear success criteria prior to the assessment forces teachers or test developers to clarify the specific and desirable levels of performance and thus structure instruction to allow students to achieve those levels; it also enables the student to know they have achieved the learning goals (Shiel, 2017; Wiggins, 1998). Thoughtfully designed

rubrics integrated into instruction promote quality performance assessments that enhance both teaching and learning.

Criterion 5: Student Directions Prompt and Resources/Materials

Criterion 5, Student Directions Prompt and Resources/Materials, focuses on the directions and materials provided to students in the LAA, commonly referred to as *student-facing materials*. Criterion 5A ensures that the prompt, directions, and any additional materials or resources the student must use are aligned to the intended learning outcomes and the purpose of the task (VDOE, 2019d). Although activities might be fun or engaging, that does not mean the tasks or resources provided connect to the intellectual goals (Wiggins, 1998). Thus, quality performance assessments must align with the intended purpose, so that they can also provide meaningful feedback to improve student performance and pedagogical practices (Khattri et al., 1998).

Criterion 5B evaluates the clarity and appropriateness of the language in the prompt, directions, and resources as well as organization of the assessment (VDOE, 2019d). Unclear instructions, overly difficult to decipher materials, or poorly organized tasks inhibit students from being able to demonstrate the intended learning outcomes, as students struggle with making sense of the instructions. Finally, criterion 5C measures the cultural sensitivity and bias of the prompt and resources to ensure that students are not prevented from demonstrating proficiency in the intended learning outcomes due to inherent biases of the assessment (VDOE, 2019d). Criterion 5 is designed to prevent poorly written or constructed tasks from preventing students from being able to demonstrate what they have learned, thus preventing inaccurate depictions of student levels of proficiency that could lead to flawed decisions about how to structure instruction.

Criterion 6: Accessibility

Criterion 6, Accessibility, evaluates the accommodation or differentiation of the LAA to ensure all students are able to engage with it. Criterion 6A measures the extent to which performance assessments accommodate the participation of all students through the provisions for accommodations or other supports for students of varying abilities. Criterion 6B evaluates the degree to which the LAA allows differentiation based on the principles of Universal Design for Learning, allowing students multiple means of engagement, representation, and expression fitting student abilities and needs (CAST, 2018). Similar to Criterion 5, the emphasis in Criterion 6 is to ensure that performance assessments measure the intended learning outcomes they were designed to measure, and that student performance on an assessment is not actually due to other factors, namely differing student needs.

Criterion 7: Feasibility

Finally, Criterion 7, Feasibility, emphasizes the feasibility of students completing the task assigned in the LAA. Criterion 7A addresses the existence of student-facing materials, including prompts, resources and scoring tools, and teacher ability to access the resources required for the task. The lack of any of these student-facing materials would prevent consistency in implementation of the assessment, and that variation could result in the assessment no longer aligning to the intended learning outcomes. Criterion 7B notes the existence of an indication of the duration of the assessment and whether that time frame is realistic for the complexity and scope of the task. Criterion 7C applies to assessments that are implemented over multiple lessons, requiring a schedule for how the assessment is implemented and fits withing the learning sequences, as well as how the assessment fits with student prior knowledge (VDOE, 2019d). Variations in timing or scheduling could affect student ability to adequately demonstrate skills or

knowledge, thus minimizing the usefulness of the performance assessments for teaching or learning.

Problem Statement

The Virginia General Assembly's passage of House Bill 895 and Senate Bill 336 has removed five state end-of-course assessments, including three social studies SOL assessments in Grade 3, USI, and USII, to be replaced with local alternative assessments with the intent to expand this policy to other disciplines and grade levels (VDOE, 2014). The subsequent passage of ESSA has given states like Virginia greater autonomy in defining academic standards, including allowing states to not rely on end-of-course multiple-choice tests, but to instead use multiple interim assessments throughout the course, which can include performance assessments (Stosich et al., 2018). In addition, the new *Profile of a Virginia Graduate* requires the teaching of the 5 C's: creativity, collaboration, critical thinking, communication, and citizenship.

In practice, Virginia has decided on a grassroots policy approach, allowing school divisions to develop locally based alternative assessments to replace the SOL tests in five courses (VDOE, 2014). While Virginia provides quality criteria for performance-based assessments, there is no standardized procedure guiding how the local alternative assessments are to be developed and implemented. As a result of this autonomy, school divisions have the freedom to develop assessments that correspond to the unique needs and settings of each division. My research is a descriptive study of how school divisions have responded to the mandate in terms of how the performance assessments for the LAAs were developed, the role of those performance assessments in the balanced assessment plan, and how well those assessments meet the VDOE definition of quality when given this level of autonomy from the state. By focusing the study on

performance assessments for USI and USII, I tried to control some issues of validity and reliability in the research process (as described in Chapter 3).

Insight into the development process and the types and qualities of local alternative assessments in Virginia has both local and national implications. Locally in the state of Virginia, the Code of Virginia states that the local alternative assessments should "ensure that students are making adequate academic progress" (VDOE, 2019c, p. 1), and the VDOE seeks comparable "rigor and quality across the state" for performance assessments (VDOE, 2019d, p. 1). A study of the authenticity and quality of the local assessments may reveal the extent to which these policy objectives are being met. Nationally, as other states are similarly allowing local school divisions autonomy to develop assessments, insights into the types and quality of locally developed assessments being developed in Virginia can provide insights and direction for state help to increase teacher and school division capacity to develop quality performance assessments.

Research Questions

In this study of locally developed performance assessments, I explored the developmental process used by school divisions to construct the performance assessments for LAAs, the role of the performance assessments in the larger balanced assessment plan, and the extent to which local assessments meet the quality criteria of performance assessments. To that end, I investigated the answers to the following three research questions.

Research Question 1. How have Virginia school divisions approached the process of developing local alternative assessments to replace the removed SOL tests in US I and US II?

Sub-question A: Who is responsible for creating the local alternative assessments in each division, and are the assessments division-wide or school-specific (e.g., teachers, administrators,

a combined group of teachers and administrators, or obtained from an outside publisher or consultant)?

Sub-question B: What steps were taken to prepare for the process of creating the local alternative assessments (e.g., regional workshop by the state, VDOE sponsored workshops, workshops by a professional association such as the School-University Resource Network (SURN), Virginia Association of School Superintendents (VASS), consultant or invited presenter to the division, internally-led PD)?

Sub-question C: What processes were utilized in the design and development of the assessments to ensure quality performance assessments? (e.g., use of templates, comparison to the VQCT, piloting of the assessment, review of student work samples, use of tables of specification, external expert review, or other strategies as indicated by study participants)?

Research Question 2. How many and what type/s of assessments have divisions selected to constitute their locally developed alternative assessments for USI and USII, and what is the role of performance assessments in their respective plans?

Research Question 3. To what extent do the locally developed performance assessments developed by individual school divisions in Virginia for USI and USII meet the seven quality criteria and 17 distinct sub-criteria established by the VDOE?

Significance of the Study

At the state level, the Virginia Assembly legislated that LAAs were to be one measure to ensure that students were making "adequate academic progress in the subject area and that the Standards of Learning content is being taught" (VDOE, 2019c, p. 2). Teachers, schools, and divisions will use LAAs to make inferences about student learning, to measure whether the SOL content is being taught, and to make modifications to curriculum and instruction. Given the

important inferences drawn from the assessments, I measured the quality of the assessments and thus the degree to which divisions can use the assessments to adequately draw conclusions about student learning and progress on the SOLs. In addition, since the use of LAAs is one method by which the Commonwealth of Virginia is seeking to promote deeper learning and college and career readiness skills, it is important to examine whether the LAAs being created under this mandate are, indeed, of sufficient quality.

Nationally, states, such those in the Innovation Lab Network, are currently implementing performance assessments. Of the 12 Innovation Lab Network members, 11 currently use performance assessments for classroom use; three use them for graduation requirements, seven for school accountability, and one, New Hampshire, for federal accountability. Thus, this study may provide data from one state on what existing locally developed performance assessments for state accountability can look like and whether teacher-made assessments meet the criteria for quality performance assessments. The findings of this study may reveal directions for better teacher training and preparation to develop quality performance assessments and the development of policy to more successfully implement performance assessments to improve the quality of teaching and learning.

Definitions of Terms

Since school divisions in Virginia are being trained on and mandated to meet the Virginia Department of Education's definition of performance assessments, the study will use definitions that correspond with the criterion to which schools are being held accountable.

Authentic Assessment: an assessment that includes tasks with a real-world context and/or are authentic to the academic discipline (VDOE, 2019a).

Deeper Learning: A set of "skills and knowledge that students must possess to succeed in 21st Century jobs and civic life" including the six competencies of mastering core academic content, thinking critically and solving complex problems, working collaboratively, communicating effectively, learning how to learn, and developing academic mindsets (Hewlett Foundation, 2013, p. 1).

Local Alternative Assessment (LAA): an assessment created by individual school divisions in Virginia to be administered in place of eliminated SOL tests in five subject areas (VDOE, 2019a).

Performance Assessment: an assessment that requires students to perform a task or create a product and is scored using a rubric or set of criteria (VDOE, 2019a). While experts distinguish between performance tasks and performance assessments, these two terms are often used interchangeably (Wren & Gareis, 2019). In the Guidelines for Local Alternative Assessments under the heading of definition the VDOE uses both terms writing, "performance assessments generally require students to perform a task or create a product" and "it is up to the local school division to determine whether a performance task is authentic" (VDOE, 2021a, p.2). Interviewees often used the word task to describe their performance assessments or to focus the conversation on the actual task the student was engaged in and thus this paper will use the two terms, performance task and performance assessment, interchangeably.

Rubric: provides a set of criteria for measuring or evaluating student work on a performance assessment (VDOE, 2019a).

Standards of Learning (SOLs): The VDOE (2022b) has outlined the expectations for "what students should know and be able to do" in the required courses for students in Grades K-12 in

the SOLs, specifying the knowledge and skills that students should possess by the end of each course (para. 3).

.

CHAPTER 2

REVIEW OF RELATED LITERATURE

Educational researchers, educational policy makers, and the Every Student Succeeds Act (ESSA) promote the implementation of performance assessments as a means of promoting deeper learning for students, better measuring student learning, and strengthening classroom instruction. As the Virginia Department of Education (VDOE) encourages and requires the increased use of performance assessments for both classroom use and accountability, this chapter will examine the research behind performance assessments and their implementation. Since the locally developed alternative assessments in this study are for state accountability purposes, I begin this chapter with a summary of the empirical studies conducted on previous state, district, and local accountability systems that employed performance assessments and the resulting concerns with reliability and validity of performance assessments. I then examine the literature focused on the traits of quality performance assessments and the procedures for developing them, including a focus on the role of the teacher in this process. Because most of the research on performance assessments focuses on math and language arts, and because Virginia is beginning this process primarily in the social studies, I conclude this chapter with an examination of the research specifically on performance assessments in the social studies.

Empirical Studies

The rationale behind the use of performance assessments, as explained in Chapter 1, is to improve classroom instruction and promote deeper learning for students. Research studies during the late 1990s and early 2000s focused on states such as Maryland, Kentucky, and Vermont that

were implementing state accountability systems that used performance assessment with the goal of improving instruction (Koretz, Barron, et al. 1996; Koretz, Mitchell, et al. 1996; Stecher & Mitchell, 1995). Much of the empirical research surrounding performance assessments focused on either the effects of performance assessment reforms on classroom instruction and assessment or on the validity or reliability of performance assessments for measuring student outcomes.

Effect of Performance Assessments on Classroom Instruction

Survey Studies

Numerous studies on the effects of performance assessments on classroom instruction and learning were interviews or surveys in which teachers, principals, and/or students reported their perceptions of changes in the classroom due to performance assessment reforms.

Performance Assessments for State Accountability. Most of the empirical studies on performance assessments focus on states that have implemented state-wide performance assessment initiatives designed to improve instruction, such as Kentucky, Maryland, and Vermont (Koretz, Barron, et al., 1996; Koretz, Mitchell, et al., 1996; Stecher & Mitchell, 1995). Across all three states, researchers found teachers reporting positive changes to instruction resulting from the state performance-assessment program, but researchers were also concerned about teachers' lack of understanding of skills and the emphasis on tested skills in instruction to the neglect of other skills and content.

Stecher and Mitchell (1995) conducted written surveys and phone interviews of a representative sample of 20 fourth-grade teachers implementing the Vermont math portfolio program. The researchers found that teachers felt the portfolio assessments increased their knowledge of math problem-solving, and the teachers reported changing their instruction to increase time on problem-solving skills. While teachers reported the portfolios improving

instruction, the researchers felt teacher comments raised questions about how well teachers understood problem-solving and thus teacher ability to implement the reform. The teacher responses revealed a lack of a common understanding of problem-solving, that teachers used a variety of problem-solving skills in class, and, when given a math problem, a lack of agreement on the problem-solving demands of the specific task (Stecher & Mitchell, 1995). In addition, researchers argued that teachers chose classroom tasks based on how closely the task aligned with the portfolio rubric, rejecting useful problems that would not score high on the rubric, and teachers also pre-taught portfolio tasks (Stecher & Mitchell, 1995). The researchers concluded that, while instruction was changing, the nature and degree of change was questionable given the teachers' lack of understanding of problem-solving and limiting of instruction to only the skills designated on the rubric.

Koretz, Barron, et al. (1996) found similar results regarding Kentucky's KIRIS program in their mail and telephone surveys of 115 principals and 216 teachers. While teachers and principals largely agreed that KIRIS was a useful tool for positive instructional change leading to more writing and problem-solving, researchers found teachers used the term "problem-solving" to refer to widely varying activities. In addition, researchers reported that over 90% of the teachers reported deemphasizing or neglecting untested areas of the curriculum to focus on the tested elements, while 52% reported that the increased emphasis on writing had actually led to students being tired of writing.

A similar study in Maryland using mail and phone surveys of 112 principals and 226 teachers also reported perceived positive effects on instruction with more cooperative work, writing, problem-solving, and thinking skills with similar lack of agreement as to what constitutes problem-solving skills and an over-emphasis of instruction on tested areas and test

preparation (Koretz, Barron, et al., 1996). Researchers found in Maryland, as in Vermont and Kentucky, that "problem-solving skills" was a broad term that was used to describe a wide range of instructional practices and researchers call for more investigation into the actual instructional activities in the classrooms. With a large percentage of teachers crediting practice tests and test preparation materials for MSPAP score gains over broader improvements in student knowledge and skills, the researchers felt the data raised questions about the usefulness of the MSPAP for accountability measures concerning student growth.

Further surveys were conducted concerning the MSPAP in the early 2000s that supported the survey findings of Koretz, Mitchell, et al. from the mid-1990s. Lane et al. (2002) surveyed teachers, principals, and students in 115 classes across 59 elementary schools and 95 classes in 31 middle schools Maryland concerning their familiarity with and beliefs about MSPAP and the impact of MSPAP on instruction. The responses showed 94% of principals and 76% of teachers felt MSPAP was a useful tool to create positive change in instruction. Teachers reported increased emphasis on math problem-solving, reasoning, and communication as found in MSPAP, but the researchers found that emphasis differed between tested grades and non-tested grades. In elementary schools, 67% of tested grade teachers reported an increased emphasis on problem solving and 45% asked students to perform tasks similar to MSPAP tasks at least weekly, while non-tested grade teachers were 57% and 27% respectively on both questions. The gaps in middle school were greater with 51% of tested subject teachers reporting increased emphasis on problem-solving skills and 43% using MSPAP type tasks at least weekly compared to 26% and 14% for non-tested area teachers (Lane et al., 2002). Researchers noted that while 89% of teachers reported general improvement to daily instruction to help prepare for the

MSPAP, the focus was on format and content of the MSPAP rather than more general improvements in instruction.

In a separate Maryland study, Stone and Lane (2003) used questionnaires to study the relationship between increased MSPAP scores and a variety of factors including classroom instruction, student and teacher beliefs about MSPAP, and school characteristics such as free and reduced lunch. The researchers conducted separate surveys for social studies, science and math, and language arts. Teachers and students in 86 elementary and middle schools were surveyed about math and language arts, while 111 elementary and middle schools were surveyed about social studies and 116 schools about science. Researchers found a positive relationship between teacher reports of instruction more strongly reflecting MSPAP problem types and improved MSPAP scores. The effects were greater for reading and writing scores than other content areas, and scores in science and social studies had declined in the last year of the study. The researchers found that student reports of increased use of MSPAP-like tasks in science and social studies were actually negatively related to student performance on MSPAP tasks in these subjects. The researchers concluded that the lack of MSPAP gains in science and social studies may be due to increased emphasis on tasks that resemble the assessment, teaching to the MSPAP, rather than a classroom focus on a variety of process-learning outcomes and instructional strategies (Stone & Lane, 2003).

In a later study focused on just science and social studies, Parke and Lane (2007) surveyed 19,000 fifth- and eighth-grade students across 160 schools in Maryland using both Likert-type questions and constructed-response items; they also found increased emphasis on MSPAP-like tasks in those classes. In this student-focused study, students reported that while they preferred multiple-choice assessments and found them more interesting than MSPAP items,

they believed the MSPAP was cognitively more demanding, using phrases like "think harder" and saying the MSPAP better demonstrated what they had learned. Fifty-five percent of fifth-graders and 42% of eighth-graders reported using MSPAP-type activities throughout the year, and 21% reported doing them weekly, suggesting greater integration of MSPAP skills in instruction. Teachers reported higher alignment between instruction and MSPAP skills than students reported, which Parke and Lane attributed to teachers not making their goals clear to students and student inability to distinguish the skills required by particular classroom tasks.

The researchers in these studies did not review samples of classroom activities to verify the accuracy of the teacher responses. As stated in several of the studies, teachers and administrators had diverse understandings and definitions of performance assessments and constructs such as higher-order thinking, communication, or problem-solving, which could affect their perception of changes in instruction. Also, students could lack metacognitive awareness of the tasks they are performing. These different understandings may have resulted in accurate perceptions and characterizations of classroom instruction and activities. Furthermore, the new behaviors of increased writing, group work, or MSPAP/KIRIS-type tasks did not inherently indicate that students were engaging in critical thinking and real-world skills. Thus, the self-reported changes may or may not have reflected actual changes in instructional practices, but the methodology of the studies prevented any verification of teacher responses.

Performance Assessments as a School-Based Reform. While most studies on the effects of performance assessments on classroom instruction and assessment have been focused on large, state-wide programs, Sivalingam-Nethi (1997) examined a school-based reform independent of state accountability measures. The researcher sought to determine whether the new science program increased students' positive attitudes toward the subject and increased

instructional emphasis on understanding, problem-solving, and higher-order thinking skills. Using a cluster sample of 20 classes, the researcher provided questionnaires to 366 students and interviewed 13 teachers. Similar to the research by Stecher and Mitchell (1995), Koretz, Barron, et al. (1996), and Koretz, Mitchell, et al. (1996), Sivalingam-Nethi found that teachers differed in their conceptual definitions of student understanding, problem-solving, and higher-order thinking. Also similar to the studies on state-based performance assessments, this study found that the new science program had a limited effect on instruction. While more activities focused on understanding, problem-solving, and higher-order thinking, the researcher noted that those activities were small in number compared to the extent of teacher-directed instruction still going on. While the previous studies that surveyed students had found fairly positive student responses, Sivalingam-Nethi concluded that the use of performance assessments did not change student attitudes toward science.

Summary of Survey Studies. All seven of these survey studies found that the use of performance assessments in state or local reform movements led to instructional changes regarded as positive by teachers and principals. The studies from Maryland, Kentucky, and Vermont also reported greater use of tasks similar to those on the MSPAP or KIRIS. Thus, the studies show that the use of performance assessment reforms promoted instructional change to increase higher-order thinking, problem-solving, and deeper learning for students. The studies revealed that potential drawbacks to this emphasis on assessment-type tasks may be more focused on test preparation and practice rather than promoting a variety of process-learning outcomes and the possible exclusion of non-tested curriculum or problem types (Koretz, Barron et al., 1996; Koretz, Mitchell, et al., 1996; Stecher & Mitchell, 1995; Stone & Lane, 2003). Koretz, Barron, et al. (1996) found that the perceived pressure by teachers to improve scores led

to test preparation materials that resulted in higher scores but did not integrate the desired skills of MSPAP into the broader curriculum. As a result, the researchers felt the score gains on the MSPAP were misleading indicators of success in improving instruction or higher-order thinking.

While teachers reported using more problem-solving, writing, and higher-order thinking skills in the classroom in alignment with the task demands of the performance assessments, several of the researchers indicated that the degree to which classroom actually met these standards is suspect since teacher understandings and definitions of the concepts varied considerably (Koretz, Barron, et al., 1996; Koretz, Mitchell, et al., 1996; Stecher & Mitchell, 1995. Survey and questionnaire data rely on self-reporting, and it may be that respondent perceptions of their actions and the reality differ. In a five-year study of a performance-based math reform in nine Michigan school divisions, Spillane and Zeuli (1999) provided questionnaires to 283 teachers about their teaching practices and then observed 25 of those teachers. The researchers found that while all 25 teachers reported that they taught and used instructional strategies consistent with the problem-solving, real-world focus, and use of manipulatives in the reform, results showed that only four of the teachers were actually teaching that way. Researchers in Maryland felt that respondents may have similarly reported greater alignment of instruction with reforms than actually happens (Parke et al., 2006).

Although teachers may report more problem-solving or higher-order thinking, teacher definition of those terms and understanding of how that translates into student activities and assessment may affect the findings of these studies; therefore, the changes in instruction being reported may not have matched the intent of the reform nor the intent of performance assessments. The differing teacher definitions of what constitutes reasoning, problem-solving, or

higher-order thinking combined with a tendency to overstate alignment with reform agendas raises questions about the nature and extent of instructional changes reported in these studies.

Triangulation Studies

Although survey studies are limited by accuracy of self-reported data, other researchers studying the effects of performance assessment reforms on classroom instruction have sought to triangulate survey data with classroom artifacts and/or observations.

Performance Assessments for State Accountability. Parke et al. (2006) studied the impact of the MSPAP on classroom instruction on the teaching of reading and writing. Through stratified random sampling, the researchers surveyed teachers, principals, and students at 59 elementary schools and 31 middle schools, and then 51 teachers were asked to submit classroom instruction and assessment materials for the study. Similar to the survey-only studies, the surveys revealed that most teachers and principals agreed that MSPAP tasks were an improvement over multiple-choice assessments and that the MSPAP was a useful tool for implementing positive changes to instruction. The analysis of the artifacts submitted by teachers revealed that the majority of classroom activities reflected at least some characteristics of MSPAP. Researcher evaluations showed only 25% of reading activities and 29% of writing activities were not at all like the MSPAP, and 42% of reading activities and 31% of writing activities had low alignment with the MSPAP. Comparing these results to teacher questionnaires, the researchers reported that teachers self-reported greater alignment with MSPAP reforms than the classroom activities reflect. Students reported a lesser emphasis on MSPAP problem types than did teachers, and student perceptions were more aligned with the artifacts presented than with teacher perceptions. Researchers concluded that while the teachers believed in and supported MSPAP reforms, teachers were not able to implement the reforms in practice.

Parke and Lane (2008) conducted a similar study of impact of MSPAP on classroom instruction focused this time on mathematics instruction. This study collected 3,948 instructional and assessment activities being used across Maryland schools to evaluate the extent to which mathematics activities aligned with the learning outcomes of MSPAP. Using a stratified sample of 51 schools, the researchers requested five instructional activities, five assessment activities, and one scoring scheme from teachers in December and again in June. Because teachers self-selected what to submit, researchers could not determine how representative the activities were of the entire classroom experience. However, Parke and Lane (2008) reported that the activities varied widely, and they did not feel teachers necessarily chose the best examples or those most aligned with standards. They found that 48% of instructional activities and 64% of assessment activities did not align with the skills and intellectual processes intended by MSPAP. Additionally, only 10-12% of all activities, classroom or assessment, had a medium to high alignment with the MSPAP outcomes.

Parke and Lane raised the caveat that the documents provided to them could have been implemented at a lower cognitive level than the examples submitted indicated depending on teacher delivery and encouragement of higher-order thinking, which would not be evident in this study. Thus, researchers suggested that the teachers' perceptions that their instruction reflected the processes of MSPAP was overstated when compared to the skills evident in the activities submitted. Similar to the earlier study on reading and writing, while the mathematics classroom materials and documents did not reflect substantial instructional changes, interviews and surveys suggested that state performance assessments were positively influencing classroom experiences.

While Parke and Lane (2008) and Parke et al. (2006) had conducted stratified random samples to measure effects on instruction broadly, Goldberg and Roswell (2000) chose to focus

more narrowly on the instructional changes by teachers who had been trained in and participated in scoring the MSPAP. The researchers sought to examine the effect of scoring MSPAP responses as a professional development strategy to increase teacher appropriation and implementation of MSPAP reforms. Fifty teachers were given pre- and post-questionnaires before and after the MSPAP scoring session. From that 50, 12 teachers were chosen for semi-structured interviews, classroom visits, and submission of classroom artifacts. Similar to the survey-only studies described previously, the questionnaires showed that while teachers were familiar with performance assessment and performance-based instruction, teacher definitions of those terms were often partial or superficial, and two held seriously flawed definitions. Analysis of the artifacts led researchers to conclude that teachers reduced performance-based instruction to one or two of the most obvious and easily adopted elements: hands-on activities and group work.

Goldberg and Roswell (2000) did find that participating in the scoring changed teacher attitudes and instruction. Teachers who participated in the scoring resolved to do more performance activities, use more or better rubrics, utilize more reading or writing, and have students do more explaining or elaborating. Teachers who participated in the scoring process did better at establishing a context and purpose for student tasks and making them more real-world than teachers who had not scored the MSPAP responses. Still, researchers found that even after scoring, teachers tended to think of problem-based activities in limited ways and created mock-MSPAP activities that had the appearance but not the substance of good performance-based instruction. Teachers tended to craft activities that looked like MSPAP activities and were interesting and engaging but had little or no connection to state-mandated learning outcomes.

Thus, the researchers concluded that the perceived professional development benefits of scoring MSPAP tasks were not translating into greater positive impacts in classrooms.

All three of these studies demonstrated teacher-reported commitment to performance-based reforms and attempts to incorporate those reforms in their classrooms, but analysis of classroom activities and artifacts revealed that teachers were not able to fully implement the reforms. The studies that triangulated self-reported data with observations and classroom artifacts found that the instructional changes attributed to performance assessments often focused on the measures being tested or included more writing, rubrics, and group work. Although all of these strategies have the potential to promote higher-order thinking and deeper learning, none of the studies here empirically show that the instructional changes demonstrated increased critical thinking or real-world skills. All three researchers highlighted the need for greater professional development on the purpose and concepts for teachers to better implement performance-based reforms.

School-Based Study. While the previous three studies focused on performance assessment programs for state accountability purposes and the effects on classroom instruction, Pfeifer (2002) created an experimental study that combined surveys with observations and artifact analysis to analyze the impact of performance tasks on student attitudes. In this study, 22 teachers in 13 Lutheran elementary schools were assigned to either a treatment or a control group. Both groups completed pre- and post-study surveys, and a subsample of teachers was selected for interviews. Teachers in the treatment group were given 6half-day training sessions on the rationale and implementation of authentic assessments and were then asked to use authentic instruction and assessment tasks for one nine-week quarter. During this quarter, teachers were observed three times, and the teachers submitted two of their assessment tasks to

be scored on their authenticity to confirm whether authentic instruction was taking place. The findings from this study are limited given the small sample size as well as the construction of the study.

Given this research, 6 half-day trainings do not seem sufficient to train teachers on authentic tasks and instruction. In addition, Pfeifer (2002) did not account for what teachers may have been doing prior to the study that utilized elements of authentic tasks or authentic instruction, nor for teacher prior knowledge of authentic instruction. Despite these limitations, the focus of the study was student attitudes, and the Pfeifer found that authentic pedagogy did not improve student attitudes toward social studies, often students' least favorite class, but students did realize that authentic projects required them to construct knowledge, and students preferred projects over traditional tests because they felt projects were a better way to demonstrate what they learn.

While Pfeifer (2002) focused on student attitudes, part of the study also involved evaluating teacher-made performance assessments according to seven criteria posited by Newmann et al. in 1995. The researcher and a second trained evaluator scored assessments from each teacher in the study on a scale of 1-3 or 1-4 on each of the seven criteria from Newmann et al. (1995) for a total sum with a possible range of 7-23. Since each teacher submitted two assessments, their two scores were averaged; the 14 teacher scores ranged from 16-20 out of the possible 7-23 points. The researcher and a trained evaluator did the same scoring of recordings of classroom observations, and the 14 teacher scores ranged from 11.33-16.67 out of a possible 4-20 points. The researcher than summed the assessment score with the observation score and grouped teachers by overall scores, with two teachers ranked Low Authenticity, two ranked High Authenticity, and 10 ranked Middle Authenticity. These rankings were used in comparing

student survey scores. Since the focus was on student attitudes, the researchers did not elaborate on nor discuss the teacher-made performance assessments beyond reporting the teacher's composite score, but it is unique in that the study used and evaluated teacher-made assessments.

Summary of Triangulation Studies. Studies that triangulated teacher self-reported data with other indicators of quality or change support the findings of survey-based studies. Most teachers in performance assessment related reform initiatives recognized the benefits of performance assessments and felt they are beneficial tools for improving instruction to promote deeper learning for students. The researchers analyzing instructional materials found that teachers' perception of the actual changes they are making to instruction in alignment with that reform is greater than the actual changes in instruction. Researchers found that teachers need more training to go beyond superficial modifications to fully implement the processes and principles of performance-based instruction and assessment.

Case Studies

Survey studies and surveys combined with artifact analysis are all limited by participant self-reporting of data, selection of documents, and the limitations of the documents to reveal how the activities were actually implemented. The research data reveals gaps between teacher perception of instructional changes and actual classroom activities, and researchers still raise questions about how the classroom context of the activities submitted might affect the cognitive demands actually required by students (Parke & Lane, 2008). Although more limited in the number of participants and thus generalizability, case studies allow researchers further insight into classroom experiences.

Firestone et al. (1998) conducted an embedded case study of two schools in Maryland and five in Maine to study the effect of state-mandated performance assessments on math

instruction. Maryland and Maine were chosen for comparison because Maryland's MSPAP has sanctions attached to it while Maine's tests do not. In addition, Maryland had larger central offices with curriculum specialists, while the Maine districts had less than one full-time central office person on curriculum issues. The researchers hypothesized that, given the higher stakes of the MSPAP, there would be greater instructional change in Maryland than in Maine.

For each school, the researchers interviewed a school board member, district administrator, principal, department heads of English, math and social studies, and math teachers. The researchers then observed two classes and interviewed the teachers. As the researchers hypothesized, three times as many teachers in Maryland described instructional changes due to the reform compared to the Maine teachers, and the Maryland teachers referenced more teaching of MSPAP activities. Yet, similar to the studies discussed previously, the subsequent classroom observations found that the reality in the classroom did not match teacher perceptions. In both states, researchers found math teachers used twice as many activities that involved simply computing answers (small problems) as complex problems that involved reasoning skills (large problems), with most activities focused on practice. Despite expecting to see more large problems and non-practice in Maryland, Firestone et al. (1998) found most math instruction in both states focused on small problems with teachers telling students the procedures, not having students develop them. The researchers concluded that while performance assessments can change specific behaviors and procedures, they do not change teacher paradigms and understandings of the content and how it is taught.

Khattri et al. (1998) also conducted a multi-state case study, but their study went beyond just assessing the effects of performance assessments to analyze the facilitators and barriers to assessment reform as well as analyze the key characteristics of performance assessments used by

teachers. In a 3-year study of 16 school sites with a variety of assessment projects and grade levels, the researchers gathered sample performance assessments, policy documents about assessment and related reform efforts, and evaluation and research reports as well as newspaper reports regarding the assessment. The researchers then conducted phone interviews with state and local education offices, the school sites, and external assessment reform organizations followed by 1–2-day site visits consisting of interviews, classroom observations, and professional development observations.

Similar to the survey studies, Khattri et al. (1998) found that the term "performance assessment" was being used to refer to a wide range of instruments, and the only commonality among teacher definitions was that performance assessments are not multiple choice, but are pedagogically useful (Koretz, Barron, et al., 1996; Koretz, Mitchell, et al., 1996; Sivalingam-Nethi, 1997; Stecher & Mitchell, 1995). Beyond the emphasis on constructed-response, the definitions of performance assessments varied by form and demand on students from on-demand tasks and extended tasks to demonstrations and portfolios. Teacher definitions also varied by dimensions such as time demands, application of problem-solving skills, metacognitive demands, and student control. The researchers felt this lack of common understanding and constructs limited the evidence of performance assessment validity. The researchers concluded that before implementing an assessment system, leaders must articulate a clear statement of purpose, clearly aligning the assessment form to that purpose.

Similar to the other studies, Khattri et al. (1998) used classroom observations and interviews to examine the effects of performance assessments on classroom instruction. The researchers found that students were being asked to write and do more project-based assignments, but while the use of writing increased, the quality of that writing was open to

debate. Writing assignments can be descriptive without requiring critical thinking, and students can work in groups on low-level tasks. Students were found to be more motivated and engaged in project-based tasks, but the researchers warned that the change in format of assessments is not enough; the content must be challenging as well. As students in one school responded, performance assessments were better suited to low-performing students.

The changes to instruction found in the study varied in degree. Sites with portfolios showed the most extensive impacts, and sites with on-demand assessments showed the least change, with the most visible change being the use of rubrics. Similar to other studies, as a result of incorporating assessment reforms, teachers reported more in-depth coverage of certain curricular topics and the neglect of other curricular topics. The researchers concluded that the lack of linkages by teachers of content, performance, and assessment strategies made pedagogical change based on the assessment project difficult.

While other studies had also noted the gap between intent and practice, Khattri et al. (1998) used their findings to identify a list of potential facilitators or barriers to assessment reform. The researchers' recommendations included policy strategies such as ensuring the coordination and compatibility of the assessment reform with other existing reforms and policies, establishing reasonable timelines for implementation, and ensuring sufficient professional development for teachers. The researchers also stressed the importance of communication, as the perceived technical soundness of the assessment and public perceptions of the assessments' fairness could also be barriers to the success of an assessment reform, as can the lack of clearly outlined standards. The researchers emphasized that successful assessment reform required appropriation of the reform by teachers. Teacher appropriation could be increased by involving teachers in the designing and implementing of the assessments, providing sufficient professional

development, and the time and ability to experiment with the assessments and modify them to their own classroom in a more loosely prescribed environment. Thus, the researchers concluded that for assessment reforms to be successful, teachers need time and support to build capacity to effectively work with performance assessment techniques.

Khattri et al. (1998) argued that the lack of integration of assessment reform with other existing policies can be a barrier to implementation of assessment reform. This conclusion is supported by the work of Moon et al. (2005). With a goal of designing authentic assessments that promote meaningful learning, the researchers developed four authentic tasks based on state and national standards that were real-life and allowed some student choice. They then conducted a case study of implementing the assessments in a single school setting. After having 46 individuals representing higher education, state-level education specialists, teachers, and division-level representatives review the tasks, each of the tasks was used in one or two classes.

Moon et al. (2005) concluded that performance assessments can provide consistent information on student learning, but the only statistics reported and analyzed were statistics on interrater reliability. No other data on student gains were given, and the connection between the researcher-constructed tasks and the classroom instruction was not defined by the researchers; instead, most teachers reporting having students do most of the work on the task outside of class time. Teachers and students were then asked to reflect on their experiences with the assessments developed by the researchers. Similar to Khattri et al. (1998), Moon et al. (2005) reported study both teachers and students expressed positive responses to the tasks with students finding the rubrics most helpful in guiding their initial process, as a reference point throughout the task, and to check for completion at the end. Moon et al. (2005) did find that teachers were mixed about using the assessments in the future; while some wanted to try to incorporate the tasks into their

instruction more, other teachers were resistant due to the time the tasks required not fitting with the demands of state testing. Although the size of the study limits its generalizability, the study shows the need for assessment reform to align with other existing initiatives for teachers to appropriate the reform.

While the previous case studies examined how performance assessment affected instructional change in terms of classroom activities, Abbott and Wren (2016) focused more directly on the ways that teachers used performance assessments to inform instruction through a case study of one school division. The participating school division had been using locally developed performance tasks for 5 years at the time of the study and was using a protocol for staff to analyze student responses to the locally developed performance tasks. After scoring student responses, a professional learning community (PLC) including administrators, school improvement specialists, resource teachers, and classroom teachers, would complete a form called the Plan for Analyzing, Communicating, and Using Results (PACR). The PACR required school staff to first list their findings from the analysis of student responses and from that identify instructional interventions and next steps. The PLC then decided how the locally developed performance task results would be shared with the rest of the staff and with students, and they developed a plan for student reflection on their performance.

Abbott and Wren's (2016) study was based solely on a content and thematic analysis of the PACR forms from across the division. The researchers concluded from the first step on the PACR forms that a data-driven process was necessary to enhance the use of performance assessments, but the only evidence and discussion to support this conclusion was that teachers had worked in professional learning communities, including resource teachers, to analyze student scores and complete the PACR. From the list of suggested instructional interventions listed on

the PACR, the researchers concluded that the skills in the locally developed performance tasks, such as critical thinking and problem-solving, were incorporated into daily class practice. The only evidence given to support this claim was the list of potential instructional interventions based on the analysis of student scores, such as the use of Paul's Reasoning Model, Socratic seminars, document-based questions (DBQs), and analyzing political cartoons, but the researchers did not examine actual classroom instructional activities to document that this was occurring.

Abbott and Wren (2016) did note that the limitation to the study is that there was no follow up to determine whether any of the listed strategies were utilized and whether the PACR resulted in any effect on teaching and learning. The researchers found schools used different methods of sharing scores, but that students did reflect on their strengths and areas that they felt needed improvement. While the researchers claim that performance tasks helped change the focus of classroom instruction, that conclusion is not supported by the data presented, as the PACR only lists intended actions and not actual classroom events. Despite the limitations of the study, the data analysis tools used within PLCs to analyze student responses to performance tasks and to make informed pedagogical decisions may be strategies that other divisions could utilize.

Summary of Case Studies. Whether survey, document analysis, or case study, these studies on the effect of performance assessments on classroom instructions had similar findings. Teacher and student interview responses reveal instructional changes resulting from the use of performance assessment systems that teachers and administrators regarded as positive. If teachers are incorporating more activities like the assessments and the assessments focus on

important learning outcomes and higher-order thinking skills, than more of those skills are being integrated into the classroom.

But practice might be affected by the nature of loosely coupled policy in which the understanding and implementation of the policy at the local level may not match the intent of the state legislation (Weick, 1976). While the mandates and school policies were designed to promote deeper learning and increased use of performance tasks throughout the classroom experience, most of the studies find limited actual changes to instruction and classroom practice. Although states adopted performance assessment reforms to promote instructional changes, in practice teachers lack common understandings of performance assessments, problem-solving skills, and higher-order thinking required to fulfill the intent of the mandate. In addition, teachers remain wary of performance assessments due to a lack of time and resources required to fully integrate performance assessments into the classroom (Firestone et al., 1998). Thus, the research findings may result from a gap between policy and the implementation (or lack thereof) of professional development, resources, and clear communication of the reform to teachers, not necessarily a flaw in performance assessments.

Studies of Planning for Implementation. The previously described studies illustrate the importance of planning and preparation in successfully implementing assessment reform to improve classroom instruction, and several studies provide insights into strategies that states and school divisions have undertaken when embarking upon the process of assessment reform. Marion and Leather (2015) studied the process chosen by the state of New Hampshire when it implemented its Performance Assessment of Competency Education reform. New Hampshire chose to start its reform deliberately to ensure meaningful incorporation of performance tasks into the accountability system by starting the program in a set of pilot districts. To be accepted

into the pilot, districts had to demonstrate a high-quality set of K-12 course and grade competencies to participate effectively. Rather than provide state-created assessments, the Performance Assessment of Competency Education sought to allow teachers to develop and use their own classroom assessments. Prior to the pilot, the state had run a 3-year training initiative on performance assessments with a group of teachers. Beyond training teachers, the program also started the New Hampshire Test Bank, a repository of quality performance tasks. By the start of the pilot, the district then had models to draw on as the schools developed their own performance assessments. Once drafted, district assessments were submitted to a peer and expert review process for evaluation. Once evaluated and subsequently revised, the pilot districts were then to administer the assessments in all subjects and grades. Districts were not isolated in the process of task, rather the state provided professional development support from the state, and the pilot schools participated in work groups with other districts to share ideas and analyze work. The researchers do not specify the nature of the professional development nor the structures for the work groups but do identify two strategies for policy-setting bodies to consider when mandating reforms: providing support and resources, such as the New Hampshire Test Bank, and ensuring districts were in positions to engage in the reform successfully. The other limitation to this study for researchers and other school divisions is that there was no follow up to this study to measure the success of the preparation taken in New Hampshire.

While Marion and Leather described the state-level policy-making process for planning a reform, Abbott (2016) focused on how one school division prepared to undertake a state-mandated performance assessment reform. Abbott (2016) conducted a descriptive case study of one large Virginia school division's leadership's process for developing and enacting a program of alternative assessments when the General Assembly announced the policy to remove SOLs

and replace them with LAAs. To learn how this division approached implementation of the policy, the researcher conducted semi-structured interviews of four central office personnel involved in overseeing the implementation of LAAs and/or managing the development processes of LAAs as well as attended a regional meeting of 12 school divisions concerning LAAs and a local school board meeting on the same topic. The researcher found that the division already had the development of performance assessments and balanced assessments in the strategic plan, thus the division started the process confidently, feeling the state mandate simply accelerated the division's existing plans. The division first conducted an internal audit of assessments to identify what already existed that met the state guidelines and where there were areas of need. The division then invited teachers to apply to serve on an Alternative Assessment Development committee. Drawing on the work of the Stanford History Education Group and regional collaboration, the Alternative Assessment Development committee worked to create performance tasks and criterion rubrics.

The interviews with central office personnel revealed a concern over teacher ability to use various types of assessment, thus central office staff began planning for professional development in assessment literacy and practice scoring student tasks. The division also worked to develop clear and consistent communication to educate shareholders on the reform. The researcher identified limits of the study as the small number of potential interviewees and the lack of teacher or classroom-level voices. The timing of the study was the year that the division was planning the steps and professional development, so the study was unable to cover the actual training events or the outcomes of this implementation plan; however, the steps taken by the division match with the recommendations of Khattri et al. (1998) to provide professional development for teachers and to communicate with stakeholders.

The studies by Marion and Leather (2015) and Abbott (2016) studied state and local entities embarking on a reform and did not report subsequent actions or outcomes, but O'Brien (1997) focused on teacher responses to a state assessment reform after the first year of the reform. O'Brien's study focused on Kansas, where the reform began with pilot teachers before being fully implemented across the state. The University of Kansas Center for Testing and Evaluation had worked with the Kansas Board of Education and an advisory committee of educators and curriculum specialists to develop performance tasks for social studies that were implemented in Grade 5, 8, and 11 classes. O'Brien surveyed the pilot teachers (n = 2,838) at the end of the first year to measure teacher responses to the reform. The responses revealed only 54% of teachers feeling supportive or very supportive of performance assessments. The teacher concerns noted by the researcher included lack of time to prepare and implement the reform, the poor fit with the existing program, the lack of resources and communication from the Board of Education, and the lack of a coherent and consistent staff development. O'Brien (1997) concluded that while most teachers were receptive to the reform, the success of the reform would depend on securing teacher commitment and more training to better integrate performance tasks into instruction.

Similar to O'Brien, Bandalos (2004) also conducted a case study at the end of the first year of a state-level assessment reform in Nebraska. While the Nebraska School-Based Teacher-Led Assessment and Reporting System (STARS) program did not focus on performance assessments like Vermont and Virginia in the two previous studies, Nebraska did allow each school district in the state to construct its own portfolio of assessments to meet state accountability measures in a manner similar to the Vermont portfolio and the Virginia Balanced Assessment Plan. Bandalos used focus groups and written surveys to analyze teacher attitudes

about the implementation of the STARS program after the first year. Bandalos (2004) found negativity from all districts as they reported feeling confused and frustrated in the beginning, especially as the school districts felt the rules from the state kept changing. All of the districts wanted more guidance from the state on writing different assessment types and on the quality criteria. Beyond that there was variation, as Bandalos found that the school districts' approaches to implementation varied, and those districts that had been working on identifying and creating assessments further ahead of the start of the program (some as early as 3 years prior) felt more satisfied with the reform and had higher quality ratings on their assessments than districts who had not started the process early. Similar to the work groups Marion and Leather (2015) discussed in New Hampshire, Bandalos (2004) reported that districts that decided to join together in consortiums to share ideas, distribute the work amongst more people, and pool resources to hire consultants also had higher satisfaction and higher quality ratings.

Brookhart (2005) also studied the Nebraska STARS program but focused solely on the quality of the local assessments developed and being used for accountability under the Nebraska STARS program. The researcher took a random sample of 300 assessments, half of them reading and the other half math assessments. The Department of Education in Nebraska had an established set of quality criteria for the local assessments developed for STARS:

1. Alignment to content standards,
2. Opportunity for students to learn content prior to testing,
3. Appropriateness of the assessment in terms of development levels and reading levels,
4. Freedom from bias and sensitive situations,
5. Consistency in scoring, and
6. Appropriate cut scores for mastery.

Brookhart's (2005) team developed their own rubrics to evaluate the quality of the assessments. The rubrics were comprised of five criteria:

1. Alignment to the standards,
2. Sufficiency meaning the essence of the standard is measured with enough points,
3. Clarity of task and directions,
4. Appropriateness in terms of length, cognitive level, and bias, and
5. Scoring procedures are clear and can be applied consistently.

Each assessment was scored on a scale of 1-6 points for each of the five quality criteria. Brookhart (2005) found Alignment, Clarity, and Appropriateness scores to have means over 5 and Sufficiency just under 5 at 4.73. The Scoring criteria had a mean of 2.87 since many schools did not submit scoring rubrics.

Based on these standards, Brookhart (2005) concluded that most teacher-developed assessments were of sufficient quality and that teachers were successful at matching assessments to the standards, but that the assessments needed better rubrics. The study does not discuss the types of assessments being analyzed or whether they were selected-response, constructed-response, or performance assessments. The STARS program allowed districts to determine their own set of assessments, specifying that those assessments could be norm-referenced, criterion-referenced, or classroom assessments (Roschewski, 2004). Most districts employed locally developed criterion-referenced tests along with a standardized norm-referenced test such as the Terra Nova or Stanford Achievement Test (Dappen & Isernhagen, 2005). The lack of description in Brookhart's study of the type and format of test being analyzed limits the conclusions drawn from the study.

Summary. These five studies of the initial implementation of state-mandated assessment reforms reveal common themes that need to be considered when engaging in assessment reform at the state or local level. First, time is needed to prepare for the reform. In New Hampshire and in the more successful Nebraska schools, division staff and teachers had been working on developing quality assessments for three years prior to implementing the reform. These efforts were furthered by sharing the work across districts. New Hampshire intentionally created work groups with structures for sharing ideas and comparing student work, while the more successful Nebraska districts created consortiums on a more grassroots level. In both cases, even though given autonomy to develop their own assessments, districts benefited from peer support as they navigated the new reform.

Secondly, these studies stress the importance of policy makers and educational leaders establishing clearly defined policies and goals that align with existing policies and initiatives. Both Moon et. al. (2005) and O'Brien (1997) found teachers were reluctant to embrace performance assessment reforms because the time required to implement them prevented teachers from meeting the demands of other initiatives and curricular demands. Bandalos (2004) found frustration and confusion arose from unclear policies and the subsequent attempts to clarify policy being perceived as constantly changing the expectations of teachers. Clearly explained and communicated policies that coordinate with other policies and initiatives allow teachers to understand and adopt the reform as well as communicate the reform to other stakeholders (Khattri et al., 1998).

The other common theme is the centrality of professional development to prepare teachers to implement the reform. New Hampshire had a 3-year training program with designated teachers before implementing the reform in the pilot schools. The Virginia school

division in the Abbot (2016) study was planning professional development and opportunities to practice scoring student responses as an early step in its implementation plan. Marion and Leather's (2015) and Abbott's (2016) research took place at the outset of the reform there was no follow up to investigate the outcomes of the plans and the effect on teacher attitudes and classroom instruction. Both Bandalos (2004) and O'Brien (1997) reported teacher discontent and frustration with lack of clear communication and staff development from the state-level policy-making body that made implementation more confusing and difficult for teachers. The study by O'Brien illustrated the problems with not preparing and developing teachers, as the teachers implementing the reform revealed their unfamiliarity with a skills-oriented system. The studies stress the importance of time for teachers to learn about and plan for assessment reform and how to integrate the reform with existing instruction.

Reliability Studies

While the previous empirical studies focused on the extent to which performance assessment reforms were having the desired effects on classroom instruction, other empirical studies have focused on the reliability of the scores obtained from the performance assessments in these reforms. When implementing performance assessments for accountability at either the state or local level, student scores are being used to make decisions about the student and/or the school. For performance assessments to provide useful data, the assessments must be reliable, providing consistent, dependable results that reflect student outcomes rather than chance, systematic error, or bias (Gareis & Grant, 2015). One of the issues examined in the literature is the generalizability and reliability of students' scores on a given task.

To analyze sampling variability and generalizability of scores on performance assessments, Shavelson et al. (1993) focused on analyzing student scores from three studies: one

involved 186 fifth and sixth graders on three science tasks, one with 105 sixth graders on a math performance assessment, and the third with 120 students performing problem-solving science tasks at five stations. The analysis of student scores revealed that a major source of error was the task students performed, with 59% of total variability tied to the task. Even with tasks that were designed to assess similar student understandings, students performed better on some tasks than others. The researchers concluded that many tasks were needed to obtain generalizable measures of achievement. In a similar study the next year focused only on science performance tasks, Gao et al. (1994) randomly selected 15 students from each of 40 randomly selected schools in California. The students then went through five stations performing performance tasks at timed intervals. Teams of teachers then scored the student responses. The researchers found that performance task scores were inconsistent across tasks and raters. As a result, the researchers concluded that many tasks were needed to obtain generalizable measures of achievement.

McBee and Barnes (1998) found similar results to Gao et al. in a smaller study of 73 eighth graders in a single middle school. McBee and Barnes also used a set of tasks, four math tasks, to analyze how task similarity affects the ability to generalize the results of performance assessments. The students completed the four tasks and then were retested three weeks later with no discussion of the tasks in the interim. While interrater reliability was acceptable (.77 to .91), even with highly similar tasks examinee performance was influenced by the task. Two of the tasks were completely interchangeable other than the setting changed from a basketball camp to a space camp. Still, task sampling was the major source of error and 95% of the time student scores varied as much as 2.7 points on a scale of 0-4. The researchers concluded that with task consistency so low, generalizability of scores could not be obtained and that the use of performance assessments in high stakes testing was not appropriate.

Stecher and Klein (1997) found similar results with science tasks in a state assessment reform. In a study focused on the cost of hands-on science tasks, researchers administered and scored two fifth-grade tasks of 45-55 minutes each and four sixth-grade tasks lasting half a class period across 38 classrooms in 16 schools for a total of 1100 students. Students also took the science section of the Iowa Test of Basic Skills. Similar to the three previous studies, the researchers found high task variability among the tasks. Based on the student scores, Stecher and Klein concluded that it took three to four class periods of performance tasks to produce a score for an individual student that was as reliable as a one period-long multiple-choice assessment.

Although Stecher and Klein (1997) reported on reliability findings, the study was designed to compare the cost of hands-on assessments to multiple-choice assessments. The researchers found the average cost of one class period was $100 per student, and fixed costs accounted for 70% of that amount. Stecher and Klein pointed out that the high task variability of scores contributed to these higher costs, as performance tasks required several rounds of piloting and revision to ensure the adequacy of directions and clarity of the task. While multiple-choice assessments also require revision, the researchers argue that the piloting and revision process for multiple-choice measures costs 2/3 less than the development process of performance tasks. Thus, one class period hands-on tasks cost 60 times as much as a multiple-choice test. Most of those costs recur annually because, due to the more memorable natures of the tasks, they cannot be reused as often as multiple-choice questions. Stecher and Klein argue that given the variability of performance task scores and the lack of generalizability of performance task scores, it may be difficult for schools to justify the costs of administering that many performance tasks when it is not clear how to interpret student scores.

Webb et al. (2000) also examined the importance of occasion as a source of error variance in performance assessments. With a sample of 662 seventh and eighth grade students across five schools in a division, the researchers gave students two assessments: one with manipulatives and one with similar tasks but only on pencil and paper. Students took the same two assessments a month later. Mean performance did not differ by rater, the scores on the pencil and paper version were similar to those on the manipulatives version and means improved from the first testing to the second. The major source of score variation found by the researchers was the interaction of task sampling and occasion sampling, and the relative standing of examinees was not consistent from one task to another. Like the previous studies, the researchers concluded that dependable measures of student performance may require greater task sampling, but they add that taking occasion into account changed the level of generalizability, so tasks should be administered on different occasions.

Reliability of student scores depends not only on task and occasion, but also on the scoring of the tasks. Kan and Bolut (2014) analyzed the impact of rubric use and teaching experience on teachers' performance assessment scoring behaviors. For their study, 17 math teachers, ranging from one to 26 years of teaching experience, scored the math performance task scores of 50 students. The teachers scored the tasks multiple times: once with no rubric, three weeks later with a rubric, and 10 weeks later with the same rubric. The researchers found higher interrater reliability when using the rubric. But the researchers found that the less complex questions had greater consistency, even with the rubric, than the more complex tasks.

Although Kan and Bolut (2014) examined the effects of rubrics on student scores, Taylor (1998) explored the effects of three different methods of mathematics scoring on student scores: holistic, trait, and item by item scoring. Thirty districts in Washington state volunteered to

participate, and from that pool, one classroom in each of five divisions was randomly selected for a total of 202 students. The students all completed a math performance task that was scored by a team with each of the scoring methods. The researchers found that the different scoring methods measured different elements of student performance creating variability of scores. Taylor argues that different scoring methods exist for different purposes and have different outcomes. The researcher warns that narrow labels of the skills measured by a particular task may result in a loss of information about student outcomes. Thus, scoring methods must be considered when evaluating the reliability of performance task scores.

Summary of Reliability Studies. Student scores on performance tasks have been found in several studies to vary considerably from task to task depending on the nature and context of the task. Student scores have been found to vary depending on the nature of the prompt, the nature of the task, the day, or the setting; this makes it difficult to generalize a student's ability to transfer knowledge to new situations or reliably state whether a student understands a particular skill (Gao et al., 1994; McBee & Barnes, 1998; Shavelson et al., 1993; Webb et al., 2000). Thus, researchers have found it takes a set of performance tasks to produce a more reliable, generalizable score of student performance. The studies reviewed here complicate the use of performance assessments for teachers, because the length and complexity of performance tasks make it difficult to construct assessments that include a large enough number of tasks to establish a reliable score (Stecher & Klein, 1997).

The reliability of student scores is further complicated by the lack of interrater reliability due to the complexity of performance tasks, meaning student scores could vary depending on who scored the work and the scoring method used. As tasks become more complex, researchers have found it more difficult to get scorer agreement (Baxter & Glasser, 1998; Dunbar et al.,

1991; Taylor, 1998). The use of rubrics can decrease variation. The challenge for assessment reform is constructing rubrics that clearly define the constructs to be assessed and that clearly delineate the student behaviors or responses that demonstrate those constructs. Training teachers in the use of the rubric and in developing consistent application of the rubric can strengthen interrater reliability and produce more reliable student scores.

Validity Studies

Besides reliability, another concern surrounding performance assessments is whether the tasks actually measure the designated learning outcomes, providing adequate information to infer what students have learned, thus measuring what the assessment was intended to measure (Gareis & Grant, 2015).

Construct Validity Studies. Several studies have analyzed performance assessments to determine whether or not the performance tasks demand that students engage in the skills and behaviors the task purports to measure. S. G. Grant et al. (2004) conducted a document analysis of the DBQ on the New York Regents' Exam to evaluate the authenticity of the DBQ as a social studies performance assessment. The researchers found that DBQ performance tasks fail to engage students in the tasks of a historian. The researchers argue that historians do not begin with a set question; rather, they follow their interests and define a question and thesis from their research. In contrast, the DBQ presents students with a highly structured question to which the student has no personal connection. Historians select their own sources and verify those sources by comparing them to other sources, while the DBQ presents students with a heavily edited set of sources that inconsistently identify the author, location, or date. The researchers conclude by questioning whether any classroom-based assessment can be authentic. The DBQ is often used as a performance assessment in social studies, not only on the New York Regent's Exam, but also

on the AP history exams, to assess student ability to summarize, contextualize, evaluate, and use documents in an argument. But this study raises questions as to whether DBQs as constructed actually require students to employ and demonstrate those skills and what teachers can conclude about a student performance on a DBQ.

Just as the structure of a DBQ may or may not require students to actually engage in higher-order thinking and demonstrate the skills teachers seek to assess, the use of visual images is often used in performance assessments without actually requiring students to demonstrate the skills teachers are trying to measure. Suh and Grant (2014) analyzed the usage of visual images on the National Assessment of Educational Progress (NAEP) assessments to determine what cognitive level was required to answer each question. Only 49 of the 246 NAEP questions used visual images, which included primary sources, photographs, cartoons, paintings, and posters; most of the questions using visual images tended to be short constructed-responses items. Suh and Grant found that visual images were more often used for historical analysis and interpretation level questions than for knowledge and historical perspective. While the questions with visual images were of a higher cognitive level, the researchers found that easy items with a visual image were more frequent than an easy item without a visual image, though the researchers stated that the scaffolding provided by the images might make complex ideas more accessible. When looking at the content of the questions, those with visual images focused mostly on observation, association with events, or summarizing and stating the purpose of the image. Suh and Grant (2014) noted that the questions failed to ask students to evaluate the validity and reliability of the images, creating the assumption that the images were trustworthy. The researchers concluded that the images, a strategy seen as a way to engage students in higher-

order thinking and demonstrate the skills they have learned, only measured basic knowledge and failed to adequately demonstrate historical thinking.

Similarly, O'Brien (1997) in his aforementioned study of the pilot of the Kansas social studies performance assessment reform found that performance assessments may not reflect the student outcomes that developers intended. In addition to surveying teachers, O'Brien also collected student work samples from the pilot in fifth, eighth, and eleventh grade classes. Students were asked to read a prompt, create a plan to address the prompt, draw information from a variety of sources, and write a response. O'Brien found that there was often no connection between the student's planning of the project and the work produced. Similar to the findings of the previous two studies, O'Brien found that students engaged in a performance assessment failed to engage in the skills the assessment was designed to measure. Students summarized information from sources accurately, but they did not engage in higher-order thinking to compare the sources, and most students struggled to distinguish different perspectives. Most students simply submitted a research paper that did not make an argument nor integrate the information from the sources meaningfully. O'Brien argues that this could be due to the poorly written nature of the question, and he suggests that revisions to the directions would improve the assessment moving forward.

The possibility of task construction reducing construct validity was also addressed by Goldberg and Roswell (2001). Using the MSPAP scores, the researchers analyzed 60 activities used by MSPAP that contain elements of both writing/language arts and another content. The goal of the study was to analyze which student outcomes the items measure, for while the items were currently scored for one student characteristic, the researchers felt that each of the 60 items could potentially be multiple-measure items. The researchers critically analyzed the tasks to

determine what skills and understandings were required of students to complete the task. While the study sought to determine whether items could measure two different student outcomes, what the researchers discovered in their analysis of the items was a concern about construct validity. Goldberg and Roswell (2001) found that some MSPAP questions were cognitively more challenging, but other questions were difficult because they were poorly worded and formatted, thus preventing students from adequately demonstrating skills and understandings. The researchers argued that all performance assessments need a close and critical analysis of the effect of the item language and format on student performance.

These studies reveal that performance assessments do not necessarily require students to engage in the level of problem-solving, critical thinking, and deeper learning that are intended in a performance assessment. Writing assignments can be descriptive without being critical (O'Brien, 1997). On the surface, using images or documents can seem like a performance task, but the structure of the task may not require students to go beyond knowledge in their responses. Thus, tasks that purport to be or are accepted as performance assessments may not meet that label in their current structure. Teachers and developers need to consider the nature of the task in order to construct performance assessments that actually enable and require students to engage in critical thinking, thus establishing construct validity. Once they are constructed, tasks, including the directions and the resources provided to students, need to be critically analyzed to ensure that the wording and format do not reduce the cognitive level of the task nor create barriers for students to demonstrate understanding of the constructs to be measured. If activities are highly teacher-directed, in which the teacher defines the problems and provides the resources, then students are not engaging in the skills of historians who pose problems, search for documents, evaluate and source the documents, and create their own argument (McCann & McCann, 1992;

S. G. Grant et al., 2004). Since performance assessments provide feedback to teachers and students about student progress and may be used for accountability purposes, it is critical that the assessments measure what they are intended to measure and provide accurate evidence of student achievement (Stecher, 2010).

Concurrent Validity Studies. Other studies have sought to investigate the validity of performance assessments by determining the concurrent validity of the assessment with another assessment. The validity of a performance assessments' ability to measure student achievement has been called into question when high-achieving students, as measured by other standardized and accepted tests, did not score well on performance assessments. Koretz and Barron (1998) examined the validity of student gains on Kentucky's KIRIS assessments by comparing student scores on KIRIS to both NAEP and the ACT. The researchers took students' scores on KIRIS from 1992 through 1996 and first compared student scores to NAEP scores. Koretz and Barron argued that when KIRIS was created, the developers were strongly influenced by the NAEP assessments and intended there to be congruence between KIRIS and NAEP; however, they found that while KIRIS was all constructed-response items, NAEP consists of both multiple choice and constructed response items. The researchers found that from 1992 to 1996, students demonstrated large gains in KIRIS scores; for example, three-quarters of a standard deviation gains on the KIRIS fourth-grade reading assessment, but little to no change in NAEP scores over the same time. Thus, KIRIS scores were increasing at three to four times the rate of NAEP gains, and in some cases the large KIRIS gains corresponded to no change in NAEP scores.

Koretz and Barron (1998) then compared the KIRIS scores to the ACT. They accounted for the fact that not all students take the ACT and only compared the KIRIS scores of those students who sat for the ACT. Similar to the NAEP scores, the study found large gains in KIRIS

scores and no corresponding gains on the ACT. The researchers acknowledge that the format of KIRIS differed from the other two assessments, but they concluded that the difference in gains revealed inflation of KIRIS scores and score gains. The researchers argued that their earlier survey study reinforced this conclusion, as few teachers reported believing gains in knowledge and skill had led to KIRIS score gains; rather, teachers cited test preparation activities, practice tests, and the teaching of test-taking skills to account for artificial KIRIS score gains (Koretz, Mitchell, et al., 1996). Koretz and Barron (1998) recommended that performance assessment reforms need to set realistic goals for improvement and assessments must be designed to minimize score inflation, particularly by limiting how many tasks are reused from one year to the next.

Strong and Sexton (2000) also questioned the concurrent validity of Kentucky's KIRIS scores due to the ways the scores were being used. KIRIS assessments were designed to measure school performance, not individual achievement; however, individual scores were being reported and used. Strong and Sexton surveyed all the public high schools in Kentucky to get the names of all National Merit and National Merit Commended students; 119 schools responded. Strong and Sexton (2000) then compared the scores of the National Merit and National Merit Commended students' KIRIS scores. The researchers found that 47% of National Merit students and 73% of National Merit Commended students scored non-mastery on the KIRIS reading assessment and 44% and 70% respectively were non-mastery on the KIRIS math assessment. Strong and Sexton concluded that KIRIS had no concurrent validity, and they then questioned the validity of using performance assessments in making high stakes decisions.

Visintainer (2002) conducted a similar concurrent validity test with Maryland's MSPAP. Using the scores of 982 students from a small rural school system over three years, the

researcher compared student's fifth grade MSPAP scores with their fourth grade TerraNova scores, a nationally normed general achievement test with selected responses items. Regression analysis produced statistically valid predictions of MSPAP scores from TerraNova scores and found enough consistency to suggest a degree of concurrent validity.

On a smaller-scale, Crehan (2001) examined the concurrent validity of a district-level assessment reform compared to a national test. The school district had revised the assessment program from a purely multiple-choice test to an assessment that consisted of 45-60 multiple choice questions and two performance tasks in both English/language arts and math called the Curriculum-Based Performance Assessment. The study took the scores of 6000 third and fourth graders from across the district and compared the performance task scores to the Comprehensive Test of Basic Skills/4, a national test, and to the multiple-choice part of the Curriculum-Based Performance Assessment. While interrater agreement was rated moderate, the major source of score error was due to task sampling. Based on the analysis of the scores, Crehan concluded that there was no evidence of concurrent validity between the performance assessments and either of the other two measures. Crehan further concluded that the lack of validity evidence for the performance tasks was further problematic since sampling of skills or content on a performance assessment is far less than what can be sampled on a multiple-choice assessment.

The limitation of these studies is that they compare tests with different purposes that measure different outcomes, thus student score differences may differ based on student ability on the constructs being assessed. Brown-Kovacic (1998) compared student scores of 280 sixth and seventh graders from five middle schools in a school division in its second year of a math performance assessment project. Brown-Kovacic took student scores on eight math performance assessments and compared them with student scores on the Iowa Tests of Basic Skills and found

that math performance assessments which measure analytical thinking measure different student attributes than traditional, standardized tests, and thus student knowledge is best assessed through multiple measures. Sivalingam-Nethi (1997), in the survey study summarized previously, examined the extent to which performance assessments duplicated the measurement role of multiple choice. An analysis of the correlations between the multiple-choice portions and the performance components of the school-based assessment showed that performance tasks did not just duplicate the multiple-choice data but provided unique information on student learning. Sivalingam-Nethi (1997) did not elaborate nor explain what unique information was conveyed through student scores. The KIRIS and MSPAP performance assessments are designed to measure constructs of deeper learning based on state standards, while the national standardized multiple-choice tests tend to be centered on content knowledge of broader math and language topics. If the assessments have different purposes and measure different constructs, then the scores would not be comparable, since performance assessments are purposefully designed to measure different constructs than multiple-choice assessments.

Conclusions

The empirical studies reviewed in this study present three themes: the effect of performance assessments on classroom instruction, the reliability and generalizability of performance assessment scores, and the validity of performance tasks. The studies provide evidence that while teachers support performance assessment reforms and perceive that the reforms improve classroom instruction, teacher definitions of performance assessments vary widely, and classroom instructional changes due to the reform are not as apparent in practice. Thus, the researchers recommend more time for implementation and professional development for teachers. The reliability studies caution about drawing inferences about students from

performance task scores, as student scores vary by task, occasion, scorer, and scoring method. Researchers conclude that it takes numerous tasks from any one student to get a generalizable score. They recommend the use of rubrics and training educators scoring on the rubrics to increase interrater reliability. Finally, the literature highlights the need to carefully construct and analyze tasks to ensure the tasks and rubrics require and measure the student outcomes intended to be assessed by the performance task.

Quality Performance Assessments

While performance assessments may vary in duration, prescriptiveness, and authenticity, the literature provides common measures of quality to promote the validity and usefulness of the assessments for improving teaching and learning. In 1998, Wiggins defined quality assessments as being open with the task, criteria, and standards made clear to all stakeholders, thus credible to all stakeholders, honest yet fair, rigorous and thought-provoking, anchored in authentic performance tasks, providing data for teachers and students to self-correct, providing feedback for students and teachers, and modeling exemplary instruction while measurably improving student performance. That same year, Khattri et al. (1998) stated that quality performance assessments had a clear statement of purpose, had procedures to continually evaluate the validity, reliability, and meaningfulness of the assessment, and were fair to individual students. More recently, Chappuis et al. (2017) defined quality assessments as having a clear purpose, clearly addressing learning targets with systems to communicate results, and that promote student motivation and cause learning. All three definitions include a focus on clear learning targets and purpose, the purpose of the assessment to promote learning, not just measure achievement, and the importance of being meaningful and engaging to students. Yet each has different emphases and varying degrees of specificity in describing a quality performance assessment.

The Stanford Center for Assessment, Learning, and Equity (SCALE), Jay McTighe, and Chris Gareis have all created tools that provide more specific criteria for quality performance assessments. SCALE (2014) has seven criteria with four of the seven having two or three sub-criteria: worthwhile student outcomes including aligned to standards, deeper learning, and big ideas; performance focused with clear prompts and tasks aligned to the learning outcomes; engaging to students in terms of relevance to the audience, purpose, and discipline; providing for student choice; accessible to all students and appropriate resources; tied to the taught curriculum; and having opportunities for feedback. McTighe (2016) has eight required and three optional criteria: aligns with standards, requires higher-order thinking, authentic context, requires explanation, includes a rubric with clear criteria, has clear directions, is feasible, does not contain bias and could use technology, more than one subject, and provides some student choice. Gareis (2017) identified 10 criteria: progress student learning, focus on important knowledge, skills and dispositions, aligned to intended learning outcomes, requiring higher-order thinking, authentic tasks, containing a clear prompt, response formats matching the intended learning outcomes, requiring a verbal expression of reasoning, evaluated with a rubric, accessible to all, require instruction to promote deeper learning, and clear direction for teachers. All of these measures of quality performance assessments share an emphasis on validity and authenticity while demonstrating a shared construct of how to measure quality performance assessments.

The quality criteria established by SCALE (2014), McTighe (2016), and Gareis (2017) share critical criteria for measuring performance assessments. All three focus on the alignment of the assessment with the standards to be measured--both content and skills--and the need for the assessments to require higher order thinking and/or 21st century skills. SCALE and Gareis further emphasize the need for deeper learning competencies in additional criteria beyond what

is indicated in McTighe. All three evaluate performance tasks on their authenticity to both the purpose and audience of the task, as well as to the discipline, fitting with the definition of performance assessments that students engage in real-world tasks and engage in the skills of the discipline being studied. Finally, they all emphasize the need for a clear rubric or success criteria and clear directions for students. Gareis and McTighe both emphasize the feasibility of completing the task given the classroom limitations on resources and time and the elimination of biased language that would impede student success. Gareis and SCALE also include a focus on developmentally appropriate tasks, accommodations for different learner abilities, and an emphasis on skills that are transferable to other settings and situations (McTighe, 2016; SCALE, 2014; Wren & Gareis, 2019). These tools provide schools and teachers with guidelines for constructing quality performance assessments as well as ways to evaluate existing performance assessments to improve the validity and usefulness of the assessments for teachers and students.

Best Practices for Developing Performance Assessments

The literature argues for deliberate, thoughtful construction of performance tasks that require establishing a clear context and goals, carefully creating activities that align with those learning outcomes, and clear, meaningful scoring criteria that ensure students engage in critical thinking (Stosich et al., 2018). Assessment developers need to start the process by focusing on the construct or learning outcome to be measured and then build the performance task to ensure that the task truly addresses the construct, not start with the task and build the rationale (Linn & Baker, 1996; Messick, 1994). The first step to constructing quality, meaningful performance assessments is to identify the learning standards that the assessment is to measure and clearly identify the knowledge, skills, and understandings students are intended to learn (Brookhart, 2015; L. Grant & Gareis, 2015; Lane, 2014). This process requires consensus among assessment

developers about the purpose of the assessment and which skills and knowledge should be assessed (Khattri et al., 1998). Once the intended learning outcomes to be assessed are determined, the next step is for assessment developers to determine what evidence or responses from students would demonstrate student progress and success criteria for the intended learning outcomes (Lane, 2014; Shiel, 2017; Wiggins & McTighe, 2005).

With a clear understanding of what outcomes will be assessed and how success on those outcomes will be defined, assessment developers can start to construct quality performance tasks that are appropriate for providing the identified evidence (Brookhart, 2015; Chappuis et al., 2017). Once developed, tasks need to be reexamined to ensure that they fully capture the target learning outcome and do not include irrelevant constructs that were not intended to be measured to ensure the validity of inferences made about student learning and success (Lane, 2014). Teachers and administrators developing performance assessments must engage in a thoughtful, deliberate process of identifying the constructs and understandings to be measured and then creating the task, regardless of the length or scope of the assessment. If teachers are constructing the assessments, this will require time and training on the part of the teacher to define the constructs, what those constructs look like in student responses, and design assessments and rubrics that require students to demonstrate those skills and measure them.

Once tasks and rubrics are constructed, the assessments still require analysis before implementation with students and drawing inferences about student achievement. One step suggested by researchers is to have outside experts review the tasks and the rubrics to ensure the content and processes are being assessed and to identify potential biases in the tasks or language (Lane, 2014; Moon et al., 2005; Wren & Gareis, 2019). Tasks should then be field tested with students, and samples of student work should be analyzed by the task development team to

evaluate whether or not the tasks and rubrics produced the desired processes by students and adequately assessed the level of student understanding (Brookhart, 2015; Khattri et al., 1998; Lane, 2014; Pecheone & Kahl, 2014).

One method of determining whether the tasks required the desired skills and thought processes would be to ask students to think aloud while completing the field test items to identify the actual cognitive processes the students engage in (Lane, 2014). In addition, field tests should gather data on the length of time it took students to complete, appropriateness of the vocabulary of the questions, fairness of the prompt and materials, and the feasibility of implementing the task for teachers and students (L. Grant & Gareis, 2015). Based on the feedback from external reviewers and the insights from the field tests and student work samples, the performance tasks and rubrics should be modified and revised (Lane, 2014; Pecheone & Kahl, 2014; Wren & Gareis, 2019). This presents challenges for school divisions, as Maryland teachers interviewed as part of a study of the MSPAP felt overwhelmed by the demands of creating and field-testing complex performance tasks and desired greater resources and professional support in the process (Goldberg & Roswell, 2000).

Even when following these steps, creating effective performance assessments can be difficult for teachers and curriculum developers, many of whom lack the background experience to develop quality performance tasks (Chappuis et al., 2015; Wren & Gareis, 2019). Often teachers or developers choose performance tasks because they are seen as interesting or engaging and then the teacher or developer constructs a rationale for how the task measures desired skills (Linn & Baker, 1996; Wiggins, 1998). Additionally, the wide range of tasks considered performance tasks that range from short, possibly twenty-minute constructed response activities, to multiple week projects complicates the training of teachers. Designed assessments might

appear complex and authentic, but closer examination might reveal that the problem could be solved by memorizing a formula or that the real-world elements are simply tacked on and the problem itself is rather basic and mundane (Cumming & Maxwell, 1999; Linn & Baker, 1996). Careful attention must also be paid to whether the performance tasks are actually requiring thinking that is more critical or if the difficulty arises out of poor task construction and wording or if the question can be answered with general intellectual ability rather than a result of the instructional experience (Goldberg & Roswell, 2001; Linn & Baker, 1994).

The difficulty of constructing quality performance assessments is seen in 12 states joining together to form the Innovation Learning Network to pool their resources to develop quality assessment systems and the teams of researchers who spent five years developing authentic assessments with expert reviewers and pilot programs (Moon et al., 2005; Stosich et al., 2018). New Hampshire has attempted to assist teachers by creating a test bank of quality performance tasks where teachers from across the state can contribute and use assessments to share the burden rather than all schools developing their own (Marion & Leather, 2015). The challenge for school leaders is using performance assessments properly and best developing teachers to implement performance assessments and the requisite instructional changes correctly.

The Role of the Teacher in Performance Assessment Development

Since the purpose of performance assessments is to improve teaching and learning, teachers must understand the purpose and usage of performance assessments to integrate the desired skills and understandings into curriculum and instruction (Khattri et al., 1998; Wiggins, 1998; Wren & Gareis, 2019). Darling-Hammond (2017) recommends that performance assessments be embedded in classroom instruction where teachers analyze the assessments to diagnose student strengths and needs to give students quality feedback and modify instruction for

student success. The shift to performance assessment and resulting impacts on instruction and pedagogy require learning on the part of teachers that requires quality and sufficient professional development (Khattri et al., 1998; O'Brien, 1997; Stosich et al., 2018). Identifying the constructs to be measured based on state and national standards and then creating functional understandings of what constitutes acceptable demonstration of conceptual understanding or effective communication is overwhelming to teachers who lack experience and training in this work (Darling-Hammond & Ancess, 1996; Goldberg & Roswell, 2000; Khattri et al., 1995). Given the complexity of defining learning outcomes and constructing assessments that measure these outcomes, teachers need considerable support to effectively implement performance assessments.

Khattri et al. (1998) argued that for teachers to regularly employ the skills and higher-order thinking skills of performance assessments to enhance pedagogy and classroom instruction, teachers need to "appropriate" performance assessments (p. 85). Appropriation means that teachers believe in the value of performance assessment to provide valuable information about student progress and the effectiveness of instructional strategies and curriculum. Khattri et al. (1998) posited that teachers can only appropriate performance assessments, thus integrating the benefits of them into the learning environment on a regular basis, when teachers are involved in the design, implementation, and scoring of the assessments. In addition, the level of prescription of performance tasks impacts teacher appropriation, as tightly prescribed assessments promote a lower degree of appropriation while more loosely prescribed systems allow teachers to adapt the skills for their own classrooms and better integrate them into regular teaching practice (Khattri et al., 1998). Given the importance of teachers integrating the skills and understandings of

performance assessments into the classroom setting in a meaningful way, teachers must understand, be versed in, and be able to develop performance assessments.

The struggle to construct meaningful performance assessments is not just due to the difficulty of understanding how to construct the tasks, but also having the time and resources to devise and field test complex performance tasks (Baron, 1996; Goldberg & Roswell, 2000). As a result of the lack of time and understanding, many teacher-created assessments end up being hands-on activities that lack the depth of true performance tasks and are often a medley of assorted activities that have been inserted into existing instruction (Goldberg & Roswell, 2000; Gong & Reidy, 1996; Messick, 1994; Firestone et al., 1998). Schools and districts need to provide teachers with training, models of performance assessments, time to plan and collaborate in order for performance assessments to bring about the instructional reform envisioned by researchers, and proponents of these assessments (Goldberg & Roswell, 2000; Khattri et al., 1995). Wiggins (1998) envisions considerable time, specifically five four-hour blocks, spread out through the year to collect data, design assessments, discuss teacher experiences with the assessments, and adjust the assessments and instruction. Khattri et al. (1998) also argue that substantial release time provided throughout the year for planning and collaborating promotes teacher creativity and innovation, but lack of professional development time leads to teacher frustration and exhaustion. This extensive, resource-intensive professional development for teachers would require the commitment of time, money, and materials by leadership to help teachers become knowledgeable about assessment design, implementation, and scoring (Khattri et al., 1998; Wiggins, 1998).

In addition to time within the school year for teachers and other assessment developers to collaborate, time needs to be allowed for the reform to be fully realized. Implementation of

performance assessments takes time to develop, pilot, and refine performance tasks and for teachers to appropriate performance assessments and alter instruction to embed these strategies (Brookhart, 2015: Khattri et al., 1998; Wren & Gareis, 2019). In the interim, Khattri et al. (1998) argue that school leaders need to remain committed to the reform, not introducing contradictory or incoherent policies. When initiatives are not consistent, teachers and administrators hesitate to fully invest in an initiative like performance assessments. Given the time performance assessments take and the time it takes to teach the skills necessary for student success on performance assessments, if other policies demand teachers cover all the standards, teachers feel caught between depth and breadth and feel they must sacrifice one to meet the demands of the other. Thus, states must reevaluate the standards and expectations if they are intent on implementing performance assessments and higher-order thinking (Khattri et al., 1998). The time and resources required to implement quality performance assessments creates a challenge for states, like Virginia, who are allowing local divisions to develop their own performance assessments to ensure that all of the developers and school divisions both understand and have the resources to engage in this difficult task and that other local school division policies are consistent with the larger performance assessment reform (Darling-Hammond, 2017; Khattri et al., 1998).

Performance Assessments in the Social Studies

While most of the state performance-assessment programs and much of the empirical research focuses on math, language arts, and science, Virginia has chosen to begin the transition to performance assessments primarily in the social studies, as three of the five courses required to use LAAs are social studies. In the standardized-testing era, social studies tests were, as Wiggins (1993) says, “the worst offenders” (p. 5) in terms of reducing student knowledge to

discrete facts and often likened to preparation to play Trivial Pursuit (Biermer, 1993). The current movement in Virginia to eliminate multiple-choice SOL tests and replace them with performance assessments reflects a widespread criticism of the social studies SOL tests for emphasizing memorization and recall of facts rather than assessing historical skills (VanHover et al., 2010). Standardized, multiple-choice questions focused on the recall of specific information do little to measure student understanding of the concepts or skills of social studies (Wiggins, 1993). In a discipline that focuses on argument, interpretation, and the understanding of multiple perspectives, performance assessments promise the ability to better measure the nature and purpose of social studies disciplines (Biermer, 1993; S. G. Grant et al., 2004).

Despite the natural connection between social studies skills and performance assessments, implementing performance assessments in social studies presents several challenges. One of the greatest challenges to incorporating performance assessments in social studies is the lack of consensus about the desired outcomes to be measured. Baker (1994) gathered a team of psychologists, history teachers, university-level historians, psychometricians, and students to design and validate new performance assessments to measure history understanding. As the team engaged in the process of task, they found both lack of agreement as to what content to teach as well as epistemological differences. When attempting to define the content to build tasks around, the team reviewed all the history textbooks adopted by California high schools to identify topics that received extended or deeper treatment by the texts. Baker (1994) wanted to ensure that all students had exposure to the content of the task, but this proved impossible as they found few sets of common, detailed content across the textbooks. The other concern of the team was constructing tasks that would apply in any classroom given varying epistemological differences. History experts and particular teachers interpret events and figures

in the past differently, and Baker's task development team did not want a student's work to be scored lower because the student response did not match with either widely accepted historical arguments or those of a particular teacher (1994).

The epistemological concerns in social studies education discussed by Baker (1994) are also seen in state-level debates about the content of state social studies standards. O'Brien (1997) describes the process of the Kansas Board of Education's committee to draft state social studies standards and struggling to come to consensus about what constitutes important social studies learning. Similar struggles were seen in the development of the Virginia Standards of Learning, when there were numerous perspectives as to which events and individuals should be taught, what perspective of events should be presented, and the scope of world events to include (Van Hover et al., 2010). Thus, performance tasks in social studies need to allow students to demonstrate the skills of social studies without reliance on particular factual knowledge or interpretations of historical events given the diversity of content and perspectives to which students have been exposed (Baker, 1994).

Putting aside content to focus on the processes and skills of social studies opens another debate over the purpose of social studies education and thus the outcomes that assessments should measure. S. G. Grant et al. (2004) describe the diverse views on social studies education as ranging from "creating little historians to creating little social radicals" (p. 316). For some practitioners, social studies education is about training students in the work of historians or geographers such as writing essays, making maps, or making decisions about issues based on evidence (S. G. Grant et al., 2004). Others see social studies as providing students with the skills to utilize historical and ethical reasoning to formulate decisions, adding a moral and ethical dimension to the skills of a practitioner (Maddox & Saye, 2017). The National Council for the

Social Studies (1994) states that the "primary purpose of social studies is to help young people make informed and reasoned decisions for the public good as citizens of a culturally diverse, democratic society in an interdependent world" (p. 3). Similarly, the Kansas Board of Education committee drafting state social studies standards agreed on the need for students to be participants in civic and community affairs, with an emphasis on compromise and making decisions in uncertainty (O'Brien, 1997).

If the desired outcomes are the skills of a historian, then performance assessments should measure students' ability to critique evidence, establish historical context, and explain their own interpretations of historical events; but if the desired outcomes are citizens of an independent world, then performance assessments would have to measure completely different attitudes and behaviors. Virginia has broadened the purpose, stating that the goals of the LAAs are to "ensure all students are college, career, and future ready," have access to relevant assessment strategies that target skills, and "engage students in meaningful tasks" (VDOE, 2017, p. 1). This creates a challenge for school divisions to build assessments that meet these goals and definitions. The lack of agreement on the mission and goals of social studies education will make it difficult to construct performance assessments to measure desired outcomes for students.

Beyond defining the purpose of social studies assessments and the skills to be measured, the breadth of skills demanded in the social studies can make constructing performance assessments difficult. The social studies discipline requires students to perform a wide range of skills such as analyze graphs and data, interpret photos and cartoons, contextualize and analyze the perspective of documents, research, communicate both verbally and in writing, construct arguments based on evidence, and problem-solve (S. G. Grant et al., 2004; O'Brien, 1997; Suh & Grant, 2014). While the challenges of constructing tasks that measure higher-order skills such as

document or image analysis and analytical writing were discussed previously, social studies tasks also need to focus on the knowledge and types of analysis taught in social studies, not simply "artful writing" or the general talents of students (Baker, 1994, p. 99). The challenges of creating quality, valid, and reliable performance assessments outlined in the empirical studies are further complicated in social studies class given the range and variety of skills to be measured.

Implications for Virginia

Proponents argue that performance assessments address the shortcomings of traditional standardized testing and, when used properly, could promote the development of higher-order thinking skills and the transfer of knowledge in students. While the research shows that individual schools and teachers have had classroom success with performance assessments, state initiatives have faced concerns of validity, reliability, and authenticity of performance assessment programs (Darling-Hammond & Ancess, 1996; Khattri et al., 1998; Stecher, 2010). In this context, the VDOE hopes to gain the benefits of performance assessment to improve instruction to ensure that all students are college, career, and future ready and receive equitable educational opportunities (VDOE, 2017, 2019c). Without reliable scores and valid assessments, it will be difficult for the VDOE to assure all stakeholders that the goals of the mandate are being met and that all students across the Commonwealth are receiving equitable educational opportunities. To leverage the benefits of performance assessments while avoiding the pitfalls experienced by other state programs, the VDOE must find a way to ensure that each of the 132 school divisions are able to construct and implement quality LAAs.

Other literature focuses on defining quality performance assessments and best practices in constructing performance assessments to address the concerns around implementation and to promote the learning gains of performance assessments. The VDOE drew on the work of

SCALE, Jay McTighe, and Chris Gareis to develop an evaluation tool, the VQCT that establishes criteria for quality performance assessments (Gareis, 2017). The VQCT can be used to guide divisions in constructing performance assessments and to evaluate existing performance assessments.

Given the need for performance assessments to be embedded in the curriculum in order to reform teaching and learning and the challenges of constructing quality performance assessments, the literature emphasizes the importance of teacher professional development and appropriation of performance assessments. By employing a grassroots policy and letting local divisions develop their own performance assessments that correspond with the strengths and needs of individual divisions, Virginia policy reflects the literature's focus on loosely prescribed systems that allow greater flexibility and adaptation, promoting greater teacher appropriation (Khattri et al., 1998). The concerns in the literature about the need for greater teacher training and time to create new performance assessments are critical in the Virginia setting. Given the grassroots nature of the Virginia policy, in which local divisions are tasked with creating their own LAAs and restructuring instruction accordingly, the availability of resources to all divisions may affect the success of the policy.

Conclusion

Most of the studies reviewed focus on the effects of the performance assessment reform and the performance assessment, whether that is the effect on classroom instruction or the effect on the conclusions that can be drawn from the student scores on the assessments. However, the studies do not address the assessments themselves and the large-scale process by which schools successfully develop quality assessments. The studies focus on describing the planning for implementation and the processes schools engaged in to implement the reform and describe the

assessment development process as happening in committees or consortiums, but they do not provide many details about how the committees developed the assessments nor the quality of them (Abbott, 2016; Marion & Leather, 2015). The survey and document studies on classroom impact as well as the case studies of teacher responses focused on teacher practices after implementation of the reform as a whole, and they did not specifically address the nature and quality of the assessments.

Abbott (2016), Marion and Leather (2015), and Khattri et al. (1998) mention committees being created that involved teachers in the development process, but none of those studies give a thorough description of all the members of the committee, the roles of each member, nor the processes or steps taken by the committee to develop the assessments. There is no discussion as to how the committees went about identifying topics and tasks, what, if any, research-based best practices were utilized in the process of developing the assessments, and the extent or type of professional development provided to committee members. The only studies that described the process of task development were those of Moon et al. (2015) and Pfeifer (2002), who had performance tasks created by the research team, not by school division personnel and teachers. Thus, the studies reviewed here do not provide support nor evidence for the best practices described in the performance assessment literature being used by school divisions, nor do they provide insights for other schools engaging in similar reforms.

Only two of the studies evaluated the quality of the assessments being used. Brookhart (2015) analyzed the quality of the local assessments being used for state accountability in Nebraska, but it is difficult to draw conclusions or hypotheses based on Brookhart's work since the study does not identify nor define the type or format of the assessments being examined. The assessments could be multiple-choice, constructed response, performance assessments, or some

combination, as the Nebraska STARS requirements did not specify that assessments had to be performance tasks. Pfeifer (2002) also evaluated teacher-constructed performance assessments, but this was not the focus of the study. The study compared an experimental group who was to receive performance-based instruction to a control group who did not. The researcher's evaluation of the performance assessments was done to measure whether the experimental group had indeed been receiving performance-based instruction to ensure alignment with the research design. Since the evaluation of the assessments were not the focus of the study, the researcher devoted limited discussion to the evaluation process and findings, except to report the final evaluation score of each assessment. Thus, the research reviewed for this study provides little insight into the quality of performance assessments developed by teachers and local school divisions.

The existing research demonstrates the need to examine the individual assessments being developed more carefully. The many studies on the effects of performance assessments on classroom instruction found that teachers lack common understandings and definitions of performance assessments and deeper learning skills. These varied understandings may affect the quality of assessments across Virginia, which would complicate the VDOE (2019d) goal of "providing comparable rigor and quality across the state for performance assessments" (p. 1). In addition, the studies show that teachers are changing instruction and using more tasks in the classroom that they perceive to align with the assessment. Therefore, if teachers are modifying instruction to mimic the assessments, school divisions need to ensure the implementation of quality assessments that promote the desired learning outcomes.

The reliability studies reveal the importance of school divisions using balanced assessment plans that include a variety of performance assessments as well as other measures.

Since student scores on performance tasks can vary, divisions need to consider the number of tasks administered to provide students opportunities to demonstrate understandings and carefully consider the significance or weight given to a performance assessment. Thus, given the challenges other state performance assessment programs have faced, the goal of this study is to survey what types of LAAs school divisions are constructing and the quality of the LAAs according to the VQCT to explore how a grassroots policy to employ best practice is being implemented.

CHAPTER 3

METHODS

This research project was a descriptive study to explore the types, role, and quality of locally developed alternative assessments (LAAs) that school divisions within Virginia have developed to meet the mandates of the Virginia Department of Education (VDOE) for alternative accountability measures. Utilizing a grassroots approach to policy, the state gave each school division considerable flexibility in how to develop and select LAAs, and whether LAAs are implemented as division-wide assessments or school-based assessments (VDOE, 2014). Given the autonomy to adapt LAAs to local contexts across Virginia, this study sought to explore the different ways that school divisions have chosen to meet the requirement of replacing state-mandated, state-wide, multiple-choice tests with their own alternative assessments. The study examined the steps taken by divisions to develop the performance assessments for LAAs, analyzed the role of performance assessments in the local alternative assessment plan, and evaluated the quality of those assessments using the VQCT for Performance Assessments. To that end, the following research questions guided the study.

1. Research Question 1. How have Virginia school divisions approached the process of developing local alternative assessments to replace the removed SOL tests in US I and US II?
    i. Sub-question A: Who is responsible for creating the local alternative assessments in each division, and are the assessments division-wide or school-specific (e.g., teachers, administrators, a combined group of

teachers and administrators, or obtained from an outside publisher or consultant)?

ii. Sub-question B: What steps were taken to prepare for the process of creating the local alternative assessments (e.g., regional workshop by the state, VDOE sponsored workshops, workshops by a professional association such as SURN, VASSL, consultant or invited presenter to the division, internally-led PD)?

iii. Sub-question C: What processes were utilized in the design and development of the assessments to ensure quality performance assessments (e.g., use of templates, comparison to the Quality Criteria Tool, piloting of the assessment, review of student work samples, use of tables of specification, external expert review, or other strategies as indicated by study participants)?

2. Research Question 2. How many and what type/s of assessments have divisions selected to constitute their locally developed alternative assessments for USI and USII, and what is the role of performance assessments in their respective plans?
3. Research Question 3. To what extent do the locally developed performance assessments developed by individual school divisions in Virginia for US I and US II meet the seven quality criteria and 17 distinct sub-criteria established by the Virginia Department of Education?

Participants

Virginia is currently the only state with legislation that requires the replacement of selected state-wide, multiple-choice tests with LAAs. Therefore, the participants for this study

were drawn from the 132 school divisions in the Commonwealth of Virginia. Currently, the VDOE selects a sample of divisions for annual desk reviews and/or site visits concerning the LAAs. Divisions selected in the sample provide copies of the locally developed alternative assessments and rubrics used for scoring to the VDOE and participate in an interview process (VDOE, 2019c). Thus, while all school divisions are required to develop and implement LAAs, not all divisions have been required to submit them to the VDOE for review at this time. As a result of the grassroots nature of the policy and the relative newness of the LAAs, the researcher was concerned that school divisions would be reticent to share their LAAs for this study out of a concern of being critiqued or portrayed negatively. The difficulty of gaining access to willing participants, combined with the exploratory goal of this study to gain a general idea of the development, role, and quality of LAAs, meant this study would use a nonprobability sample (Daniel, 2012). While a nonprobability sample reduced the generalizability of the findings of this study, it met the goal of the study to provide illustrative examples of the work being done by Virginia school divisions while increasing the likelihood of finding willing participants (Daniel, 2012).

The study used a professional referral sampling method, a form of respondent-assisted sampling (Power et al., 2009). Respondent-assisted sampling methods involve asking selected participants of the study to assist with the identification and selection of other additional participants for the study (Daniel, 2012). Respondent-assisted sampling is often used with topics that are perceived as private or sensitive, which applied in this study as school divisions may have been hesitant to share their LAAs due to a fear of scrutiny (Daniel, 2012; Rothbart et al., 1982). While snowball, multiplicity, chain-referral or respondent-driven sampling are more

commonly known, these methods all identify participants who then provide names of or recruit subsequent participants (Daniel, 2012; Heckathorn, 2002; Rothbart et al., 1982; Welch, 1975).

Professional referral sampling differs from these other forms in that the first contact is not with potential participants; rather, the first contact is with intermediaries who provide professional services to identify or recommend participants for the study (Power et al., 2009). This methodology has been used in health and mental health research using professionals such as obstetric-gynecologists, pediatricians, school nurses, or substance abuse counselors to recommend participants for various studies (Power et al., 2009). This study followed the professional referral approach and contacted executive directors and professional development providers involved in the dissemination of training on locally developed performance assessments such as VDOE, Virginia Association of School Superintendents, Virginia Social Studies Leadership Consortium, Virginia Association for Supervision and Curriculum Development, and School-University Research Network. A member of each of these identified professional organizations was asked to recommend at least five school divisions to participate in the study, with the specification that the professionals feel the recommended divisions have taken a conscientious approach to developing LAAs and/or had some success in developing strong performance assessments for the LAAs. To protect the anonymity of participants and recommenders, the professionals were informed that they were not the only recommender of school divisions for the study and that the school divisions would not be told who recommended them. The rationale for this professional referral sampling method was that knowing they were respectfully recommended for their high-quality work on LAAs may make school divisions more willing to participate in the study. In addition, asking for referrals from a variety of statewide organizations and professionals in different parts of Virginia may have resulted in a greater

geographic spread of participant divisions, thus overcoming the network bias of snowball or other respondent-assisted sampling methods (Heckathorn, 2002). Although one agency was unable to participate and provide the names of divisions, six other organizations or individuals involved with, and knowledgeable about, division performance assessment development did provide lists of divisions ranging from 8–20 divisions.

Once recommendations were received from all professional recommenders, the lists were compared to identify overlap, and the number of unique divisions was identified. Although one drawback of respondent-assisted sampling can be the bias of interconnected networks of referring elements, leading to overlap or overrepresentation of certain participants, for this study the overlap was used to narrow the list of 56 provided divisions (Heckathorn, 2002). Divisions were ranked by the number of recommendations each division received and 18 divisions received two to five recommendations. My goal was to obtain a total of 10-20 divisions to participate in the study. Since this was an exploratory study of how school divisions are responding to the autonomy of creating their own assessments, a sample of at least 10 allowed a reasonable assumption of anonymity of the divisions participating while maintaining the feasibility of garnering an in-depth picture of the quality of assessments being constructed, the role of the assessments in a larger assessment plan, and the steps taken in that process (Creswell & Guetterman, 2019; Daniel, 2012).

The 18 divisions receiving two or more recommendations were contacted, the study explained, and the division was asked to participate in the study. Two divisions responded that the administrative roles or personnel had recently changed or that roles were being redefined and as a result the division would be unable to participate as individuals adapted to their new roles. Four divisions did not respond after several emails and phone messages. Twelve of the 18

divisions agreed to participate and all 12 completed an interview, but only 11 of the 12 divisions shared copies of their LAAs for the study, still meeting the minimum participant goal of the study. Thus, the sample size of the study for the interviews was 12 and the sample size of the LAAs for the evaluation process was 22, two from each of the 11 divisions that submitted LAAs.

The 12 divisions participating in the study represent the diverse geographic, size and economic characteristics of Virginia school divisions. Virginia school divisions are divided into eight superintendent regions that each represent different geographic regions of the commonwealth. The sample of this study had at least one division from seven of the eight Superintendent regions and two divisions from five of the regions, only Superintendent Region 7 was not included. In terms of size, two of the divisions in the study had less than 5000 students, while six divisions had from 10,000-30,000 students and four divisions had over 40,000 students, thus the 12 divisions in the study represent the range of different size divisions in Virginia. Economically, the sample in the study reflected a reflected as range as Per Pupil Spending of the sample ranged from $10,800 to $17,000 while across Virginia Per Pupil Spending ranges from $9965 to $22,953 with only nine of the 132 divisions spending more than $18,000 per student. Despite a sampling method that did not focus on obtaining diverse schools, the resulting sample does contain schools of different size, geography, and economic status (VDOE, 2023b).

Data Sources

I analyzed locally developed alternative assessments for US I and US II using participant interview responses, review of division LAA Plans, and evaluation of performance assessments using the VQCT.

Interviews

The staff member or teacher identified by the division as being responsible for the social studies LAAs was asked to participate in an interview concerning the development and implementation processes for the local alternative assessments (Appendix B). A table of specifications was created to ensure that the interview questions aligned with the research questions. After the interview questions were refined using the table of specifications, the questions were reviewed by two experts involved in state-level and division-level public education, and higher education, then further revised for clarity. Finally, the interview questions were piloted with school division personnel knowledgeable about the LAAs for US I and US II before a final revision and implementation with participants.

Interview Questions

The VDOE guidelines state that for the 2019-2020 school year, divisions would prepare Balanced Assessment Plans for the replaced SOL assessments that fully detail the local alternative assessment plan, indicating the types of assessments used (e.g., multiple-choice, short answer, performance assessments [namely constructed-response, stand-alone, unit-embedded, and/or project-based]). The VDOE (2019c) further specifies that while these plans may include a variety of assessment types, they must include some performance assessments. Each of these types of assessments differ in the number of intended learning outcomes measured, level of teacher support during administration, degree of student choice in expression, and duration (Wren & Gareis, 2019). Constructed response tasks, lasting only a portion of class, cover a limited number of learning outcomes with little teacher-student interaction and little student choice. Complex projects, lasting multiple weeks, measure a wide set of learning outcomes, with room for greater student choice and considerable teacher instruction, feedback and guidance for the student (Wren & Gareis, 2019). Given that the VDOE policy seeks to promote increased use

of performance assessments to promote deeper learning, identifying the extent to which divisions are implementing performance assessments within the larger assessment plan provided insight into the effectiveness of the grassroots approach to achieving the goals of the policy. As seen in the empirical literature, performance assessments measure different outcomes than selected-response tests, but no single type of assessment can provide all the data that students and schools need to improve teaching and learning. Therefore, balanced assessment systems that use a variety of assessment measures in a purposeful approach are needed (Chappuis et al., 2017; Wren & Gareis, 2019). The first interview question asked the division staff member to describe the number and role of performance assessments in the LAA plan for US I and US II. This was to provide insight into what proportion of the LAA plan is performance assessments and how the performance assessments are used within that plan.

Since the VDOE has granted school divisions autonomy in the implementation of the local alternative assessments, school divisions have approached the task in a variety of ways based on division policy and resources. In interview question one, the second follow up question asked at what level the local assessments are administered (e.g., classroom, building, division, or multi-division), revealing the extent of variation among assessments across a division and therefore across the state.

After establishing at what level assessments are administered, the second interview question focused on the processes used by the divisions or schools to develop one of the performance assessments. The first sub-question focused on who was responsible for the actual construction of the assessment. Khattri et al. (1998) argue that for performance assessments to inform instruction, teachers must be knowledgeable and appropriate the pedagogical techniques of performance assessments. Thus, identifying who developed the assessments provided insight

into teacher involvement in the process. The other sub-questions concerned the processes used in the development of the local assessment. The careful connection of tasks to intended learning outcomes demonstrated through tables of specification, the use of templates for performance assessments, piloting assessments and then reviewing student products, and external expert reviews are all means of ensuring valid, quality performance-based assessments (Brookhart, 2015; Gareis & Grant, 2015; Khattri et al. 1998; Wren & Gareis, 2019). In addition, The VDOE expectation is that each locally developed performance assessment should be evaluated using the Quality Criteria Tool and necessary modifications made (VDOE, 2019c). The responses to interview question two were analyzed to provide information about the different steps and processes each division took to develop a performance assessment for their LAAs.

Given the grassroots approach of allowing divisions to design their own LAAs, division personnel need to be prepared for the task. Many educators may lack the background to develop quality performance assessments, so the tendency of many teachers is to create assessments around favored or attractive activities that may not be quality assessments (Chappuis et al., 2017; Wiggins, 1998; Wiggins & McTighe, 2005; Wren & Gareis, 2019). The VDOE and other educational organizations in Virginia, such as VASS, have offered workshops throughout the state on performance assessments, LAAs, and the VQCT. Divisions may have had differing abilities to participate in these trainings or different levels of investment in professional development surrounding this initiative, thus divisions will be asked about the availability of professional development opportunities. Given the importance of professional development shown in the literature, interview question three asked what trainings the assessment developers were able to participate in and how available training opportunities were for the division.

Question 4 supported the evaluation of the assessments with the VQCT. Teachers, schools, school divisions, and the VDOE will use the local alternative assessments to make inferences about student learning and the meeting of the SOLs. Given the important inferences drawn from the assessments, the study sought to measure the quality of the assessments, and the degree to which school divisions can use the assessments to adequately measure student learning and progress on the SOLs, thus meeting the VDOE's stated goal for LAAs (VDOE, 2014). The VQCT Criteria 4C calls for a scoring tool being used that provides a consistent set of expectations. For a scoring tool to be applied consistently for all students, school divisions must have a method for establishing interrater reliability and consistency of student scores by any scorer. If interrater reliability and consistency are desired, then it was important to establish how many and which division personnel are involved in the scoring to ensure all are trained on the scoring rubric. To promote interrater reliability, the VDOE expects that by 2019-2020 schools would provide opportunities for cross-scoring student responses to performance assessments within and across schools (VDOE, 2019c). Given the concerns about the reliability of performance assessments addressed in Chapter 2, interview question four sought information concerning how school divisions have attempted to increase interrater reliability and consistency of student scores.

VDOE Quality Criteria Tool

The VQCT was constructed based on the work of the Stanford Center for Assessment, Learning and Equity (SCALE), the work of Jay McTighe, and the work of Chris Gareis (Gareis, 2017). The VDOE states that the VQCT is "designed to support comparability in rigor and quality across the state" in performance assessments (VDOE, 2019d, p. 1). As of the 2018-2019 school year, the VDOE expected school divisions to use the VQCT to evaluate locally developed

assessments and modify the assessments based on the results of this evaluation before implementing the assessments with students (VDOE, 2019c). Thus, the study used the same evaluation tool that school divisions are using to build and evaluate their performance assessments.

The VQCT consists of seven criteria that measure different elements of quality performance assessments, described in Chapter 1. The seven criteria are further subdivided into 17 individual scoring categories. The VQCT uses a ranking system of 0 - *No Evidence*; 1 - *Limited Evidence*; 2 - *Partial Evidence*; and 3 - *Full Evidence* for each of the 17 criteria. In addition, the rubric has a column for evaluators to indicate evidence or rationale for the ranking assigned to each element in the rubric (Appendix A). Comments made in this column can be used as part of a process to strengthen interrater reliability as well as for providing information to identify possible themes that arise from reviewers during the rating process regarding any given criterion.

Data Collection

Since this study sought to explore the development, types, and quality of assessments already developed by local school divisions, the first step was to obtain copies of one performance assessment developed as an LAA to meet the state mandate in each course, US I and US II. To identify divisions that would be more willing to share these materials with a researcher, phone calls were made to professional referral agencies and individuals that have been involved in the dissemination of training and information about the development of LAAs: VDOE, VASS, Virginia Social Studies Leadership Consortium, Virginia Association for Supervision and Curriculum Development, and SURN. These phone calls consisted of an explanation of the study, a request for the individuals to recommend at least five school divisions

in Virginia who they feel have taken a conscientious approach and/or had some success in developing strong performance assessments for the division LAAs, and a reassurance that divisions will not know who recommended them for the study (Appendix C). The responses from each call were compiled into one list to identify overlap and unique responses.

Responses were ranked based on the number of recommendations each division received and starting with the division with the most recommendations. After the list of recommended divisions was compiled, an internet search was conducted to identify the individual in each division who supervises social studies instruction, in some cases the recommending agencies provided a point of contact in the divisions they recommended. The first contact with the division was a phone call to the individual or the administrative assistant of the individual overseeing social studies instruction to explain the study, explain that their division has been highly recommended by other professionals in the state based on the division's work on LAAs, and ask if they would be willing to identify an individual in the division with the knowledge of or responsibility for developing the performance-based assessments being used as LAAs in US I and US II. Once the division personnel involved in the development and/or implementation of the LAAs agreed to participate, then a time was arranged to discuss the LAAs and respond to the interview questions, 11 of the interviews were conducted over Zoom and one over the phone based on the desires of the division personnel (Appendix D). In four divisions there was an internal approval process to conduct research in the division. For those divisions the approval form was completed and upon approval the divisions indicated who to contact for the study. If the initial call provided the name of another division employee responsible for the LAAs, a phone call was made to that individual to arrange a time to discuss the LAAs and conduct the interview. In several cases no contact was made, and a message was left, and a follow up email

was sent explaining the study and requesting a time to talk by phone about the division LAAs. This process continued until at least 12 divisions of the initial 18 with multiple recommendations agreed to share information about their LAAs for the study.

During the Zoom call, or in one case the second phone call, I again explained the purpose of the study, assured the participants of their anonymity in the study, and requested a copy of any one of the performance-based LAAs developed in each course, US I and US II, to replace the SOL tests, for a total of two performance-based assessments of the division's choice. The request was for all teacher-facing and student-facing materials, including rubrics and the balanced assessment plan, but only 11 divisions provided copies of the assessments, and while all divisions described their balanced assessment plans in the interview only two sent documents outlining their complete plan. All 11 of the divisions who submitted assessments to the study emailed electronic files of the assessments to the researcher. The interview concerning who developed the assessments and the steps taken in the development of each of the two assessments chosen for the study was conducted, and responses were recorded on a handheld audio recorder, the Zoom calls with participant faces and name identification were not recorded, and transcribed (Appendix E). After the conversation, follow-up emails were sent thanking the participants and reiterating the process for submitting the LAAs for the study (Appendix F).

For the evaluation of the quality of the assessments using the VQCT, I recruited four other social studies teachers and social studies specialists who are knowledgeable about performance assessments and the Virginia US History SOLs to review the assessments with me. All five reviewers, including myself, were trained on the VQCT and the state level of expectation for each element of the rubric to ensure that scoring matches the intent of the VDOE. The team of reviewers gathered in person and first practiced rating a set of example performance

assessments of varying types. The entire team scored an assessment against the 17 sub-criteria of the VQCT, then compared their ratings, and discussed their rationales for those ratings. This process repeated with other assessments that were not a part of the study until interrater consistency and consensus about the application of the VQCT was reached.

Once interrater agreement was achieved with the sample set of assessments, the five reviewers rated the LAAs provided by the school divisions. The copies of the LAAs provided for the study had any identifying information about the source or division removed to protect the anonymity of the divisions. The review team individually rated each LAA using the VQCT and made any notes in the notes column of the tool. The team then talked through all 22 assessments. For each assessment, each member of the team reported their ratings on each of the 17 subsections for each assessment. If the scores awarded by the reviewers did not match, reviewers discussed their rationale and reasoning. The discussion continued until consensus was reached on a score. The review team evaluated 22 assessments on 17 sub-criteria per assessment, resulting in 374 individual scores that were discussed by the team. For 172 of those initial scores, or 45.99% the review team was in complete agreement, with all five members awarding identical scores prior to the discussion. For another 97 or 25.94% of the scores four members of the team had identical scores and one member of the team was off by one. Thus for 71.93% of the scores the team was in complete, or almost complete agreement. Another 10.43%, 39 scores, three team members had the same score and the other two were off by one, but agreed with one another, such as a 1, 1, 1, 0, 0 or 3, 3, 3, 2, 2. In only 66 scores, or 17.65%, did team member scores vary by more than one as seen in Table 1.

Table 1

Initial Scorer Agreement

Scoring Pattern	No. of Scores *N* = 374	% of total scores *N* = 374
All 5 scorers in complete agreement	172	45.99%
4 scorers in agreement, 1 scorer off by 1	97	25.94%
3 scorers agree, 2 scorers off by 1 but agree with one another	39	10.43%
Scores off by more than 1	66	17.65%

Within the 17 sub-criteria the initial scores had greater rates of variation on sub-criteria 1C, 2, 5A, 5B, 5C, and 6A and more closely aligned on 1A, 1B, 3A, 3B, 4A, 4B, 4C, 6B, and 7A, 7B, and 7C as seen in Table 2. After discussing scores, in most cases a team member admitted to missing something in their reading of the assessment and on reexamination agreed that they had scored inaccurately and agreed with the rest of the team's scores. Notations were included on the final rubric of the rationale for each score.

Table 2

Initial Scorer Agreement by Sub-Criteria

Sub-criteria	All 5 scorers complete agreement (% agreement)	4 scorers in agreement, 1 scorer off by 1	3 scorers agree, 2 scorers off by 1 but agree with one another	Scorers off by more than 1
1A	12	5	5	0
1B	13	3	3	3
1C	8	9	4	1
2	6	10	2	4
3A	10	5	2	5
3B	12	7	0	3
4A	15	5	0	2
4B	12	7	0	3
4C	17	2	0	3
5A	5	7	5	5
5B	6	6	5	5
5C	5	8	2	7
6A	5	6	3	8
6B	14	4	1	3
7A	10	5	2	5
7B	11	4	2	5
7C	11	4	3	4

Data Analysis

Research Question 1

Given the grassroots nature of the policy of locally developed alternative assessments, research question one examined the processes by which Virginia school divisions have chosen to prepare for and approach developing the performance-assessments for the LAAs to replace the SOL tests. Sub-question A focused on who each division chose to be responsible for creating the local alternative assessments (e.g., teachers, administrators, a combined group of teachers and administrators, or obtained from an outside publisher or consultant). The first part of interview question one focused on the degree of uniformity across each division, whether the assessments were division-wide or school-specific. Then interview question two focused on who was responsible for developing the assessments (e.g., teachers, administrators, a combined group of teachers and administrators, or obtained from an outside publisher or consultant). A thematic data analysis of the interview responses was conducted to compare who is responsible for developing the performance assessments in each division including a discussion of the most common responses and the differences in approaches. In addition, the division policies about uniformity across the division or individual school or teacher autonomy were be described, comparing the approaches used by each division.

Sub-question B focused on the steps each division took to prepare for the process of creating the local alternative assessments. Interview question four asked if the division engaged in additional training to prepare division staff in developing performance assessments (e.g., regional workshop by the state, VDOE sponsored workshops, workshops by a professional association such as SURN, VASSL, consultant or invited presenter to the division, internally led PD). A thematic data analysis of the responses to interview question four about which trainings

division staff were able to access was conducted, including a discussion of any challenges divisions faced in accessing professional development resources and opportunities.

Sub-question C identified what processes were utilized by each division in the design and development of the assessments to ensure quality performance assessments. A descriptive analysis of the survey responses to interview question three provided by divisions was conducted to identify how many divisions employed how many processes that increase the quality of assessments, which steps were most commonly used, and which steps were least often employed.

Research Question 2

Research Question 2 examines the LAAPs developed by the divisions for US I and US II to identify the role of performance assessments within the plan. The analysis of the LAA plan focused on the format of the assessments in the plan (e.g., multiple-choice, short answer, performance assessments, namely constructed-response, stand-alone, unit-embedded, and/or project-based) and how many of each are represented in the plan. A thematic data analysis of the different formats was conducted including a summary of the pattern of most common types and formats and those formats least utilized.

In addition, the analysis looked at the role of the different assessments by examining the types of assessments (i.e., formative, diagnostic, summative) identified by the division and how each is utilized to measure student mastery and modify instruction. The format and type of assessments implemented in each division was used to provide insight into the scope and duration of the performance assessments being created as LAAs to replace the SOL test. Using the interview responses and LAAPs, if provided, a thematic data analysis of the formats of assessments being employed as well as the types and the purpose of each assessment in the LAA plan was conducted.

Research Question 3

To evaluate how well divisions are meeting the criteria of quality and meeting the VDOE's goal of equality across the state, the two performance assessments provided by each of the divisions in the sample were evaluated and scored according to the VQCT. The VQCT ranks each of the 17 subsections on a scale from 0 to 3. The individual ratings by each reviewer as well as the consensus of numeric ratings from the evaluation team was analyzed using descriptive statistics, namely mean, median, mode, standard deviation, and range to analyze the variation of scores on each of the measures. Both individual ratings and the consensus ratings were analyzed to gauge interrater reliability and to increase the validity of inferences drawn from the data. The means and standard deviations of the scores on each measure were compared to identify which measures had the greatest variation and which were more similar, as well as a description of which of the 17 measures tended to have higher scores and which of the measures tended to have lower scores across the divisions in the study. In addition, the subsection scores for each of the seven parts of the VQCT were totaled, and descriptive statistics will be run through SPSS to compare which of the seven parts had the greatest variation, the most commonality, and which of the seven sections had the higher or lower scores across the divisions. Finally, reviewer comments on the assessments were coded, categories of comments grouped together, and a thematic data analysis of the comments will be included in the analysis of the findings.

Sub-criterion 4C on the VQCT, which states that the scoring tool will be used across performance assessments in the course to provide consistent expectations to students and parents, was difficult to measure from the single performance assessment in each course. Therefore, interview responses to question four that asks about the scoring practices and procedures to build

interrater reliability and consistency were summarized. A thematic data analysis was conducted of participant responses and patterns of common practices were defined.

Table 3

Data Analysis of Research Questions

Research Question	Data Sources	Data Analyses
1: How have Virginia school divisions approached the process of developing local alternative assessments to replace the removed SOL tests in US I and US II?	Interview responses	Thematic data analysis of interview transcripts
2: How many and what type/s of assessments have divisions selected to constitute their locally developed alternative assessments for US I and US II, and what is the role of performance assessments in their respective plans?	Interview responses and copies of Balanced Assessment Plans	Thematic data analysis of interview transcripts and/or Balanced Assessment Plans
3: To what extent do the locally developed performance assessments developed by individual school divisions in Virginia for US I and USII meet the seven quality criteria and 17 distinct sub-criteria established by the Virginia Department of Education?	Ratings and descriptions of performance assessments using the VDOE Quality Criteria Tool	Descriptive statistics of all LAAs rankings on each of the seven criteria domains and the 17 sub-measures that comprise them to include mean, median, mode, range and standard deviation to analyze variance among scores. Thematic data analysis of the evaluator comments. Description of the results to identify patterns in the rankings and variations among quality criteria scores.

Note. USI = United States History I, USII = United States History II and VDOE = Virginia Department of Education

Timeline

During the summer and fall of 2021, I communicated with school divisions to obtain copies of the performance assessments and LAAPs, including all teacher-facing materials and student-facing materials that are available, and conduct interviews. During this time, I recruited and trained evaluators to review the assessments.

During the summer of 2022, the evaluation team reviewed and completed the Quality Criteria Tool for each assessment. All of the rubrics and comments were gathered. During the summer and fall 2022, the data were analyzed, and the study was completed by Winter2022.

Delimitations, Limitations, and Assumptions

Delimitations

The study focused only on Virginia school divisions who share a common state policy of replacing state-mandated multiple-choice assessments with locally developed alternative assessments. The goal of the study was to explore how different divisions have responded to this grassroots policy mandate and the degree of success each division has had in aligning to state expectations. Therefore, the study only analyzed assessments developed to meet the state guidelines. Virginia has implemented this policy in three social studies courses, but this study only focused on two: US I and US II, courses usually taught in middle school or late elementary school. Being taught to older students, the skills standards in US I and US II are more complex than in the other courses where the SOLs have been removed, better matching the goals and intent of performance assessments.

The study was limited to the state of Virginia and to social studies courses. Other states have pursued different policies for accountability during NCLB, have had different policies toward assessments (and performance assessments in particular), so the experience in Virginia

that has given rise to these particular assessments limits the generalizability of the findings to other states. This study also only evaluated those assessments developed for state accountability purposes, not alternative or performance assessments developed by teachers for their own classroom use and instruction. The intent was to examine assessments that will reflect school divisions' best examples of locally developed alternative assessments, since these are the assessments chosen for state accountability measures.

Limitations

This was an exploratory study of how school divisions have responded to a grassroots policy. As such, this study sought to describe the different processes that school divisions have chosen and were able to engage in to develop performance assessments, how divisions chose to incorporate those assessments within the larger balanced assessment plan, and the quality of assessments that divisions were able to develop. Since the goal of the study was to gain insights into how individual divisions experienced and responded to a grassroots policy approach, that created limits to both internal and external validity.

Internal validity, the ability to draw appropriate inferences from the data, was limited due to the selection process, history, and instrumentation (Creswell & Guetterman, 2019). To promote access to school division materials, the participants selected were those divisions perceived to have been more successful in the process of developing local alternative assessments. Any inferences about the correlation of the quality of division assessments to interview responses about the steps and trainings undertaken to develop those assessments might not be appropriate. The participating divisions may have had more resources or may have already been engaged in a performance assessment reform prior to the VDOE policy that contributes to higher quality performance assessments (Abbott, 2016). Besides selection of

participants, history, the lapse of time between when divisions developed their local alternative assessments and the timing of this study, also affects both the internal validity of this study (Creswell & Guetterman, 2019). Since some divisions may have developed their local alternative assessments five or more years ago when the policy was first enacted, it is possible that interviewees may not clearly remember the steps or procedures taken in the development of the assessments. The other limitation that emerged in the interview process was that since the introduction of the policy the administrator of social studies instruction has changed at least once in many of the divisions. Seven interviewees specifically stated that they had become the supervisor of social studies after the LAAs were initially developed and that they were not present at the onset of the initiative resulting in limited knowledge of the specifics of what their predecessor or predecessors had done. Thus, inferences drawn about division practice for enacting grassroots policy may be limited and thus the generalizability of the findings limited. The lapse of time from when divisions initially developed assessments to the time of this study also created an instrumentation limitation. The VDOE revised the VQCT in January 2019. Thus, several divisions explained in the interview process that they had developed their LAAs prior to the introduction of the tool and had not yet evaluated their assessments against the tool while other divisions had used the previous version of the tool in developing and evaluating their performance assessments, but this study evaluated the assessments based on the current quality criteria. Although these factors limit the drawing of inferences from the data, the goal of the study was to survey school division experiences and products as a basis for examining how divisions respond and what they are able to develop when given autonomy to implement policy.

External validity, the ability to generalize the findings of the study to other divisions or settings, was also limited by the selection of participants, the small sample size, the setting, and

history (Creswell & Guetterman, 2019). The choice to sample high-performing divisions combined with the small sample size limits the generalizability of the findings of this study to other school divisions (Creswell & Guetterman, 2019). The experiences and products of the divisions in this study may not be representative of all divisions due to a diversity of resources, leadership, or other factors. Within that small sample it is possible that setting will further limit the generalizability of the findings. Since some divisions chose to implement a division-wide common assessment while other divisions chose school-based or classroom-based assessments that created even greater variety of experience within the sample and hinder generalizability. Finally, similar to internal validity the lapse of time and possible unclear memories of interviewees may affect the accuracy of the processes followed, thus limiting the generalizability of the findings. Because this was an exploratory study to discover how divisions have approached a grassroots policy implementation and the quality of the assessments created, the feasibility of obtaining participants and conducting in-depth interviews into the experience of each participating division was prioritized. Thus, while the findings are not generalizable the study results will provide a basis for further research.

Assumptions

The first assumption I made as the researcher was that school divisions have developed performance assessments in the courses designated by the state. While divisions are required to have desk copies for VDOE review, the VDOE has adjusted the implementation timetable for these requirements, suggesting that it was possible that divisions have not yet developed assessments, but all of the participants in the study had developed LAAs. Similarly, the assumption was that divisions and individual teachers within the division are implementing the assessments and implementing them with integrity. The existence of the assessments does not

mean that students were actually engaging in the assessments or doing so in the manner outlined by the divisions.

Ethical Considerations

The VDOE has given considerable autonomy to individual divisions to replace state assessments with locally developed ones and still give students credit for meeting state standards. The greatest issue surrounding feasibility of this study was the willingness of school divisions to provide me with their assessments and LAAPs. The other concern was protecting the identity and anonymity of the divisions that provide LAAs to the study and of the individuals who recommend participants for the study. It was important to protect the anonymity of the teachers and divisions that score low on the VQCT to prevent public or VDOE scrutiny or criticism.

Since those recommending participants for the study are professional organizations that work closely with division personnel it was critical to protect their anonymity due to their close working relationships with school divisions. The names and organizations of the recommenders were not shared with participants. There are several recommenders which prevents identified school divisions from being traced to a specific recommender or recommending organization. Since the source of the recommendation was not critical to the study, when school divisions were recommended, they were added to one list that did not connect any division to a specific recommender. Finally, the full list of school divisions recommended was not revealed in the study, thus allowing for the possibility that other divisions, not contacted for the study, were recommended but just not asked to participate. Maintaining the anonymity of the recommenders prevented school divisions from feeling slighted by being left off the recommendation list and creating tension with the recommenders or their organizations.

The participating school divisions and the school division personnel responding to the interviews must also be protected to prevent school divisions from potential criticisms by stakeholders. Participating school names were kept confidential, including from the evaluation team scoring the assessments, and given pseudonyms when referred to in the study. All identifying information was removed from submitted performance assessments, balanced assessment plans and interview transcripts and replaced by a number assigned to that division by the researcher. The record of the number assigned to each division was known only to the researcher and kept in a separate location from the rest of the research materials in a secure location. I transcribed the interviews to prevent voice recognition of a participant. All data and transcripts will be stored on the hard drive, not cloud storage, of a password-protected computer and any hard copies were kept in a locked cabinet accessible only by the researcher.

I completed the College of William & Mary's Institutional Review Board approval process. After the initial email contact to obtain agreement to participate, interviewees were provided with an emailed Letter of Informed Consent (Appendix G) to sign and return. The letter informed interviewees that their participation was voluntary and that they could withdraw from the study at any time. The letter also informed interviewees of the potential risks of participating in the study which, given that divisions are recommended based on their success in this process, would be potential criticism of one or two division assessments by stakeholders. To prevent the scrutiny of division assessments I took steps to minimize this risk by protecting the confidentiality of recommenders, participating divisions, and interviewees as described above. Furthermore, all of the files of performance assessments, LAA plans, and interview transcripts will be deleted and destroyed at the completion of the study. Participants were informed that the

benefits of the study are contributing to a greater understanding of how divisions respond and the level of success they can achieve when given greater autonomy by policy makers.

In order to minimize my own bias, I was trained on the VQCT, and I had four other evaluators who were also be trained. This helped ensure that I used the VQCT as intended by VDOE and provide interrater reliability to support the rankings of each assessment.

CHAPTER 4

FINDINGS

Since 2014, the VDOE has allowed school divisions autonomy in developing LAAs to replace the SOL multiple-choice tests in US History I and US History II. Given this grassroots approach, the VDOE did not require specific structures or procedures for divisions. To gather data on how successful divisions navigated this process and met the mandate, this exploratory study used interviews followed by a rating of two division LAAs on the VQCT for Performance Assessments. The study has found that school divisions leveraged their unique resources of people, expertise, time, and external relationships to develop LAAs that meet the needs and settings of their division, teachers, and students. With each division drawing on different resources and experiences, divisions have employed different approaches to the process of developing LAAs, resulting in a variety of types of assessments and structures of assessment plans. This chapter summarizes the different processes by which divisions leveraged resources to develop their LAAs, the number and types of LAAs divisions are currently implementing within their balanced assessment plans, and finally, how a sample of those LAAs match the criteria of the VQCT.

Research Question 1: How have Virginia school divisions approached the process of developing local alternative assessments to replace the removed SOL tests in US I and US II?

Because I sought to identify the strategies for successful implementation of a grassroots policy, the participants in this study were all identified as successful in performance assessment

implementation by educational leaders in the state. Although the approaches to developing performance assessments and teacher capacity to implement them demonstrated by these divisions may not be representative of all the divisions in the state, these divisions reflect the diversity of location, size, and per-pupil spending found in divisions across Virginia. Geographically the participants in the study represent seven of the eight Superintendents' Regions in Virginia with no region represented more than twice and only the far western part of the state not included in the study, as no schools from this region received multiple recommendations from the educational agencies (see Table 4). The divisions also reflect the different-sized divisions in Virginia, with two divisions of less than 5,000 total students, six divisions with 10,000 to 30,000 students, and four divisions with greater than 40,000 students (see Table 5). Finally, the divisions reflect the economic variation within Virginia where school division spending ranged from $9,965–$22,953 total per pupil expenditures as the divisions in the study ranged from $10,500–$17,000 per pupil as seen in Table 6 (VDOE, 2023b).

Table 4

Participants by Superintendent Region

Region	No. of Divisions
1	2
2	2
3	1
4	2
5	2
6	1
7	0
8	1

Table 5

Division by Student Enrollment

No. of Students	Study		Virginia	
	No.	%	No.	%
< 5,000	2	16.67%	84	63.6%
5,000-30,000	6	50%	41	31.06%
> 40,000	4	33.3%	7	5.3%

Table 6

Division Total Per Pupil Expenditure

Total Per Pupil Expenditure	No. of Divisions in study	% of Divisions in study	No. of Divisions in Virginia	% of Divisions in Virginia
$9,9500-$11,500	5	41.67%	37	28%
$12,000-13,5000	4	33.3%	64	48.48%
> $13,500	3	25%	31	23.48%

What the divisions do have in common is they have invested time and resources in training division personnel and developing quality performance assessments early in the VDOE initiative and continue to build on those practices. The policy of replacing SOL tests with LAAs began in 2014 and these divisions all report engaging in the development of performance assessments before or during that time; however, collecting thorough data on division practice was complicated by the passage of time and the transition in leadership roles since 2014. Of the division representatives interviewed, 7 of the 12 people overseeing social studies instruction came to that role after the division began the process of developing LAAs, and two were new to the position within the last 1–2 years (see Table 7). As one leader said, “This is only my 6th year here, so I’m still somewhat new in the process.” Another leader who was new in the position said, “The supervisor before me was in the role for 3 years, and before that the person had been there for 15 years.” Thus, many of the current leaders of social studies assessments were not present for the initial introduction or trainings and could not provide details on events prior to their arrival. Other leaders’ involvement in the process of developing LAAs has been continuing since 2014 or earlier, which conveys that some of these processes took place 7–10 years ago. As

a result of being either new to the position and not knowing exactly what their predecessor (or predecessors) had done, or due to the passage of significant time, interview responses might not fully or accurately reflect all the steps and processes taken since the outset of LAA development. In addition, the interviews were conducted in the fall of 2021, following a school year where COVID protocols disrupted division initiatives; therefore, the VDOE allowed exceptions to accountability measures, and divisions had altered or paused assessment initiatives.

Table 7

Characteristics of Interview Participants

Division	Years in Role	Role in Division
A	1	Division social studies lead
B	1	Division social studies lead
C	3	Division social studies lead
D	3	Division social studies lead
E	3	Division social studies lead
F	5	Division social studies lead
G	6	Instructional Leader
H	8+	Division social studies lead
I	8+	Division social studies lead
J	8+	Division social studies lead
K	8+	Division social studies lead
L	8+	Instructional Leader

During the interviews, participants were asked to describe the process by which the LAAs are developed, including who is involved, the steps taken, and the trainings that are provided to developers. Eleven of the Virginia school divisions in this study approach the process of developing local alternative assessments to replace the removed SOL tests in US I and US II through a mix of division-wide, school-based, and teacher-selected sets of assessments developed at least in part by the teachers of the division. The divisions started the process by obtaining training from state educational organizations or educational consultants and using templates to structure their assessments. Representatives from the divisions continue to attend, conduct trainings, and maintain professional development while incorporating the VQCT to promote quality performance assessments.

Research Sub-question 1A: Who is responsible for creating the local alternative assessments, and are the assessments division-wide or school-specific?

The VDOE employed a grassroots policy allowing school divisions flexibility in the approach to LAAs. The divisions in this study present a mix in their approaches. Five divisions employ a similar grassroots approach of allowing schools or teachers to develop and choose the performance assessments to use in their classrooms, while the other seven divisions have chosen to use common, division-wide assessments. Regardless of the division approach, teachers in all but one division could be involved in developing the LAAs for state accountability.

Although seven divisions have chosen to implement common division-wide LAAs to meet the state policy, five divisions have individual schools choose the assessment for their building or individual teachers choose the assessment for their classrooms (see Table 8). Six of the 12 divisions use completely uniform assessments across the division while one division has common assessments that were used across the division. In this division, individual teachers

could choose when to implement the assessment, creating a variation in the specific content and time period covered in the assessment. These seven divisions include two smaller divisions, two mid-sized divisions, and two larger divisions; thus, the use of division-wide assessments does not correspond to division size. Two divisions have a combination of one division-wide common assessment and the rest of the LAAs are chosen by either the school-based team or each individual teacher from a division-approved menu. As one of these division leaders explains, by allowing some teacher choice, "we are honoring their autonomy, their professionalism, and their ability to come up with great things, the same way the state is giving us autonomy." The three divisions without any common division assessments take different approaches to how LAAs are chosen. One division has each school-based team develop and choose the assessments to be used at their school, creating commonality within each school but variation across the division, although teams could choose from a division bank of assessments if they wished. Two divisions, both mid-sized divisions, allow each teacher to decide what assessments to use in their classrooms as LAAs, and one of these divisions provides teachers with a division repository of assessments which teachers were not obligated to use. At the time of the study, one of these divisions was planning to move to one common performance assessment across the division with the rest remaining teacher choice. Divisions have taken different approaches in the implementation of the LAA policy, but the majority, nine out of twelve, have at least one assessment in common across the division, and a tenth division is intending to add one division-wide assessment.

Table 8

Division LAA Policy

Implementation Policy	No. of Divisions	Developers
Division-Wide Uniformity	5	Teachers
	1	Division Administrator
Division-Wide assessment	1	Teachers
Mix of common division-wide and teacher choice	2	1 division: Teachers 1 division: Division-wide assessment by administrator, others by teachers
School-Wide Uniformity	1	Teachers
Individual Teacher choice	2	Teachers

Note. LAA = Local Alternative Assessment

Ten of the divisions implemented performance assessments, division-wide or teacher-specific, created by teachers; the remaining two divisions used LAAs created by a division-level administrator. In one division, the division-level administrator has developed all the performance assessments for state accountability, and in the other division the administrator has developed the one common division-wide assessment and then school-based teams choose their second assessment from a division-approved menu of teacher-developed assessments. In the second division, while the administrator developed the assessment, a committee of teachers reviews the assessments to "tweak the language and build scaffolds" before classroom implementation, thus involving teachers in the revisions and refining of the administrator-created assessment. The other seven divisions with division-wide common assessments have teacher workgroups that develop the assessments for the division. One division leader feels that this approach of teacher-developed assessments led to "teachers having a little bit of ownership" over the process of integrating performance assessments into their instruction, and another division leader said, "I do like that [the assessment] is completely written and driven by our teachers." The remaining three

divisions do not mandate division-wide LAAs; in these divisions each individual teacher has the ability to construct and implement their own assessments rooted in the division training with the support of their PLCs or instructional coaches. One of these three divisions requires LAAs to be common within a school, but the school-based team had the autonomy to develop and implement their shared assessment. Two of these divisions, including the one with school-wide assessments, provide teachers access to a division repository of assessments. The goal of providing teachers with a bank of assessments is to allow teachers to implement the assessments as provided, adjust or edit and then implement the provided assessments, use the provided examples as models to develop their own assessments, or to not use them as all, as best fit the teachers' classrooms. The third division that allows LAAs to vary by teacher focuses on student choice with a portfolio approach, where students select the best representatives of their academic growth, including student products from teacher-created performance assessments. In all three cases, teachers receive training on performance assessments and division expectations to inform the LAAs they implemented. Whether division-wide or teacher-specific, most LAAs used for state accountability are developed by teachers in the division.

The seven divisions using teacher-created, division-wide LAAs provide avenues for teachers to voluntarily participate in the development of the division LAAs or LAA menu. Three of the smaller divisions in the study involve all or most teachers in the process, hosting professional development sessions, including both training on performance assessments and time to construct LAAs for the division. One division requires each teacher to construct a performance assessment based on the division template because the "process of creating them is beneficial" to build understanding of how performance assessments inform instructional practice. Three of the larger divisions, which faced challenges gathering all teachers in one place, used

volunteer or application processes to identify interested teachers for training and the development of LAAs for the division. One division leader felt that this accessibility to the process was important because "they developed them or could have been part of the process, they can speak to the author, have conversations about it, and this creates greater buy-in since they know the author and could have been a part of it." Eight division teacher teams worked solely as a division while four divisions chose to engage in a collaborative effort with teachers from other divisions to jointly build assessments. In these seven divisions, teachers have an opportunity to be part of the LAA development process; the division-wide LAAs that teachers' implement have been developed by their colleagues.

With Virginia's shift to a grassroots policy allowing school divisions autonomy in meeting the mandate, divisions have administered a variety of implementation plans with some common, division-wide assessments, some school-based decisions, and other teacher-specific assessments. Whatever the level of standardization across a division, 10 divisions consistently involve teachers in the development process and rely on teacher-created assessments as the LAAs.

Research Sub-question 1B: What steps were taken to prepare for the process of creating the local alternative assessments?

While divisions approached the process of developing and implementing LAAs differently, each successful division has invested in on-going professional developments from a variety of sources (see Table 9). Every division in the study has attended trainings from two or more educational organizations or consultants, giving them exposure to a variety of tools and perspectives; the divisions then employ multiple means of disseminating that training to the teachers. All the participating divisions have maintained on-going training for division staff and

teachers over multiple years, with at least three divisions engaging in performance assessment division initiatives since at least 2014, when VDOE announced the new policy for LAAs.

Table 9

Training Source by Division

Division	VDOE Training	VASS	Regional training	Collaborative group	Outside consultant
1					X (2 different)
2	X				X
3	X				X (3 different)
4		X			X
5	X		X		X (2 different)
6	X				X
7	X				X (2 different)
8				X	X (3 different)
9	X				X
10	X	X			X
11	X			X	
12		X	X		

Note. VDOE = Virginia Department of Education; VASS = Virginia Association of School Superintendents

The most common strategy employed by school divisions is bringing in outside education specialists and consultants to provide training to teachers in the division on various topics

surrounding performance assessments from models to assessment development, rubrics and scoring. Nine out of 12 divisions report they brought in consultants who specialize in performance assessments, constructing assessments, rubric construction, scoring student responses, and instructional implementation of quality performance assessments. Of those nine divisions, three focus, or had previously been trained, on Project-Based Learning (PBL) and had all teachers PBL trained while three other divisions focus on Inquiry Design Model (IDM) training for teachers. A fourth division was starting to train teachers on IDM, since the VDOE started piloting high school social studies assessments in the IDM format in 2020 (Virginia Board of Education, 2021). Three of the 12 divisions brought in two different educational consultants specializing in performance assessments at separate times. Two divisions worked with three or more outside consultants on various elements of performance assessments. For example, one division brought in a consultant for a 3-year series of trainings on instructional practices, and the following year, they brought in a different consultant to focus on performance assessment templates and development. Ten of these successful divisions approached the shift to performance assessments by investing in outside training by performance assessment experts to build teacher capacity and to provide the resources to construct quality performance assessments.

The second most common approach to prepare for LAA development and implementation has been to attend VDOE-led trainings on the policy and quality performance assessments. The VDOE held workshops in different parts of the commonwealth in 2016, 2017, twice in 2018, twice in 2019, and 2021 to assist divisions in adjusting instruction, quality performance assessments, and scoring student responses (Virginia Board of Education, 2021). Nine out of the 12 participants report that either they, their predecessor, or teachers from the division attended the VDOE workshops with one division leader replying, “Any time that [the

VDOE] offered workshops, we sent teachers." Other state educational agencies have also supported the VDOE initiative by hosting workshops on performance assessments. The Virginia Association of School Superintendents (VASS), a professional outreach organization comprised of school superintendents and business partners who promote education in Virginia, hosted performance assessment workshops that two of the divisions were able to attend, although they did not attend VDOE trainings. Thus, 11 out of the 12 schools attended a state-level education organization training. Beyond the state-level, divisions have relied on each other or more local partners, with two divisions pooling resources with neighboring school divisions to bring in consultants for regional training events on performance assessments. Two other divisions have joined with collaborative organizations such as the Virginia 3C Hub which partners teachers, museums, and academic institutions together to design historical inquiry performance assessments in the IDM. Seven of the 12 divisions in the study brought educational consultants to their divisions and attended VDOE workshops; additionally, one of the VASS divisions also brought in outside consultants.

These successful divisions engage in a variety of repeated training opportunities on performance assessments and the state initiative; all but one of the divisions attend the trainings provided by the VDOE or VASS. Most of the divisions, 8 out of 12, prepare to meet the state mandate by drawing on state educational institution offerings paired with contracting with educational consultants to provide training for their division. Every division has used more than one training source to prepare for the process of designing and implementing quality performance assessments.

Given the logistical limitations of the number of division personnel who can attend each training, combined with teacher and administrator turnover, the divisions in the study needed to

find ways to disseminate the training on performance assessments to other teachers and to continue to train new teachers in the division. Divisions in the study use a variety of strategies, such as hiring consultants for teacher workshops, divisions professional development (PD) days, the use of professional learning communities (PLCs), instructional coaches, and online learning methods (see Table 10). To disseminate these trainings, 6 out of 12 divisions committed to large-scale trainings where every teacher in the division was able to attend a training led by an outside consultant, and smaller divisions were able to send entire grade-level teams or most teachers who implemented performance assessments, to state-level trainings. Larger divisions have employed a variety of means to disseminate trainings when it is not feasible to send all teachers directly to the training. Ten of the 12 divisions used division-level professional development to train and develop teacher capacity. Six divisions have made performance assessments the focus of all professional development days. One division is already very experienced with performance assessments spending a year of professional learning on IDM, and all the divisions in the study continue to provide on-going training and support for teachers as implementation of performance assessments evolves. As one division leader explained, even after several years of performance assessment training, "work still needs to be done…it's an on-going, never-ending process," and another division planned to "tailor PD sessions where teachers will be able to group themselves based on where they feel they're still needing support." Division leaders in the study shared plans for future professional development with a variety of foci such as re-examining existing assessments for authenticity and deeper learning, using the inquiry model, training on the common rubric and the VQCT as a review tool, conducting scoring events, or offering skill-specific trainings such as teaching argumentative writing to "keep moving the needle forward."

Table 10

Division Methods of Disseminating Information

Division	Hire consultant for division-wide trainings	Division PD Days	PLCs	Instructional coaches or train the trainer	Online learning
1	X	X	X		
2		X			
3	X	X			
4		X		X	X
5		X	X	X	X
6	X	X	X		
7		X	X		
8		X	X	X	
9		X			
10	X	X		X	
11			X	X	
12	X		X		

Note. PD = Professional Development, PLCs Professional Learning Communities

Beyond the designated division PD sessions, nine divisions rely on building-level dissemination by trained teachers or team leads to coach and support other teachers in their building. Seven divisions rely on PLCs for teachers to share what they learn at training sessions as well as for team leaders to further guide a team through improving and evolving the use of performance assessments and instruction. One division uses monthly PLCs to train teachers on the performance assessment and, in subsequent meetings, “discuss what teachers are doing in

class, how the assessment works in class, and how teachers can prepare for administering the assessment." Another division leader feels PLCs are opportunities to "build [teacher] capacity on how to teach through the inquiry and how to implement the assessment." In three divisions, administrators report attending or expressing intent to attend PLC meetings provides additional support. Six divisions rely on the trainer model or instructional coaches which help designated teachers receive extensive training and are tasked to return to their buildings and train the rest of the teachers. Two divisions have also implemented online learning modules either through Canvas or a webinar format to provide greater flexibility for teachers to access the training. Like the initial trainings, 10 divisions in the study rely on multiple methods to disseminate the training to all teachers. Seven divisions use division PD supported or extended through PLCs or trained teachers leading training in their buildings. Regular PLCs meetings and division PD days allow division leaders to provide on-going supports, to continue to develop and improve performance assessments, to reflect on practice, and to introduce new approaches.

Overall, the divisions that have been successful in implementing this grassroots LAA policy have engaged in focused and on-going teacher and administrator training on performance assessment. The divisions attend workshops provided by the VDOE, and many divisions supplement the state workshops by contracting with educational consultants to provide more extensive training for their division. These divisions have structured internal division professional development initiatives to disseminate the training obtained by teacher leaders and administrators and build teacher capacity through division PD and PLCs. Not only do divisions engage in professional development at the introduction of the LAA initiative, but all of the divisions in the study continue to revisit and expand training on performance assessments as practice and policy has evolved.

Research Sub-question 1C: What processes were utilized in the design and development of the assessments to ensure quality performance assessments?

According to previous research, as summarized in Chapter 2, strategies to construct quality performance assessments include: unpacking the learning standards, outside expert review, piloting or field testing, and analyzing student work for desired student processes (Brookhart, 2015; L. Grant & Gareis, 2015; Khattri et al., 1998; Lane, 2014; Wren & Gareis, 2019). In Virginia, further guidance was provided in 2018 through the VQCT for performance assessments. When asked which of these steps the divisions follow in developing quality performance assessments, the most common responses are using templates, using teacher feedback to make revisions, and reviewing assessments against the VQCT (see Table 11).

Table 11

Strategies for Quality Performance Assessments by Division

Division	Template	VQCT	Pilot	Feedback & Revision	Unpack Standards	Student Samples	Expert reviews
1	X	X		X			
2	X	X		X			
3	X	X Simplified	X	X	X		X
4	X	X		X			
5	X			X			
6	X		X	X			
7	X	X	X	X		X	X
8	X	X		X			
9	X	X		X	X	X	
10	X	X		X	X		
11	X	X					
12	X	X	X	X	X	X	

Note. VQCT = Virginia Quality Criteria Tool

Each division in the study reports the use of a template or framework for constructing their LAAs. Five divisions have been trained in PBL while other divisions draw on the Document-Based Question (DBQ) model, which is common in social studies and is used on Advanced Placement (AP) exams. One division identified the G.R.A.S.P.S. (Goal, Role, Audience, Situation, Product, Standards) model which is a strategy for developing performance assessments. This model asks educators to first identify a real-world goal of the assessment, then

define a real-world role for the student in the task, identify the audience for the student product and the situation or context for the student task, then define what product the student will create and identify by what standards the product will be measured (Wiggins & McTighe, 2004). Five divisions that had attended the VDOE trainings or brought in consultants report they use templates from those training sessions, such as the templates developed by Chris Gareis or Jay McTighe. Seven divisions report that they are switching templates with six moving toward the IDM approach. This model incorporates an IDM template similar to the format the VDOE has been piloting in high school social studies courses since 2020 and is used by the C3 Teachers' Virginia Inquiry Collaborative. One of the divisions that does not use division-wide assessments has created its own division-level template for teachers to use. These successful divisions have all used a template in developing their LAAs to provide a structure to promote quality assessments.

The second most common strategy for constructing LAAs is the use of the VQCT, which was introduced after the divisions in the study had developed and started implementing their LAAs. With divisions retroactively incorporating the tool, nine of the divisions have incorporated the VQCT to varying extents; one division uses a modified quality tool, and two divisions have not used the tool with their existing assessments. A participant from one of the two divisions who have not yet incorporated the VQCT explains, "We had developed that task prior to the quality tool…so how do we make sure the task we've been using fits into those expectations…but we haven't had the time yet." Another division leader expresses similar concerns, feeling that the VQCT "takes a little bit of training to relate to," that "a lot of our energy went to instructional scaffolding," and that division uses a simpler criteria tool developed by an outside consultant.

The nine divisions using the VQCT vary from administrator use to more widespread teacher involvement. In two divisions, the assessments are reviewed against the (Darling-Hammond, 2017; Reed, 1993; Stecher, 2010) by division administrators only. For these two divisions, the time it takes to train teachers on the tools is a concern. One division leader stated that administrator used the tool to "ensure our pieces fit with the quality criteria, but we haven't had the larger [training] with our teacher leaders as we were waiting to see if we could get the funding"; additionally, because of the COVID pandemic, there is concern about overburdening teachers. Another division that trains teachers on the VQCT reports that division leaders have been "working with department chairs and teachers to get a better understanding of the VQCT for more than 2 years." Five divisions report they have trained teachers and administrators on the VQCT and report that teachers, as well as administrators, review existing or new assessments against the VQCT and make adjustments after assessments were constructed. The other two divisions using the tool have embedded the VQCT into the process of assessment construction. As one division leader reports, teachers and administrators use "the VQCT at the front end and back end, during the development process to inform the process [of developing performance assessments], especially with people who are not as familiar with the process to inform them of the expectations, and then at the end to vet [the performance assessments] by reviewers and use it to give feedback." Similarly, another division works with an educational consultant to build a performance assessment template "that is aligned with the Criteria tool to make sure we are looking at [the VQCT]" as teachers develop assessments, choose resources, and design the implementation. At the time of this research, 10 divisions had incorporated the VQCT into the LAA development process and five were training teachers and encouraging teachers use the tool

to evaluate all performance assessments teachers were using, not just the assessments serving as LAAs.

Feedback and revision are the last widely used strategy, but divisions vary in the frequency and processes of obtaining teacher feedback for revision of the LAAs. Four divisions have established regular means of reviewing assessments, including one division that "meets annually [with teachers] to discuss and adjust," and a second division where teachers and administrators "revise and review assessments annually." A third division uses a Google Form to gather feedback from teachers throughout the implementation window and uses that feedback to "revise over time so [teachers] see us being reflective in the process, revising the assessments based on feedback and teachers can be heard." A fourth division, which does not have division-wide assessments, requires teachers to submit their performance assessments to administrators for feedback and revisions prior to implementation; administrators review the assessments quarterly. These four divisions have created on-going, structured means of soliciting feedback that will be used to continually review and revise the assessments.

The other eight divisions have less structure and regularity but did seek teacher feedback in the process of developing performance assessments. Four divisions are implementing new performance assessments in the coming year and have plans to meet with teachers after the first use of the assessment to get feedback for revisions. A fifth division reports that assessments are "revised as the standards change." The sixth division reports that assessments have been through one revision process since development, and the seventh division has not yet revised one of their assessments but, at the time of this study, is currently in the process of soliciting teacher feedback concerning needed changes or revisions. Almost all, 11 out of 12, divisions describe a

process of gaining teacher input based on classroom experience and student responses to revise assessments since the initial implementation.

The other research-based practices for quality performance assessments reported by the divisions include unpacking the standards, piloting, professional review, and review of student samples. Only four divisions specifically report "unpacking the standards and unpacking assessments and finding out where the opportunities for richer tasks might be." It is possible, given the content of VDOE trainings or consultant trainings, that other divisions unpack the standards as part of the larger trainings divisions have attended, but they did not specify the process of unpacking the standards in the interview. Four divisions report that they will be piloting new assessments this year and intend to gather feedback and revise them at the end of the year. Three divisions use the review of student work to revise and review their performance assessments. Two divisions report that "each year they look at student work and evaluating performance assessments so that…adjustments that need to be made can be identified and revisions made." One division who has experienced the benefits of reviewing student work samples states that:

> "What really was revised more so than the assessment was the instructional practices in preparing the students for that. The big aha that came out of [student samples] was that 'Oh my god, we've got to be teaching writing more often.'"

Thus, the review process not only was used to revise the assessments but also affected how teachers viewed their practice, one of the intended outcomes of performance assessments. Finally, the least common strategy is outside review as only two divisions have outside experts review their assessments. This includes the use of museum and higher education experts to review historical accuracy, the documents used, and the questions being asked. Although all of

the divisions use templates and most used the VQCT, the use of other research-based strategies for developing quality performance assessments are less used.

While these last strategies are only used by a minority of divisions, every division in the study is using at least two strategies to promote quality performance assessments, and 10 are using 3 or more methods. In order to develop quality performance assessments, all 12 divisions in the study start with a template to structure the assessment. Following this, they review their assessments against the VQCT and use teacher feedback to revise performance assessments after the initial implementation. Nine divisions both revise LAAs based on teacher feedback and evaluate assessments against the VQCT.

The degree to which every LAA in the division is developed in accordance with these strategies depends on division practice and the degree of uniformity across the division. Divisions with common, division-wide assessments or banks of vetted division assessments could document that each of these steps are employed with every performance assessment, but divisions with more teacher-based LAAs have less ability to track the use of these strategies. Five division leaders report that some teachers administer performance assessments to "check the box, but the tasks are not rich," while other teachers in those divisions are more invested in quality performance assessments. A sixth division plans to have teachers "bring their performance tasks to [training] to look at them through the lens of is this real world, is it hitting any of the 5Cs, are they authentic." These concerns by half of the participants indicate that not all teachers are implementing steps to ensure quality performance assessments on teacher-specific assessments. To address these concerns, four divisions provide training and expect teachers to use the VQCT on their own assessments. One division requires teachers to submit performance assessments to district administration to review for quality before implementation

with students. While 10 divisions use the VQCT or something similar, and five work to train teachers to use the tool to review their own assessments, not all divisions have formal practices in place yet to ensure that all performance assessments adhere to processes that promote quality.

Divisions have implemented assessments with varying degrees of uniformity as seven divisions use common division-wide assessments, two divisions use a mix of division-wide and teacher-chosen assessments, one division uses school-wide, teacher-selected LAAs, and two divisions allow teachers to choose all of the LAAs. Despite the variation in uniformity, the divisions in this study have approached the process of developing LAAs by investing in a variety of types of training for division personnel and repeatedly disseminating that training through internal professional development and PLCs. With this training and preparation, divisions have used templates and quality review tools to develop the performance assessments and engage teachers in the process of developing and revising of the performance assessments.

Research Question 2: How many and what type/s of assessments have divisions selected to constitute their locally developed alternative assessments for USI and USII, and what is the role of performance assessments in their respective plans?

The divisions in this study selected anywhere from one to six performance assessments as the primary or, often, the only measure for state accountability in USI and USII. The assessments chosen largely consist of IDM or DBQ-style assessments, with some use of projects or other persuasive or analytical writing assignments.

The number and type of performance assessments used by the divisions in the study has evolved and changed since the policy was initiated in 2014. Prior to the LAA policy, six divisions specifically reported having division-wide multiple-choice assessments or benchmarks for data purposes in preparation for the SOL tests; the other six divisions did not describe their

prior assessment plans in the interview. When the VDOE policy shifted away from a multiple-choice SOL in 2014, three divisions began developing PBL assessments and created "one big thing" where the assessment was an "event." Since 2014, divisions have continued to attend performance assessment trainings, and many divisions have tried to move from one large assessment, like a PBL, to what one division leader described as "using all assessment types…something that could be done in a couple of days or even a day." Given these past experiences and the changing understanding of performance assessments, the LAAs and assessment plans developed by the divisions in the study reflect the continuum of assessment types including large projects, curriculum-embedded assessments, and one or several-class-period constructed responses. The different assessment types divisions in the study used require varying demands on student and teacher time, lasting from one class period to several weeks and covering different ranges of learning outcomes. The varying time demands of the different types of assessments may contribute to the varying number of assessments used by the divisions in the study (see Table 12). The divisions using mostly longer, embedded assessments such as IDMs or research projects tended to have fewer assessments than the divisions with more stand-alone assessments that focus on fewer learning outcomes and were shorter in duration. For example, the division with six common LAAs uses shorter assessments such as map analysis, document analysis, or a single DBQ-style writing prompt which can be completed within a single class period; in contrast the three to four IDMs used in three divisions require three to four class to complete the individual formative tasks and then respond to the compelling question for each IDM.

Table 12

Format of the Balanced Assessment Plan

Assessment Plan	No. of Divisions	No. of Performance Tasks in Assessment Plan	Format of performance assessments
All Performance Tasks	9	1	Research project with written product
		3	Varied Constructed responses, DBQs
		3	IDMs
		3	Scaffolded DBQs
		4	Simplified/scaffolded IDMs/DBQs
		4	Mix of projects & constructed responses/IDMs
		4	Various types as chosen by teachers, at least one integrated writing assessment
		5	Various types as chosen by teachers including writings such as brochures, journals, slide shows, etc.
		6	DBQs/IDMs/Document and map analysis
Mixed Multiple Choice and Performance Tasks	3	1	Various types by teacher
		2	1 modified/scaffolded DBQ, 1 various types by teacher
		>1	Student portfolios where students choose presentative work that can be multiple choice but at least one performance assessment

Note. DBQ = Document-Based Question, IDM = Inquiry Design Model

The division assessment plans for VDOE accountability primarily focus on performance assessments. While three divisions maintain multiple choice assessments in their assessment plans to provide data that allows teachers and administrators, as two administrators reported, to "feel secure that [teachers] taught and [students] learned the standards," the other nine division plans focus purely on performance assessments and are chosen, as one division leader explained,

to “focus on greater emphasis on skills, not covering content.” Three of the 12 divisions have assessment plans that blend multiple-choice assessments and performance assessments. Two have division-wide multiple-choice tests and then require an additional one or two performance assessments for state accountability. The third division uses a student-chosen blend of teacher-constructed multiple-choice and performance assessments. The other nine divisions do not require any multiple-choice assessments and meet the state accountability mandate solely through performance assessments; the nine division plans range from one to six assessments per course, with three divisions requiring three performance assessments and three divisions requiring four. Of those nine divisions, four specify that multiple choice tests are given by teachers but are not part of the assessment plan for state or division accountability; two others require teachers to provide evidence or document how they taught and assessed every standard but did not specify the nature of the other assessments. With the VDOE shift to performance assessments for state accountability, nine divisions have chosen to focus solely on performance assessments for their division assessment plans.

The division leaders’ description of the format of their division-wide assessments, combined with an examination of the sample assessments submitted for the study, reflects a distribution of projects, research assignments with written products, IDMs, DBQs, and other types of creative writing such as roleplaying as historical figures, but all of the LAAs require some degree of written explanation or justification. Division-wide common assessments more commonly use some form of DBQ, or the IDM model as seen in seven of the nine divisions that use at least some common assessments while the two divisions with division-wide assessments use research projects. The IDM assessments submitted to the study consist of two to four formative tasks which require students to use maps, graphs, images, and/or readings to answer

questions. The assessment then concludes with students using the experiences in the formative tasks to write an argumentative paragraph or essay on an overarching compelling question. The DBQs submitted to the study provide students with a set of documents, images, maps, and/or graphs, usually scaffolded with questions to help students analyze the resource. Once provided with these documents, the students use the sources to construct an argumentative paragraph or essay in response to a prompt. Given that six of the divisions allow teachers to develop and implement at least some of their own performance assessments, division leaders are not able to specify all the formats of assessments being used to meet the state mandate. Examples of teacher-created assessments submitted to the study include students using information gained in class, through independent research, or through reading provided documents to write letters, journals, or news articles from the perspective of someone living in the past. Other assessments have students conduct research and write brochures or other texts that include persuasive content on an issue, event, or topic. All the assessments described by participants or submitted to the study require students to provide some degree of written explanation or justification based on learned or researched material.

As a result of the varying format of the assessments, the duration of these varies from a single class period to several weeks. While IDM tasks require a specified 2-4 class periods, some of the DBQs require one to two periods for preparation and writing, and other writing assessments could be completed within a single class period. Not all the assessments submitted to the study specify the duration of the assessment and one division leader states that there is a "dichotomy of how long [teachers] spend implementing the [common] assessments." Regardless of the format, every division uses the assessments in a summative form, while those divisions

implementing IDMs use the intermediary formative tasks formatively and the final product as a summative assessment.

While all divisions use the performance assessments as summative assessments, divisions use the data from the assessments differently; some are more focused on student growth and progress while other divisions focus more on teachers' instructional practices. Six divisions specifically cite usage of the performance assessment data to "look at student growth, how their skills improved and increased over time." One division requires students to write a reflection on "their own growth and understanding," and another division states that the "biggest thing is to provide feedback to students on their growth." One of those divisions also uses the data "to inform where we were going to go next and approach the next assessment," and another division reports using the data to "create groupings of students for differentiation" or "to do any remediation or enrichment." Although the exact methods differ, all six of these divisions report using the assessments as a data source for student growth.

Five divisions, including two who also describe student-focused purposes, report using the assessments to focus on, as one division leader reported, "instruction and curriculum, on more meaningful instruction." Another division reports that they used the performance assessments "to encourage teachers to do more teaching of skills" and another leader wants the assessment to give "teachers strategies for helping students write and write and read like a historian." Still, divisions report that the effect on instruction varies by teacher. Two division leaders share that some teachers "use the data to inform instruction, but some just check the box" and administer the LAA on the assigned due date. A third division leader emphasizes that, while the goal of using the performance assessments was "to shape instruction and not just go over a list of bullets," many teachers struggle with this and "veteran teachers even are just tied to

teaching the list of bullets." A second leader agrees that it can be a "struggle to get [performance assessments] more ingrained throughout the entire instructional process as opposed to a drop in tasks." Six divisions specifically describe intentions or plans for future professional development to continue providing support for teachers on better integrating and using the performance assessments to improve instruction.

The divisions in this study have replaced the single, multiple-choice SOL test with a set of assessments, with nine divisions choosing solely performance assessments in their assessment plans and three divisions using a mix of multiple-choice and performance assessments. Divisions use from one to six performance assessments to meet the state requirements of various types, but all require a written expression of student learning. Performance assessment research argues that performance assessments can take a variety of forms, and the school divisions in this study reflect the continuum of performance assessments in the assessments they have chosen LAAs (Darling-Hammond, 2017; Reed, 1993; Stecher, 2010). The data from the assessments both measures student growth and informs instruction.

Research Question 3: To what extent do the locally developed performance assessments developed by individual school divisions in Virginia for US I and US II meet the seven quality criteria and 17 distinct sub-criteria established by the Virginia Department of Education?

To review the quality of the submitted assessments, the review team utilized the VQCT for Performance Assessments which measures assessments on seven criteria: Standards/Intended Learning Outcomes, Authenticity, Language Use, Success Criterion, Student Directions, Accessibility and Feasibility. To measure the extent to which the LAAs meet the quality criteria specified in the VQCT, I asked each division to submit two assessments, one from USI and one

from USII, to be reviewed by a team assembled by the researcher. Given that divisions use from one to six performance assessments per course and five of those divisions allow some degree of teacher-selected assessments, the sample reviewed by this study represents only a small fraction of the assessments being used as LAAs in these divisions.

The VQCT provides 17 sub-criteria that are each ranked from 0, No Evidence, to 3, Full Evidence. The members of the review team independently scored each assessment and then discussed the scores on each of the seventeen sub-criteria for those assessments. The review team evaluated 22 assessments on 17 sub-criteria per assessment, resulting in 374 individual scores that were discussed by the team. As discussed in Chapter 3, for 172 of those scores, or 45.99%, the review team was in complete agreement, with all five members awarding identical scores prior to the team discussion. For another 97 scores, or 25.94%, four members of the team had identical scores, and one member of the team was off by one. Thus, for 71.93% of the scores the team was in complete or almost complete agreement when scoring independently. For any scores where the team was not in full agreement, the team discussed the scores and their rationale until consensus was reached and a summary of the rationale for the score was agreed upon. The team was able to achieve consensus for all 374 scores and the scores used in the analysis are the consensus scores.

While all five members of the review team were in agreement for 45.99% of the scores and off by only one point by one scorer for an additional 25.94% of the scores as seen in Chapter 3, that level of agreement was not consistent across all 17 sub-criteria. Although the team was in complete or almost complete agreement (one scorer off by one point) for 71.93% of the scores, only nine of the sub-criteria met or surpassed that level of agreement (see Table 13). Sub-criteria 4A had the highest level of almost complete agreement with 90.9% followed by 4B and 4C at

86.4%. All parts of Criterion 1 and Criterion 2 as well as sub-criteria 6B were also above 72% of almost complete agreement.

Table 13

Initial Scorer Variability on Scores

Sub-criteria	Almost total agreement*	% in almost total agreement	Scorers off by more than 1	% Scorers off by more than 1
1A	17	77.3%	0	0
1B	16	72.7%	3	13.6%
1C	17	77.3%	1	4.5%
2	16	72.7%	4	18.1%
3A	15	68.2%	5	22.7%
3B	19	86.4%	3	13.6%
4A	20	90.9%	2	9.1%
4B	19	86.4%	3	13.6%
4C	19	86.4%	3	13.6%
5A	12	54.5%	5	22.7%
5B	12	54.5%	5	22.7%
5C	13	59.1%	7	31.8%
6A	11	50%	8	36.4%
6B	18	81.8%	3	13.6%
7A	15	68.2%	5	22.7%
7B	15	68.2%	5	22.7%
7C	15	68.2%	4	18.1%

* All 5 scorers complete agreement or 1 off by 1

The other 8 sub-criteria had lower levels of agreement and more disparate scores where reviewers differed from one another by more than one point. Sub-criteria 6A had the lowest level of agreement at 50% and the highest number of disparate scores at 36.4%, followed by all elements of Criterion 5 and Criterion 7 as well as sub-criteria 3A. The lower agreement arose from different interpretations of the VQCT wording and different perceptions of the review team members. The evaluation team spent time debating what constituted scaffolding for sub-criteria 6A, whether outlines, graphic organizers, or guiding questions or student choice in products which were provided to all students constituted scaffolding or did the VDOE intend for 6A to mean additional guidelines to teachers for instructional supports for individual students beyond those supports provided to all students. For 5A, 5B and 7A the review team discussed the meaning of "realistic" (7A), "accessible" (5B), and "clear" (5A) in terms of the age group intended for the task. Those reviewers with more experience with middle school students tended to score those three sub-criteria lower than review team members whose experience was primarily with older students. Similarly, 5C had different scores due to different perceptions about culturally insensitive responses. While some review team members read the assessment as intended and assumed that students would answer in the desired ways and scored 5C higher, other team members who had extensive work with middle school students quickly imagined the worst-case scenario and scored 5C lower. A similar difference in interpretation arose in 3A which asked if the assessment contained "multiple means of accessing…academic and disciplinary language" (VDOE, 2019d, p. 3). Some team members felt that this meant a variety of types of sources beyond written text to include maps, graphs, online or visual resources, while other members interpreted it as primary and secondary sources would be sufficient for full credit, as well as difference in opinion of what constitute "academic" language. Finally, 7C has low

agreement since the review team did not know how to weight the two different components in the sub-criteria. The first sentence in sub-criteria 7C asks if the assessment is implemented over multiple days is a schedule provided for that multiday implementation, but the second sentence measures whether the assessment provides information about how the assessment fits with a student's prior learning. Thus, when scoring an assessment that had a schedule of multiple days but that provided no explanation for how the assessment connected to or fit with prior learning, team members struggled to identify how to score the assessment since it only did one of the two things listed in the sub-criteria resulting in varying scores. Even on the items with high variability team members presented their perspectives and discussed how they interpreted the assessment and the VQCT until all review team members agreed on a consensus score that was used in the study.

The 17 sub-criteria are individually scored from 0-3, but the VQCT document provides no space nor indication of an intent to total the sub-criteria scores. For this study the sub-criteria scores for each assessment have been added together for an overall composite score to compare the overall quality of the assessments. A performance assessment with Full Evidence, or 3, for each of the 17 sub-criteria scores 51 points. The 22 assessments evaluated in this study have overall scores ranging from 18 to 46 out of the possible 51 points, with a mean of 29.64, median of 30.5, and mode of 31 ($\sigma = 7.65$) as seen in Table 14.

Table 14

Total Scores on the Virginia Quality Criteria Tool

Score out of 51	*f*	%	USI or USII assessment
18	1	4.5	USI
19	1	4.5	USII
20	1	4.5	USI
21	2	9.1	USI
22	1	4.5	USII
26	1	4.5	USI
27	1	4.5	USI
28	1	4.5	USI
29	1	4.5	USII
30	1	4.5	USII
31	1	4.5	USI
31	2	9.1	USII
32	2	9.1	USII
33	1	4.5	USI
36	1	4.5	USII
37	1	4.5	USII
39	1	4.5	USI
43	1	4.5	USII
46	1	4.5	USI

The USII assessments scored slightly higher than USI with a range of 19-43, but a mean of 31.09 and a median and mode of 31 ($\sigma = 6.61$). While USI had the highest single scoring assessment at 46, more of the lower scoring assessments were USI, where the scores ranged from 18-46 with a mean of 28.18, a median of 27, and a mode of 21 ($\sigma = 8.64$). The biggest gaps in the two courses, USI and USII, were in Criteria 1B, 1C, 2, 3A, 4A, 4B, and 4C. In Criterion 4C, USI means were at least .5 lower than USII; in Criteria 4A and 4B, USI means were at least .5 higher than USII.

Comparing the Seven Quality Criteria

Overall, divisions in the study scored higher on Criterion 3: Language Use; Criterion 2: Authenticity; Criterion 5: Student Directions; and Criterion 1: Standards/Intended Learning Outcomes. The scores were lower on Criterion 6: Accessibility and Criterion 7: Feasibility, with Criterion 4: Success Criterion in the middle (see Table 15). Each Criterion has from one to three sub-criteria, signifying that the maximum scores for each Criterion range from three to nine. Criterion 2 has no sub-criteria and a maximum score of 3. Criteria 3 and 6 have two sub-criteria and maximum scores of six. Criteria 1, 4, 5, and 7 each have three sub-criteria and maximum scores of nine. The percentage of possible points each mean represents is included in the data tables.

Table 15

Criterion Descriptive Statistics

Criterion	Possible Points	*M*	% of possible points	*Mdn*	Mode	*SD*	Variance
1	9	5.72	63.56%	6.00	8.00	2.37	5.64
2	3	2.00	66.67%	2.00	3.00	1.06	1.14
3	6	5.04	84%	5.00	6.00	.997	.998
4	9	4.59	51%	5.00	5.00	1.53	2.35
5	9	5.82	64.67%	5.50	4.00	1.59	2.54
6	6	2.27	37.83%	2.00	2.00	1.49	2.21
7	9	4.81	53.44%	5.00	3.00	2.65	7.01

The data in Table 16 shows the divisions have created LAAs with assessments that require students to employ higher-order, authentic skills of the social studies, and use language

skills to present historical arguments in assessments that are clearly described to students. The lower scores usually focus on logistical items such as schedules, documentation of standards, specified instructions for scaffolding and differentiation, and issues with a common generalized rubric, rather than the intellectual tasks being required of students.

Table 16

Sub-Criterion Descriptive Statistics from Highest Mean to Lowest

Sub-criterion	Short Description	*M*	*Mdn*	*SD*
3B	Use of Language	2.86	3.00	.351
7A	Feasibility: Resources Present	2.27	2.50	.827
1B	Task Goes Beyond Recall	2.18	3.00	1.10
3A	Access Academic Language	2.18	2.00	.795
5B	Directions Clear & Accessible	2.13	2.00	.774
2	Authentic, Real-world	2.00	2.00	1.07
1C	Uses Deeper Learning	2.00	2.00	1.02
4C	Rubric Consistency	2.05	3.00	1.28
5A	Directions/Resource aligned	1.80	2.00	.834
5C	Directions Bias/Sensitive	1.75	1.50	.967
6A	Accessibility: Scaffolding	1.65	2.00	.812
1A	Standards Present & Align	1.55	2.00	1.05
4A	Rubric Tightly Aligned	1.35	1.00	.745
4B	Rubric Audience-Friendly	1.20	1.00	.523
7C	Feasibility: Schedule	1.05	1.00	1.19
7B	Feasibility: Duration	1.05	1.00	1.15
6B	Accessibility: Differentiation	.600	.00	1.05

Criterion 3. Language Use for Expressing Reasoning has the highest relative mean of 5.05 out of 6, the least variance of any of the criterion of .998 out of 6, and the lowest range of 3 ($\sigma = .998$) with no 0 scores. Criterion 3 has two of the top 3 sub-criteria scores with 3B, which requires students to use various language forms to express learning, having a mean of 2.86 and median of 3 ($\sigma = .351$). Sub-criterion 3A, which requires students to use a variety of types of

sources, had a mean of 2.18 and median of 2.0 (σ = .793) as shown in Table 17. As previously discussed, every assessment submitted to the review team requires students to express their understanding in writing, whether that is an argumentative essay, a brochure, or a historical letter/journal/news article. Seventeen of the 22 assessments require students to use a variety of sources including maps, graphs, images, readings, videos, and/or websites in the process of preparing their written product.

Table 17

Criterion 3 Descriptive Statistics

Criterion	Possible Points	*M*	% of possible points	*Mdn*	Mode	*SD*	Variance
3	6	5.15	85.83%	5.50	6.00	1.04	1.082
3A	3	2.30	76.67%	2.50	3.00	.801	.642
3B	3	2.85	95%	3.00	3.00	.366	.134

For sub-criterion 3A, 50% of assessments scored "Full Evidence," and 30% scored "Partial Evidence," as the review team agreed that these assessments "used academic language in the prompts and scaffolding questions" and "required the use of multiple types of sources including images, maps, and documents" as seen in Table 18 (VDOE, 2019d). The few low scores on this measure are related to the "developmentally appropriate" part of the criterion with the review team recording both individually and in the consensus conversation that the prompts, concepts, and resources were too complex for middle school students (VDOE, 2019d). The review team agreed that five of the questions are "above grade level readiness" as they ask

students to discuss complex concepts or issues that the students lack sufficient evidence to address, or the sources were "too high for middle school" in terms of length, quantity, and/or overly complex vocabulary and syntax.

Table 18

Sub-Criterion 3A Frequencies

Score	*f*	% of total LAAs
3	9	40.9
2	8	36.4
1	5	22.7
0	0	0

Note. LAA = Local Alternative Assessment

All divisions scored either a 2 or 3 (Partial or Full Evidence) on sub-criterion 3B (see Table 19), although the review team spent considerable time discussing the meaning of 3B. Criterion 3B reads, "The performance assessment should require students to use one or more forms of language," and the second sentence states that "the performance assessment may provide access to…various forms of language media" (VDOE, 2019d). The team deliberated whether one form, written language, is sufficient to be "Full Evidence" since the criterion reads "one or more" and only stated "may allow" various forms. The team agreed that the wording in the rubric means one form of language is the minimum required to receive full credit, therefore the team decided that an argumentative essay in which students explain their thinking fully met the criterion. While the team gave full evidence to written constructed responses, 14 of the assessments evaluated allow students multiple ways to express themselves either through formative tasks that allow different means of expression, multipart products that include maps as

well as constructed responses, discussing feedback with peers, or being allowed to present their final product in a variety of ways.

Table 19

Sub-Criterion 3B Frequencies

Score	*f*	% of total LAAs
3	19	86.4
2	3	13.6
1	0	0
0	0	0

Note. LAA = Local Alternative Assessment

Criterion 2. Authenticity has the second highest relative mean of the seven criteria of 2.0 out of 3 and a low variance of 1.048m but does include the full range of scores from 0 to 3 (σ = 1.07). Almost half (40.9%) of the assessments scored a 3/3 Full Evidence, and 31.8% scored a 2/3 Partial Evidence (see Table 20 and 21).

Table 20

Criterion 2 Descriptive Statistics

Criterion	Possible Points	*M*	% of possible points	*Mdn*	Mode	*SD*	Variance
2	3	2.10	70%	2.00	3.00	1.07	1.143

Table 21

Criterion 2 Frequencies

Score	*f*	% of total LAAs
3	9	40.9
2	7	31.8
1	3	13.6
0	3	13.6

Note. LAA = Local Alternative Assessment

Criterion 2 measures the extent to which the assessment is "relevant to the real-world" or "asks students to do work authentic to the discipline" (VDOE, 2019d, p.3). Criterion 2 is defined by a statement with two bullets which were not separated by an "and" or an "or," making it unclear whether the assessment needs to meet both bullets or just one to score Full Evidence. The first bullet defines authenticity as "relevant to the real-world, students' interests, future careers, or other meaningful context" and the second bullet measures authentic as "authentic to the discipline, what adult practitioners of the discipline do" (VDOE, 2019d, p. 3). The review team debated the intention of the tool, whether the assessments must be both relevant to the real-world and authentic to the discipline to earn full evidence, or if meeting one of those two bullets is sufficient. The decision of the review team is that the assessment only needs to meet one of the two bullets, to be either "real-world" or "authentic to the discipline." However, the review team did reason that tasks authentic to the discipline of the social studies are relevant to the real-world and that the two bullets seem to express the same construct.

During the consensus discussion, the review team noted that most of the assessments, 16 of the 22, were "real world and required social studies discipline skills," such as requiring

students to use “diverse evidence to develop claims, source documents, and make an argument with evidence.” The assessments that earned a 2 instead of a 3 were given “Partial Evidence” due to the resources provided or the structure of the assessment. The review team agreed that two of the assessments scoring a 2 provide students with “limited documents that do not allow for defensible arguments.” The other five Partial Evidence assessments have instructions vague enough or prompts broad enough that the review team agreed the assessments either “could be completed without analyzing the sources and have the potential for responses to be made up,” or involved students writing narratives from the perspective of a historical figure, which the review team agreed is “not truly authentic to the job of historians.” The review team agreed that 72.7% of the assessments, if implemented properly, require students to “do work authentic to the discipline…analyzing and evaluating historical sources” and/or “are relevant to the real-world” as stated in Criterion 2 (VDOE, 2019d). The four assessments that score a 0 or 1 on this Criterion are “creative” assignments that do not require research. For example, an assessment may ask students to write a letter or journal as a historical figure but be structured in such a way that the review team agreed students “don’t need to use historical sources to complete the task.” Additional low-scoring assessments consisted of research projects structured in such a way that the team reported were “factual but do not require analysis or historical thinking,” and thus lacked authenticity to the discipline and real-world connections.

Criterion 5. Student Directions, Prompt, and Resources/Materials has the next highest mean with 5.82 out of 9 ($\sigma = 1.59$, variance = 2.54). The overall mean of Criterion 5 is affected by the disparities in the sub-criteria scores. Sub-criterion 5B, the quality of student-facing materials, scored higher with a mean of 2.14 ($\sigma = .934$). Sub-criterion 5A, the alignment of the prompt and materials, scored a 1.86 ($\sigma = .889$), and 5C, bias and sensitivity, scored 1.82 ($\sigma =$

.958) as seen in Table 22. Three-fourths of the assessments have Full or Partial Evidence of clear student-facing materials (5A) that students could proceed through and create a product, but the review team agreed those products may not align to the standards provided or could be biased or lack cultural sensitivity.

Table 22

Criterion 5 Descriptive Statistics

Criterion	Possible Points	*M*	% of possible points	*Mdn*	Mode	*SD*	Variance
5	9	5.82	64.67%	5.50	4.00	1.59	2.54
5A	3	1.86	62%	2.00	2.00	.889	.790
5B	3	2.14	71.33%	2.00	2.00	.774	.600
5C	3	1.82	60.67%	2.00	1.00	.958	.918

Sub-criterion 5A measures the alignment of the student-facing prompt, directions, and resources to the intended learning outcomes and the performance expectations being assessed. Only 27.3% of assessments scored Full Evidence, while 36.4% had partial evidence and 31.8% had Limited Evidence (see Table 23). The lower scores on this sub-criterion, like in sub-criterion 1A, are in part due to seven assessments not having intended learning outcomes provided to the review team. Without indicated standards, it was hard for the team to determine alignment, resulting in some of the 1 or 0 scores. The other factors that led to lower scores in 5A include the alignment of the resources provided to the prompt and the standards or the alignment of the formative tasks to the final summative task. On four assessments, the review team wrote that the

"documents were not adequate for the task," leaving out perspectives of groups necessary to fully answer the prompt; on five assessments, the team agreed it was "not clear from the prompt that students needed to use the documents." The instructions and structures of the assessment do not require students to critically analyze or even use the sources, but the standards listed on the assessment include analyzing sources; this has resulted in a lack of alignment between the assessment and the standards provided and, therefore, lower scores on 5A. In five multi-step assessments, the review team agreed that the "compelling question doesn't apply to the formative task"; thus, while the formative tasks might be aligned to the standards, the compelling question is not. The team also acknowledged that in three other multi-step assessments the compelling question aligned with the standards, but three assessments had formative tasks that did not align to the standards. Low scores on sub-criterion 5A resulted from a combination of factors, but primarily the lack of identified standards and loosely constructed assessments that do not ask students to engage in the required skill standards.

Table 23

Sub-Criterion 5A Frequencies

Score	*f*	% of total LAAs
3	6	27.3
2	8	36.4
1	7	31.8
0	1	4.5

Note. LAA = Local Alternative Assessment

Sub-criterion 5B measures whether the "student-facing task prompt, directions, and resources/materials are clear, complete, written in accessible language appropriate to the grade level, and organized for students in an accessible format" (VDOE, 2019d, p. 5). While 36.4% of assessments in the study demonstrate Full Evidence, 40.9% score Partial Evidence, and 31.8% score Limited Evidence. This is due to issues with clarity and age-appropriateness of the prompt, dense resources that may not be grade-level appropriate, or the lack of clarity for how students should proceed through the assessment (see Table 24). Five of the low-scoring assessments ask students to research and create products, but the review team agreed the provided instructions lack "enough structure to ensure meaningful learning" or are "unclear with broad websites that students cannot navigate on their own" in a middle school setting. For the assessments that lack structure, the review team found that the assessment's instructions state students should use primary sources without providing suggested documents, the assessment lacks instructions on how students should use the provided documents in the response, or the rubric does not measure the degree to which primary sources are used in the student response. In each case, the lack of clear instructions allows students to complete the assessment without analyzing primary documents, and therefore does not engage in a task authentic to the discipline nor meet the skill standards listed on the assessment. Other low-scoring assessments instruct students to research a topic and provide students with a URL to a large repository of information and resources, such as the National Archives site, which would be overwhelming or difficult for a middle school student to locate relevant and useful material in a manageable amount of time. The final factor leading to low scores involves multi-part assessments which require students to proceed through a set of formative tasks before reaching the summative task. The connection between the formative tasks and the summative or compelling question was unclear to the review team, and

the team agreed middle schoolers would be "unclear how to pull the steps together to answer the compelling question."

Table 24

Sub-Criterion 5B Frequencies

Score	*f*	% of total LAAs
3	8	36.4
2	9	40.9
1	5	22.7
0	0	0

Note. LAA = Local Alternative Assessment

Most of the comments about 5B from the review team are based on the prompts provided to students and concerns about the prompts being clear enough and age appropriate. For nine of the assessments, reviewers agreed prompt is "not age-appropriate," and for three others, the review team agreed that the "sources are overly lengthy and challenging for middle school students." The prompts the team agreed are not age-appropriate focus on complex, abstract principles that practitioners and experts debate extensively; middle schoolers lack sufficient understanding and thorough definitions to adequately respond to these prompts in a single class period. For five assessments, reviewers were concerned that, based on the materials and sources provided, "students do not have the information to answer the prompt," which also contributes to concerns for 5C. Despite these concerns, six assessments did get comments from the reviewers

such as “great for age” and “clear and accessible,” and 5B had the fifth highest mean of the 17 sub-criteria.

The lowest of the three sub-criteria is 5C with a mean of 1.82, median of 2, and mode of 1 ($\sigma = .958$). Sub-criterion 5C measures whether “the task prompt/directions, topic, context, and materials/resources are sensitive to the community and free of bias” (VDOE, 2019d, p. 5). Only 31.8% of assessments have Full Evidence, while 40.9% score Limited Evidence. The review team was concerned about socially charged questions, questions that lead students toward a particular response, questions that could result in culturally insensitive student products, and questions that could perpetuate misconceptions and stereotypes, especially given the lack of knowledge of middle school students (see Table 25).

Table 25

Sub-Criterion 5C Frequencies

Score	*f*	% of total LAAs
3	7	31.8
2	5	22.7
1	9	40.9
0	1	5.0

Note. LAA = Local Alternative Assessment

One concern of the review team that resulted in lower scores on the bias measurement is the number of prompts and/or resource sets that lead students toward a particular perspective or interpretation of the past. In six assessments, the structure or wording of the questions push students to a particular response; in these assessments, the question is worded in such a way that

even a student who wants to respond to the contrary would not feel that is an acceptable answer. In at least one case, the direction of the question is designed to overlook the experiences of marginalized groups in the past who would have a different perspective on the issue, thus reducing the complexity of studying history and the authenticity of the task. In other cases, the prompt is open-ended enough to answer in a variety of directions, but the resources and documents provided to the students present a particular perspective and lead students to answer the question in certain way. In these cases, other sources could be provided to give students a more nuanced assessment and be more authentic to the task of historians. Limiting or leading student responses is not authentic to the discipline and could be potentially problematic for students or parents who hold different views or perspectives.

The review team spent considerable time discussing the performance assessments that ask students to take on the role of historical figures or groups, especially marginalized groups, and write from or about those perspectives. The review team felt the sources provided lead students to write on behalf of or, in some cases, from the perspective of people from a different time and/or culture with superficial knowledge. First, most students are not members of the marginalized groups, and in the assessments viewed, only one of them provides a document created by the members of the marginalized groups. The one that did was too limited and brief to provide students with an understanding of the group's experiences or perspective. Even assessments that ask students to write from the perspective of majority groups in the past lack sufficient resources for students to gain a deep understanding of those experiences. Without sufficient information or perspectives, the review team members discussed the possibility, that students could focus on stereotypes, anomalies, or inaccuracies. As an illustrative example, when studying westward movement, students could focus on anomalies like the cannibalism of the

Donner Party or inaccuracies, such as that all wagon trains were attacked by indigenous peoples, to build a product that is not accurate of the Overland Trail experience. The team was concerned that these assessments risk producing student work that may be offensive to particular groups, or perpetuating misconceptions or stereotypes about people in the past, rather than promoting deeper learning and aligning with the standards. These concerns resulted in lower scores on 5C.

Criterion 1. Standards/Intended Learning Outcomes has the next highest relative mean of 5.73 out of 9, a mean of 6, and a mode of 8, but it has the second highest standard deviation of 2.37 and the second highest variance of 5.639. Like Criterion 5, the variation in Criterion 1 is partially due to the differences in the sub-criteria scores; sub-criterion 1A has one of the lowest means of 1.55 out of 3 ($\sigma = 1.06$), while sub-criteria 1B and 1C are two of the higher means at 2.18 ($\sigma = 1.10$) and 2.0 ($\sigma = .1.02$) out of 3, respectively (see Table 26). The three sub-criteria focus on different elements, as Criterion 1A focuses on the existence of standards and the alignment to those standards, while 1B and 1C measure the complexity of the task and possibility for deeper learning. Sub-criterion 1A focuses on the logistics of the documentation of the performance assessment, and 1B and 1C focus on the nature of the task students are asked to complete.

Table 26

Criterion 1 Descriptive Statistics

Criterion	Possible Points	*M*	% of possible points	*Mdn*	Mode	*SD*	Variance
1	9	5.73	63.67%	6.00	8.00	2.37	5.636
1A	3	1.55	51.67%	2.00	2.00	1.06	1.12
1B	3	2.18	72.67%	3.00	3.00	1.10	1.20
1C	3	2.10	70%	2.00	3.00	1.02	1.05

The lowest scoring portion, sub-criterion 1A, measures whether the Virginia SOLs are "clearly listed" as well as the "performance assessment components, resources/materials, and student products are aligned to the listed SOLs" (VDOE, 2019d, p. 2). While 68% of assessments show full or partial evidence, 27.3% of the assessments do not list the standards (see Table 27). The lack of standards could be a result of what divisions choose to share with the review team, with several divisions only providing the student-facing materials that do not include the standards. It is possible there are other internal division documents that list the standards and were not shared. Fifty-nine percent of divisions scored a 1 or 2 because the reviewers agreed that, while standards were listed, the assessments do not tightly align to those standards. For example, eight assessments list numerous parts of Standard 1, such as 1a, c, d, f, and j; but the assessment does not require students to demonstrate all of the skills listed in the assessment.

Table 27

Sub-Criterion 1A Frequencies

Score	*f*	% of total LAAs
3	3	13.6
2	12	54.5
1	1	4.5
0	6	27.3

Note. LAA = Local Alternative Assessment

The assessments evaluated perform stronger on 1B, which measures if the "assessment goes beyond recall, elicits evidence of complex student thinking" and 1C, "the performance assessments provide an opportunity for students to develop and demonstrate deeper learning competencies…life-ready competencies" (VDOE, 2019d, p. 2). Sixty-eight percent of assessments show Full or Partial Evidence on 1B, and 68% showed Full or Partial Evidence on 1C (see Table 28 and Table 29). For 1B, the review team stated that 15 of the assessments "required higher-order thinking skills such as analyzing documents, photos, and maps to make arguments." The review team agreed that for the six lower-scoring assessments (1 or 0 scores), "students could complete the task with only recall," or that the task requires students to simply "summarize facts and report content." The structure of lower-scoring assessments and the instructions provided to students does not require students to analyze the sources nor engage in higher-order thinking; rather, it would be feasible for a student to construct a response that meets the rubric solely using information from direct instruction. The review team agreed that six of these assessments are "creative, but not in ways that support the standards or would only require recall of content."

Table 28

Sub-Criterion 1B Frequencies

Score	*f*	% of total LAAs
3	13	59.1
2	2	9.1
1	5	22.7
0	2	9.1

Note. LAA = Local Alternative Assessment

Assessments that score low on 1B tend to also score low on 1C. Sixteen of the 22 assessments had the same score on 1B as on 1C, including the two assessments that score 0 on both 1B and 1C. The remaining 6 score either a 3 and a 2 or score a 1 and a 2 on 1B and 1C, with no assessment scoring more than one point different on 1B and 1C. The five lower scores on 1C (25%) are those that the review team agreed lack "rigor, students could complete the task by repeating taught material and did not need research or the use of the documents," or the product requires students to "just list content in the [curriculum framework document] or research just factual material" (see Table 29). The seven assessments that lack higher-order thinking or deeper learning competencies lower the mean for 1B and 1C, but the majority of the assessments reviewed are rigorous, and meet the criterion of authenticity to the social studies disciplines and require students to engage in deeper learning competencies and higher-order thinking. The lower overall mean for Criterion 1 is highly affected by the omission of printed standards, despite the general evidence of assessments that meet the criterion for complex student thinking and learning.

Table 29

Sub-Criterion 1C Frequencies

Score	*f*	% of total LAAs
3	9	40.9
2	6	27.3
1	5	22.7
0	2	9.1

Note. LAA = Local Alternative Assessment

Criterion 4. Success Criterion for Students scores on the lower end of the seven criteria, with a mean of 4.6 out of 9 and median and mode of 5 out of 9 ($\sigma = 1.57$). Similar to Criterion 1 and 5, the three sub-criteria of Criterion 4 have disparate scores with low means for 4A and 4B of 1.35 ($\sigma = .745$) and 1.2 ($\sigma = .523$) respectively, both median = 1, and a higher mean for 4C of 2.05 and median of 3 ($\sigma = 1.276$) as seen in Table 30.

Table 30

Criterion 4 Descriptive Statistics

Criterion	Possible Points	*M*	% of possible points	*Mdn*	Mode	*SD*	Variance
4	9	4.59	51%	5.00	5.00	1.30	1.71
4A	3	1.36	45.33%	1.00	1.00	.848	.719
4B	3	1.32	44%	1.00	1.00	.646	.419
4C	3	1.91	63.67%	3.00	3.00	1.31	1.71

In 2020, the VDOE introduced common rubrics for history and social science (Appendix H). Ten of the 12 divisions have or are actively moving towards the common rubric, which is used by 15 of the assessments. Six of those 10 divisions specifically state, "We had individual rubrics for each assessment but now we have adopted the common rubric," and one division is developing new LAAs and has "built the assessments to fit the rubric." Of the two divisions who are not using the state rubric, one has "committed early and we leaned in on teacher learning around the assessment and the rubric had not come full circle…so we did not take the rubric and retroactively apply it to tasks that already exist." The other division said they have their "own rubric that is a direct reflection of the state rubric with a couple of little adjustments." This difference in rubric usage, state rubric or local rubrics, contributes to the differences in scores on Criterion 4.

Given that Criterion 4 and the consistent use of the rubric across all assessments in the division would be difficult to evaluate when examining only one assessment, divisions were asked about the use of rubrics in the interview process. In every interview, the conversation turned to the new common rubrics from the VDOE. While one division feels that "If your assessments align to the [VQCT], then it should match up well to the rubric," and another division leader "appreciate[s] the iterative nature of the state rubric because we can use it in every inquiry," four division leaders expressed initial struggles adapting to the common rubric. One division leader commented that "the teachers are really struggling with the vagueness of the language and what does that mean in terms of this specific performance assessment." Another noted that the common rubric "is not accessible to students…it's not written in a student or parent friendly way." One division that has spent "lots of conversations on writing good rubrics…lot of time developing rubrics" responded that the state rubric "doesn't feel as good as

what we were doing before." The differing opinions and usage of the common rubric by the divisions in the study affected the scores on Criterion 4, as eight divisions have committed to the VDOE common rubric, and four divisions still use their own.

The common rubric is written to fit any performance task and is therefore broad enough to encompass a variety of performance tasks, but sub-criterion 4A states that the assessment "includes a rubric that is tightly aligned to the performance expectations of the intended learning outcomes within the performance assessment" (VDOE, 2019d, p. 4). As the review team compared the intended learning outcomes and assessment tasks to the common rubric, reviewers agreed the rubric is "too vague" and "does not align well with the performance assessment." As a result, the team gave limited evidence, with a score of 1, to the 13 assessments using the common rubric and scores of 2 and 3 to divisions that provided more task-specific, teacher or division created rubrics (see Table 31). The review team's comments that "the vagueness of the language" makes it hard to discern "what does that mean in terms of this specific performance assessment" match the concerns of some of the division leaders about the alignment of the generic rubric to project-specific intended learning outcomes.

Table 31

Sub-Criterion 4A Frequencies

Score	*f*	% of total LAAs
3	3	13.6
2	4	18.2
1	13	59.1
0	2	9.1

Note. LAA = Local Alternative Assessment

Similarly, concerns about the common rubric affect sub-criterion 4B, which states that the "scoring tool is written clearly and concisely, with audience friendly language…to provide feedback to students" (VDOE, 2019d, p. 4). When giving quality rating to assessments using the common rubric, the review team repeatedly assigned low ratings for "not audience friendly language." Additionally, the review team stated that "it would be difficult for a middle school student to understand the feedback," and "how will parents and students understand how to improve the next time?" The 13 assessments (59.1%) that used the common rubric, along with one of the division-created rubrics, scored a 1 for using language that is not student friendly and which makes it difficult for students to gain meaningful feedback (see Table 32). Divisions are working to make the rubric more accessible to teachers and students in a variety of ways. One division has "created an unpacked version of the rubric, a student-friendly version, that we use internally." While another division is providing training so that teachers "understand how to use it, what it means exactly," and another division is providing more opportunities for teachers "to work with it even more and start utilizing it more so that they get more comfortable." The struggle for teachers, division leaders, and this review team has been, as one division leader

explained, that “it’s a generalized rubric and necessarily ambiguous because it had to apply to all different tasks and we have to translate it for teachers for a specific task.” The review team’s comments echo teacher complaints that the common rubric is, as one division leader shared, “hard to make sense of and hard to share with students.” Throughout the review process, performance assessments that use the state common rubric routinely scored low on 4A and 4B for failing to be tightly aligned and not in audience-friendly language.

Table 32

Sub-Criterion 4B Frequencies

Score	*f*	% of total LAAs
3	1	4.5
2	6	27.3
1	14	63.6
0	1	4.5

Note. LAA = Local Alternative Assessment

Sub-criterion 4C states that the scoring tool used for assessment “should be used across performance assessments withing the course…to communicate a consistent set of expectations” (VDOE, 2019d, p. 4). Since the study only looked at one assessment per course, it is impossible to tell if the scoring tool is used for other assessments, unless divisions are using the VDOE common rubric. For the divisions not using the common rubric, the review team studied the language of the scoring tool and considered how easily the language could be changed to fit another assessment. For example, in one division, the two assessments submitted to the study within the two scoring rubrics are nearly identical, with only task-specific terms changed to tailor

the rubric to the assessment. The assessments using the common rubric received the highest score of 3 (54.5%) for sub-criterion 4C, which states "the scoring tool should be used across performance assessments within the course," since the rubric is designed to be used with all performance assessments (VDOE, 2019d, p. 4). The review team was frustrated by sub-criterion 4C because the reviewers state that the rubrics that are the most tightly aligned to the task and written in student-friendly language that could be used by students to improve in the future score a 0 in 4C, since the rubric is so task-specific (see Table 33). One reviewer writes, "Criterion 4C punishes good rubrics that are tied to the task simply because they are not generic enough." The tensions between being aligned to specific learning objectives and being generic enough to be applicable across different types of assessments led to disparate scores on the three sub-criteria of Criterion 4.

Table 33

Sub-Criterion 4C Frequencies

Score	*f*	% of total LAAs
3	12	54.5
2	1	4.5
1	4	18.2
0	5	22.7

Note. LAA = Local Alternative Assessment

Sub-criterion 4C also measures consistency of scoring across teachers and schools, which is not apparent as an outsider reviewing copies of the LAAs. To examine the inter-rater reliability of scoring, participants were asked in the interview to describe any training or

protocols used by the division to establish inter-rater reliability between teachers scoring the performance assessments and any opportunities for cross-scoring student responses amongst teachers in a school or across the division. One division reports holding "cross-scoring days prior and had good conversations, most teachers scores were the same or adjacent." Seven divisions are planning or beginning a cross-scoring process. One division had "pulled teachers representatives from every school for a calibration event" prior to COVID and had those teachers train "school teams to calibrate and score together." Another division has a scoring event where "we read the same ones, everyone scores and then you compare scores, talk about scores." A third division has gone through the same process on a smaller scale and had PLCs score together, "talk about a sample or two and then score from there, then they share some quality samples from the PLC with central office." Other divisions are less structured, with one division asking "teachers to give samples to other grade level teachers and discuss," and another stating "they do get together as a department and score but we don't have a formal process." Three divisions explained that they are just embarking upon or planning "upcoming PD to train department leads on a scoring protocol." Three divisions have devoted their "energy to learning the [VQCT] and making meaningful performance assessments and now we are going to start to train on the scoring." Two other divisions express concerns about the "logistics of time, how do you make the time for this heaviness of grading, and we cannot compensate teachers for that." The interview responses revealed that divisions are at different levels of experience with cross-scoring events, especially given the recent disruption of COVID. Even those divisions that have started calibration processes to ensure inter-rater reliability feel, as one division leader stated, that they "needed to get to more training," or as another leader shared, there is a need "to provide guidelines to help teachers when they go through the calibration process." Scoring calibration is

the area that more than one division leader expresses that "of all the pieces we have in place, the scoring is the piece we need to do more work on," and another concurs, stating, "We're not there yet, but it's something to grow towards." Since the documentation for a performance assessment does not describe the policies and procedures within a school division concerning inter-rater reliability and scoring calibration, the interview data on scoring events does not affect the scores on Criterion 4 from the review team. The requirement in Criterion 4 for "consistent use" was difficult for the review team to document and measure as outside observers are only viewing a limited set of assessments and are unable to determine the consistency of scoring and feedback to students.

Criterion 7. Feasibility has one of the lowest means, with 4.82 out of 9, a median of 5, and mode 3 ($\sigma = 2.65$), mostly due to a lack of evidence for sub-criteria 7B and 7C (see Table 34). Eight divisions score well on sub-criterion 7A, which had a mean of 2.27 ($\sigma = .827$), but 7B and 7C re the two lowest means of the 17 sub-criteria with both having a mean of 1.27, a median of 1, and a mode of 0 ($\sigma = 1.20$).

Table 34

Criterion 7 Descriptive Statistics

Criterion	Possible Points	*M*	% of possible points	*Mdn*	Mode	*SD*	Variance
7	9	4.82	53.56%	5.00	3.00	2.65	7.01
7A	3	2.27	75.67%	2.50	3.00	.827	.684
7B	3	1.27	42.33%	1.00	.00	1.20	1.446
7C	3	1.27	42.33%	1.00	.00	1.20	1.446

Sub-criterion 7A, which measures the inclusion and realistic nature of student-facing prompts, directions, resources/materials and scoring tools, scores high. It has the second highest mean of the 17 sub-criteria at 2.27, a median of 2.5, and a mode of 2 ($\sigma = .827$). 50% of assessments scored full evidence (see Table 35). The 27.3% scoring Partial Evidence, or a 2, had comments focused on the requirement that "performance assessments are realistic and easily accessible to teachers" (VDOE, 2019d). The review team has concerns that "directions and steps are presented, but not clear," or that with "execution unclear, what do teachers and students do with each part." These concerns could affect the feasibility of teachers properly implementing the assessment. Other concerns of the review team are that there are "too many resources" or "questionable research materials" that made the task overwhelming for teachers to implement. Like Criterion 5, these low-scoring assessments often have a vague URL to websites with large repositories of resources extending beyond the scope of the student task. No assessment was scored a 0, but the limited evidence scores (22.7%) require students to do research with vague instructions to the teacher. The review team agreed it is unclear in these cases "what research and how, where to start and how would teachers help students in the process." It is unclear whether the instructions allude to other texts being used with no instructions or elaboration, or it did not specify which texts were used. In other cases, the low score in 7A resulted from a disconnect between formative tasks and the summative compelling questions. The teacher instructions made it unclear how to help students connect the formative tasks to one another and relate them to the compelling question to construct a coherent response.

Table 35

Sub-Criterion 7A Frequencies

Score	*f*	% of total LAAs
3	11	50.0
2	6	27.3
1	5	22.7
0	0	.0

Note. LAA = Local Alternative Assessment

The majority (77%) of the assessments in the study have Full or Partial Evidence of 7A, realistic, accessible prompts, and resources; this corresponds with the scores on 1B (complex thinking), Criterion 2 (authenticity), and sub-criterion 1C (deeper learning), as seen in Table 36. The alignment of these four scores demonstrates that the divisions in the study have, for the most part, constructed quality performance assessments for students, but the documentation of the logistical details of duration (7B) and schedule (7C) were either not shared with the review team or are less fully developed than the actual task given to students.

Table 36

Alignment of Designated Criteria

Criterion	% with Full or Partial Evidence
7A	77.3%
1B	68.2%
1C	68.2%
2	72.7%

Sub-criteria 7B and 7C are missing in 36.4% of the assessments provided. This could be a result of divisions choosing not to provide this material to the review team, or it may be information that is shared with teachers by other means, such as meetings or professional development and not officially documented. Sub-criterion 7B looks for an indication of the duration of implementation and if that duration is realistic for the assessment. Although 20.7% of assessments score full evidence, the 40% that score Partial or Limited Evidence have issues of unclear or unrealistic durations (see Table 37). Six assessments reference vague durations such as “at least one day” or “several days,” which could be interpreted by teachers in different ways and result in either rushed or overly extended time frames. Other assessments score lower because the review team, several of which previously taught or supervised middle school, agreed the students “would take far longer than indicated” or “research, product creation, and writing not feasible in the time indicated.”

Table 37

Sub-Criterion 7B Frequencies

Score	*f*	% of total LAAs
3	5	36.4
2	4	22.7
1	5	18.2
0	8	22.7

Note. LAA = Local Alternative Assessment

Sub-criterion 7C measures the existence of "a schedule indicating how the performance assessment is implemented across the lessons," as well as "information about students' prior learning and how the performance assessment fits within a learning sequence" (VDOE, 2019d, p. 6). Like 7B, 36.4% of assessments do not provide this information and scored a 0. The review team agreed that 22.7% of the assessments are "well described, with an explanation of the tie to prior learning," scoring a 3, Full Evidence (see Table 38). One of the two scores Partial Evidence, "clearly identified the standards to teach before" the assessment, but the schedule is vague. The other three give a clear schedule, but the "connection to prior learning and the connection from one formative task to the other is unclear." The review team recorded that the 22.7% scoring Limited Evidence (1) "listed a series of events but not clear how to implement in class," or, alternatively, they did not "indicate tie to prior knowledge or where to fit in the learning schedule." Three of the assessments allow teachers to implement the assessment any time during the year, thus making the expectation of schedule and fit with prior learning less clear.

Table 38

Sub-Criterion 7C Frequencies

Score	*f*	% of total LAAs
3	5	22.7
2	4	18.2
1	5	22.7
0	8	36.4

Note. LAA = Local Alternative Assessment

While the duration, schedules, and connections to prior learning, 7B and 7C, may be things that experienced teachers know and do not need explicit instructions, less experienced teachers or teachers with less understanding of performance assessments may end up, as one division leader shared, feeling "okay we're at the end of the unit, time to give the assessment, but it wasn't even in the unit that…was just taught so the alternative assessment wasn't anything [the students] had just learned." Another division leader shared, "There is that bigger dichotomy of how long schools were implementing the assessment as some schools were really embedding the assessment into the unit," while others would more quickly complete the assessment "because it is the end of the nine weeks." The integrity of the performance assessments and the ability of the assessments to meet the research-based expectations to improve teaching and learning require consistent implementation as outlined in 7B and 7C.

Criterion 6. Accessibility had the lowest mean of the seven criteria at 2.27 out of 6 and a median and mode of 2 ($\sigma = 1.49$) in Table 39. Only 13.6% of the assessments score Full Evidence for 6A and 9.1% for 6B, while 63.6% provide no evidence of 6B (see Table 39). Sub-criterion 6A measures the accommodation of participation from all students and the inclusion of

directions for teachers that "identify appropriate supports or alternatives to facilitate accessibility" while sub-criterion 6B measures whether the assessment is "accessible and allows for differentiating the ways that students demonstrate their knowledge" (VDOE, 2019d, p. 5). While sub-criterion 6B specifically identifies the use of Universal Design for Learning (UDL), the wording in the Quality Criteria Tool is "such as through the application" of UDL. Since none of the assessments clearly specify or demonstrate UDL and contained wording like "such as," the review team decided that the use of UDL is not required but should be provided as an example, and it did not reduce the rating based on the absence of UDL. The review team also discussed the differences between 6A and 6B, specifically what is being asked of divisions in each sub-criterion. The review team decided that 6A focuses on the existence of tools, graphic organizers, or other teacher instructions to better support the students who may not have the skills to tackle the assessments as written. Sub-criterion 6B focuses on differentiation for students with varying abilities, such as students with learning disabilities, multilingual students, and gifted students.

Table 39

Criterion 6 Descriptive Statistics

Criterion	Possible Points	*M*	% of possible points	*Mdn*	Mode	*SD*	Variance
6	6	2.27	37.83%	2.00	2.00	1.49	2.21
6A	3	1.59	53%	1.50	1.00	.796	.634
6B	3	.682	22.73%	.000	.00	1.04	1.08

Sub-criterion 6A scores considerably higher than 6B with a mean of 1.59, which was higher than 6 of the other sub-criteria, a median of 1.5, but a mode of 1 ($\sigma = .796$) as seen in Table 16. Only 13.6% scored Full Evidence, and with the review team commenting on the existence of document analysis tools, graphic organizers to prepare written responses, and other specific instructions for teachers to support struggling students (see Table 40). The 40% of assessments with Partial Evidence provided scaffolding and supports, usually in IDM or DBQ formats. The review team agreed that, while "pre-lessons or formative tasks are well scaffolded, the summative task was not." The review team also noted that organizing argumentative writing and making the leap from analyzing evidence to argumentative writing can be difficult for students and that scaffolding is missing or limited for the final writing task. The 45.5% that received Limited Evidence are often assessments which students could complete based on learned content or pre-existing misconceptions. The review team agreed there are "no instructions on how to present to students to ensure that historical skills are utilized," or "no scaffolding to support kids approaching the task thoughtfully."

Table 40

Sub-Criterion 6A Frequencies

Score	*f*	% of total LAAs
3	3	13.6
2	8	36.4
1	10	45.5
0	1	4.5

Note. LAA = Local Alternative Assessment

Sub-criterion 6B scores the lowest of the 17 sub-criteria, with a mean of .68 and a median and mode of 0 ($\sigma = 1.04$) in Table 41. As the review team debated the meaning of 6B, a question was raised about the VQCT and whether anyone using it, teachers, division administration, or reviewers, should assume that teachers are following all Individualized Education Program (IEP) and 504 plans and differentiation is happening. The review team decided that to fit the VQCT the assessment should specify acceptable differentiations to be made for students of differing abilities; this decision by the team may have affected the scores on this sub-criterion. The review team agreed that two of the assessments do provide "effective tools for differentiation," including "supports and organizers for students with disabilities or language barriers." Both assessments are IDMs with formative tasks and compelling questions; the team agreed that those supports are better integrated in the formative tasks and missing in the summative tasks. One assessment which scores Partial Evidence has "SPED/ML considerations" (special education and multilingual) indicated on the teacher-facing materials but does not have detailed instructions. The other assessment which scores Partial Evidence has activities for the full class, such as graphic organizers, outlining, and peer editing that the team agrees would "help [special education and multilingual] students to clarify their thinking and organize response" but these activities are provided for all students, not specifically as a special education or multilingual accommodation. The three Limited Evidence assessments allow student choice in the final product, which provides the opportunity for students of different abilities to choose a format for expressing their learning that might better fit their abilities. Fourteen of the 20 assessments did not identify any specific supports or guidelines of differentiation for gifted students, students with learning disabilities, or multilingual students.

Table 41

Sub-Criterion 6B Frequencies

Score	*f*	% of total LAAs
3	2	9.1
2	3	13.6
1	3	13.6
0	14	63.6

Note. LAA = Local Alternative Assessment

Scaffolding and differentiation may also be discussed within PLCs or other trainings, but it may not be specified in the assessment documentation that was shared with the research team. Still, the steps for scaffolding and differentiation of the performance assessment taken by the teacher could affect the authenticity, deeper learning, and overall quality of the performance assessment, affecting the conclusions that can be drawn from the student products. Given that these performance assessments are being used for state accountability and to inform teaching and learning, documentation of the appropriate scaffolds and differentiation strategies need to be as fully developed as the initial performance assessments to provide consistency of implementation.

Summary of Findings

When given the autonomy to develop their own local alternative assessments to replace the multiple-choice SOL tests, 11 school divisions have involved teachers in the development process. Although nine divisions have implemented common assessments division-wide, five divisions have extended the autonomy granted to them by the state to schools and teachers, allowing building-based or teacher-based assessments, two divisions have a mix of division-wide and teacher-chosen. Eleven divisions also use teacher feedback as one of the primary strategies

for improving and achieving quality performance assessments. The widespread involvement of teachers in the development and revision process combined with the research-based goal of performance assessments to inform and improve teaching and learning correspond with the long-term, on-going professional development all of the divisions have engaged in to build teacher capacity. The divisions have invested in both outside trainings for division personnel and in internal professional development time and resources on performance assessments. Even years into the process and being recognized by professional organizations around the state for success with the LAAs, these divisions still describe themselves as improving and evolving both teacher understanding of performance assessments and the assessments themselves, with seven divisions moving to new types of performance assessments and all divisions planning for future professional development focused on furthering teacher understanding and practice.

While the division assessments and assessment plans continue to evolve, divisions have replaced the singular multiple choice SOL tests with from one to six performance assessment of a variety of types, but with 17 of the assessments submitted incorporating student use of resources to construct a written response. The performance assessments were designed and intended to engage students in authentic social studies tasks that utilize deeper-learning skills such as analyzing a variety of documents, maps, and/or graphs and charts and utilizing the resources and their knowledge to construct a written response to a prompt. Accordingly, the assessments score higher on the VQCT measures that focus on the nature of the task given to the students while the lower-scoring criteria are those that focus on teacher instructions for accessibility, accommodations, and logistical measures of timing and duration. While the performance assessments are designed to be authentic and higher-order thinking, the lack of clear, thorough teacher instructions combined with the loosely coupled system of many teachers

implementing the assessments within their individual classrooms creates the potential, in some cases, for students to complete the assessments without engaging in the intended deeper-learning and authentic skills. Similarly, incomplete instructions combine with some prompts that call for students to roleplay or speak from the position of a historic figure, raising the possibility of culturally insensitive responses.

To strengthen and improve the quality of their performance assessments, 10 divisions use the VQCT for performance assessments as well as some process of teacher feedback on their prior implementation of the LAAs. The VQCT provides guidance for divisions in the process of improving their assessments as it stresses deeper learning and authenticity in two of the seven criteria and accessing and utilizing various forms of language in the third criterion, feedback and rubrics in Criterion 4, and the logistical elements of teacher instructions in the last three criteria. For the divisions in the study the VQCT was introduced after their initial development and implementation process, but 10 divisions have incorporated the tool as means of evaluating existing and new assessments. Still, the limited documentation and explanation of the VQCT combined with varying division familiarity with the tool, creates the potential for varied interpretations of the terminology, such as authentic and deeper learning, as well as different interpretations of how to apply the seven criteria to assessments. Both school divisions and the review committee have deliberated the intent behind the wording of multiple bullets in three of the criteria.

This exploratory study of divisions developing local alternative assessments for state accountability finds that successful divisions respond to this autonomy by investing in professional development, drawing on state-provided resources such as VDOE trainings and the VQCT, and continuing to expand division capacity around performance assessments. For all the

divisions in the study, the LAAs and performance assessments in general, continue to be a work in progress as divisions revise their LAAs, provide on-going and expanded professional development to teachers surrounding performance assessments, and incorporate new and evolving VDOE resources such as the VQCT and the common rubrics.

CHAPTER 5

RECOMMENDATIONS

In this study, I explored how successful divisions have responded to the autonomy given them by the VDOE in developing their own local alternative assessments for state accountability requirements and to identify the types and quality of assessments that the divisions have developed as a result of those efforts. The divisions in this study involve teachers in the performance assessment initiative, provide on-going professional development and use research-based strategies to develop quality performance assessments, including the use of the Virginia Quality Criteria Tool for Performance Assessments (VQCT), consisting of tasks that intend to engage students in authentic, higher-order, deeper-learning competencies. This chapter begins with a discussion of how these findings correspond with the educational literature on performance assessments. The goal of this study was to identify effective strategies other divisions could leverage to further their own success and strategies that the VDOE could use as the initiative expands to provide support for divisions across Virginia. Therefore, this chapter provides recommendations for division and VDOE practice and conclude with recommendations for future research.

Summary and Discussion of Major Findings

In 2014 the VDOE chose to move from a more tightly coupled accountability system of state-created and state-scored multiple-choice tests to a more loosely coupled system where individual school divisions determined the number and types of assessments they would implement and were responsible for scoring them (Fusarelli, 2002). While other states such as

Maryland, Kentucky, and Washington had previously implemented performance assessment reforms, those states had followed more state-mandated, tightly coupled approaches where the state created the performance assessments as well as the scoring and success criteria. Though the Virginia General Assembly and VDOE mandated the shift from multiple-choice, end-of-course tests to local alternative assessments to include performance assessments, the VDOE was more grassroots in its implementation of the policy in that initially divisions were given complete autonomy in approaching the performance assessment reform. While the VDOE continued to set the content and skill standards for student learning, in 2014 divisions were given the freedom to develop their own assessments and assessment plans to replace the SOL tests. In 2016, after divisions had 2 years to adjust to performance assessments, the VDOE began offering annual (and then biannual) workshops on performance assessments. A year later, in 2017, the VDOE introduced a quality criteria tool that was finalized in 2019 as the VQCT and at that time the VDOE stated the expectation that divisions use the tool to review the quality of their performance assessments. The Virginia performance assessment policy has evolved over time but began with a very loosely coupled system where divisions had autonomy over their assessments and the only direct oversight by the state was the requirement for divisions to certify that the assessments had been administered.

This grassroots approach in Virginia differs from other state performance assessment reforms, such as Washington state. The Washington Assessment of Student Learning (WASL) began in 1992 with the state creating a commission to develop standards and assessments (Stecher et al., 2000). The state then introduced the standards for four subjects in 1995 and four additional subjects in 1996. Starting in 1997, divisions could volunteer to take the assessments newly developed by the state commission for the first four subjects. The state commission

viewed student products from the voluntary administration of the assessment and set standards of what level of performance would demonstrate student achievement of the standards (Stecher et al., 2000). This process continued for the other subjects over a 10-year period. In this more tightly coupled system, the state determined the assessments that were taken then by all students across the state and the state set common levels of performance that all students were held to. Comparing this to Virginia, the implication of Virginia's more loosely coupled system with a more grassroots evolution from division-developed assessments and assessment plans and scoring standards to increasing guidelines from the state over time may be a more uneven implementation of the policy across Virginia than in a more top-down implementation (Fusarelli, 2002). Thus, I explored how divisions approached adjusting to a large-scale policy change when given such autonomy, what assessment plans they developed and the quality of the assessments they developed for state accountability.

Discussion of Research Question 1 Findings: The Process of Developing LAAs

The divisions in this study approached the development of LAAs by involving teachers in developing and revising the performance assessments and engaging in ongoing training and professional development on performance assessments for teachers and division leaders. Ten out of 12 school divisions in this study use teacher-constructed LAAs, where teachers either construct their own assessments or serve on a committee of teachers that developed division-wide assessments. The 11th division uses one division-wide, administrator-developed assessment and one teacher-selected assessment; however, teachers were involved in revising the division-wide assessment before implementation, providing them the ability to shape the selection of resources used, the wording of the task, and the building of scaffolds. Of those 11 divisions, six allowed teachers to choose at least some of the performance assessments implemented in their

classrooms, either from a division-vetted menu or from individual teacher-developed performance assessments. One division leader specifically stated that the goal was to involve teachers because "they developed them or could have been part of the process, they can speak to the author, have conversations about it, and this creates greater buy-in since they know the author and could have been a part of it." The involvement of teachers in the development process corresponds with the research by Khattri et al. (1998), who argue teachers only appropriate performance assessments when teachers are involved in the design of the assessments. In addition, providing teachers with some degree of choice and more loosely prescribed assessment systems allows teachers to adapt the assessments for their own classrooms and better integrate them into instructional practice which can increase teacher efficacy (Khattri et al., 1998; Weick, 1976). The experience of these successful divisions, supported by the existing research, demonstrated the research-based importance of engaging teachers in the process of assessment creation and development.

When given the autonomy to develop their own performance assessments for state accountability, all but one division out of 12 involved teachers in the process of developing the LAAs. Half of the divisions in the study grant a similar level of autonomy to their teachers allowing teachers to choose and develop their own assessments. While these 12 divisions are a small, unrepresentative sample, their experiences may be indicative of larger patterns of practice across Virginia. The distributed nature of LAA development in these divisions, with the involvement of numerous teachers in the process, requires administration and teachers to have a common understanding of performance assessments and a common set of structures and strategies for developing quality performance assessments.

The Role of PD. All the divisions in the study prepared teachers and division staff for the process of developing and using performance assessments through repeated, ongoing training from a variety of sources. Training was then disseminated and supported throughout the division by professional development, PLCs, and instructional trainers.

The shift from a single, multiple-choice SOL test to LAAs created a need for divisions to train teachers on the purpose and structure of performance assessments, as well as their role in instruction. One division leader stated, "Sometimes there's confusion on what exactly [teachers] have to do [with performance assessments], that can be uneasy for some." Another division leader reported that "it's hard; even veteran teachers are just tied to teaching the list of bullets." These sentiments from division leaders are supported by the literature. Empirical studies of the implementation of performance assessments have shown that teachers and administrators have diverse understandings and definitions of performance assessments and constructs such higher-order thinking that may be partial, superficial, or flawed (Goldberg & Roswell, 2000; Koretz Mitchell, et al., 1996; Parke & Lane, 2008; Parke et al., 2006; Stecher & Mitchell, 1995). Khattri et al. (1995) and Darling-Hammond and Ancess (1996) argue that due to a lack of training, teachers lack clear understandings of complex skill constructs and what constitutes an acceptable demonstration of student understanding of those skills. As a result of the underdeveloped understanding of the performance assessments and skills, coupled with a lack of time and resources to develop tasks, research shows teachers often devise performance tasks that are hands-on without depth, or the tasks were simply inserted into the course schedule without affecting teaching and learning (Baron, 1996; Goldberg & Roswell, 2000). These concerns from the literature are shared by one division leader who said, "We have some teachers who are doing a task or tasks and they are checking that box, but I don't believe that the tasks are rich" and

another who said, “We need to look at the assessments through the lens of is this real world and how do you know?” The division leader comments about the lack of teacher understanding confirms the literature findings that many teachers do not have clear conceptions of what constitutes a performance assessment or higher-order thinking as well as the role of performance assessments in the larger instructional environment (Goldberg & Roswell, 2000; Koretz, Mitchell et al., 1996; Parke & Lane, 2008; Parke et al., 2006; Stecher & Mitchell, 1995). To successfully implement the VDOE policy of LAAs for state accountability, the division leaders in this study stressed the importance of focused, on-going professional development surrounding performance assessments.

To address the need for teacher training, most divisions prepared for the development and implementation of LAAs by attending VDOE training and hiring outside educational consultants. Nine out of 12 divisions attended the VDOE-led training that is offered annually, and two others attended training provided by the state-level school Superintendent organization, VASS. In addition to state-level training, nine of the 12 divisions contracted with at least one outside educational consultant to provide training on professional assessments. Seven divisions reported that they had the entire division go through PBL training, IDM training, or a long-term process of rotating all teachers through instructional strategies training for performance assessments. This level of training requires these divisions to dedicate a considerable amount of time and resources to performance assessments, hiring consultants, and devoting the entire division professional development calendar for a year or several years to a focus on performance assessments. Most divisions in the study invested in a variety of training opportunities, including state-provided training and consultant-led training, to prepare teachers for performance assessments; this is supported by the literature that argued for performance assessment to

positively affect instruction quality sufficient professional development must take place (Khattri et al., 1998; O'Brien, 1997; Stosich et al., 2018). While these successful divisions have devoted the time and resources to hire consultants and provide professional time for teachers on performance assessments, not all divisions in Virginia may be able to or may not choose to invest these same resources. For example, one division leader reported that their division was not a member of VASS and thus unable to use those resources. In a loosely coupled system with divisions seeking their own professional development from a variety of sources, to ensure the movement of all divisions in a common direction the VDOE needs to clearly articulate their goals (Weick, 1982). This underscores the need for the VDOE to continue and possibly expand its training on performance assessments in order to ensure adequate access to consistently messaged performance assessment training to all divisions and teachers in the state and for all 132 divisions to participate in these trainings. The active role of the state in providing training is supported by the research of Bandalos (2004) in Nebraska and O'Brien (1997) in Kansas. Both researchers argue that clear communication from state-level policy-making bodies was critical to reduce teacher discontent and frustration, and thus promote the success of the state initiative. Even the divisions in this study admitted that they still need additional training and resources for teachers despite their current levels of success. Thus, the VDOE might need to continue to invest in these trainings as the performance assessment initiative continues and expands to other courses and disciplines to support the continued growth of successful divisions and provided needed support to other divisions. In addition, encouraging and ensuring that all divisions can and are accessing these VDOE-led trainings should be a focus of the VDOE to ensure equitable and comparable implementation of the policy and thus the equitable measurement of student progress on the standards across Virginia.

Beyond attending training outside of the division, division leaders report planned or existing on-going training, typically taking place throughout the year at regular PLC meetings or division professional development days. Ten of the 12 divisions reported using division-wide professional development to build teacher capacity on performance assessments, and one of the remaining two divisions was small enough for all teachers to attend consultant training or VDOE training. This means 11 of the 12 divisions provided means for all teachers to access training formats, enabling consistent messaging to all teachers in the division. The division-led emphasis on a performance assessment initiative corresponds with the literature, which argues that teachers are reluctant to invest the time required to implement quality performance assessments when they feel pressured to meet a variety of division demands and initiatives (Moon et. al., 2005; O'Brien, 1997). Six divisions had mapped out PLC meeting topics and division-wide professional development days for the upcoming year to systematically engage in the development, implementation, and revision of performance assessments, as well as scaffolding instruction through those skills and tasks. As one division leader reported, work on performance assessments is an "ongoing, never-ending process." This strategy of training throughout the year was supported by the work of Wiggins (1998) and Khattri et al. (1998) who argued that teachers need time throughout the year for planning, collaboration, and making adjustments. The literature and experience of the divisions in the study demonstrate a need for on-going opportunities for teacher development, distributed throughout the school year to provide supports for teachers as they engage with performance assessments.

The successful school divisions were finding ways to expand the scope and reach of their training beyond division-wide professional development. Seven of the 12 divisions used PLCs, and an additional three schools used instructional coaches for building-level training and support

for teachers when implementing performance assessments. One division leader reported using instructional coaches to provide individualized support for teachers "to fill in any gaps in teacher understanding." Two other divisions discussed the challenges of increasing building-level administration's understanding of performance assessments to support teachers in the classroom. One division leader described the challenge of building administrator understanding in order to allow teachers to move away from multiple choice to performance assessments and still meet division evaluation criteria. Another division leader was planning monthly meetings for building administrators on performance assessments with the argument that trained administrators could support and increase the effectiveness of teacher use of performance assessments. The long-term, continued investment in teacher training on performance assessments seen in the divisions in this study aligned with the literature which argued that divisions working with teachers over spans of at least three years were more successful in engaging in performance assessment reform and constructing better quality assessments (Bandalos, 2004; Brookhart, 2005; Marion & Leather, 2015). Both the existing literature and the experiences of these divisions demonstrate the successful implementation of a performance assessment initiative requires deliberate, long-term planning by division administrators to identify the needed training and to map out cohesive plans for providing this support in a meaningful and on-going basis.

Previous research and the divisions in this study demonstrate that on-going training, support and designated time to collaborate on performance assessments throughout the year is necessary for the success of performance assessments. The successful divisions invested in outside training from both the state and performance assessment specialists, then planned and structured internal professional development programs to establish and increase the capacity of all teachers around performance assessments. These divisions demonstrated that a performance

assessment initiative cannot be quickly introduced, embraced, and integrated; these divisions have been investing in extensive training for 6–7 years and still reported teachers needing additional supports, teachers not fully integrating performance assessments in instruction, and other areas of needed growth and improvement. If this is the experience in more successful divisions and recognizing that all divisions may not have the resources to invest similarly in professional development, the challenge for the VDOE in this loosely coupled approach is how to implement performance assessments evenly across the state and ensure equitable, quality assessments for all students (Fusarelli, 2002). As the VDOE continues and expands the performance assessment initiative, the state and individual divisions need to recognize and commit to a long-term, concerted focus on professional development of teachers and administrators for the initiative to improve teaching and learning (Bandalos, 2004; Brookhart, 2015; Khattri et al., 1998; Marion & Leather, 2015; Wiggins, 1998).

Strategies for Quality Performance Assessments. The VDOE initially chose a grassroots approach where the development of assessments and assessment plans were left, in the words of the VDOE, "to the discretion of the school division" rather than the more tightly coupled system seen in states like Washington where the state created the assessments (VDOE, 2014, p.3). While the VDOE later offered professional development and added the VQCT, the evolution of the policy allowed localized adaptations to the policy with more self-determination by teachers and divisions (Weick, 1976). As seen in the interviews, this process of assessment development was led by division leaders who were themselves still gaining an understanding of performance assessments. As one leader said, "This is only my 6th year here, so I'm still somewhat new in the process," five other leaders had been in their role for 3 years or less, and several leaders shared that teachers were still working to embrace performance assessments years

into the process. A grassroots approach where the development, implementation and scoring of performance assessments rely on local leaders who are still developing capacity on performance assessments, may contribute to uneven implementation of the policy across the state (Fusarelli, 2002). Besides varying levels of understanding by local leaders and teachers, in Virginia's loosely coupled system with individual divisions and teachers developing assessments, those assessments are also not undergoing the same process of item development seen in other high-stakes accountability measures such as WASL or AP exam questions. Washington state hired a committee to construct assessment items and then conducted a data driven pilot before full implementation (Stecher, 2000). AP exam questions, such as the DBQ on history exams, are constructed through a five-step process beginning with setting the exam specifications, such as the length, and clear identification of the student learning the assessment is to elicit. A committee of content experts and assessment specialists then write prompts/questions which are then reviewed by another committee for alignment with the course framework, content accuracy and other criteria. Finally, a mini pilot is conducted to test the questions and then revisions are made, and the exam assembled (College Board, 2023). In a grassroots approach local divisions do not have the experience nor the resources to employ similar processes in performance assessment development as seen in large-scale top-down assessment initiatives, thus this study sought to explore the strategies divisions used to meet the VDOE goal of quality performance assessments.

Beyond the training of teachers, the divisions in the study employ a variety of research-based strategies to ensure the development of quality LAAs to replace the SOL tests. Each division identifies three or more strategies in use, with the most commonly used strategies being the use of templates, the VQCT to evaluate assessments, and teacher feedback to inform revision. While divisions report using several means of promoting quality LAAs, the distributed

nature of the development of the assessments across numerous teachers combined with the time required by these strategies may have affected how consistently these quality measures were fully implemented.

Structures for Promoting Quality Performance Assessments. Every division reports using a template in the process of developing the performance assessments, whether that was IDM, PBL, the DBQ format, or a division-created template. Since performance assessments require careful planning, the structure of a template assists developers in attending to all components of a performance assessment and promotes aligned, quality assessments that increase student achievement (Center for Collaborative Education, 2013). Thus, the division practice and the literature support the use of a template which provides a structure that helps teachers and administrators focus on student learning goals and then guide developers through a process of performance assessment development that focuses on instruction and assessing student understanding of the intended skills and learning outcomes (Center for Collaborative Education, 2013; Wiggins & McTighe, 2005).

The other common strategy, used by 10 of the 12 divisions, is to review the assessment using the VQCT. The use of a set of criteria to develop and review assessments is supported by performance assessment researchers such as Jay McTighe, Chris Gareis, and SCALE, all of whom had developed quality criteria for performance assessments that influenced the VQCT (Gareis, 2017; McTighe, 2016; SCALE, 2014). The use of the VQCT, like the use of a template, requires developers to focus on the many facets of quality performance assessments including the nature of the task, the alignment of the task to learning outcomes, and technical elements of schedules and procedures. The 10 divisions using the VQCT have invested in years of training on quality performance assessments; however, several divisions still find understanding and

using the tool to be challenging. As a result, division leaders report different levels of integration of the tool. Two divisions reported only administrators used the tool, while a third division was building teacher understanding of the tool. Five divisions have trained teachers to use the tool to review a completed assessment, and two divisions have teachers using the tool at the beginning of assessment development, throughout the process, and again as a review tool at the end.

The consistent use of the VQCT across divisions and across teachers within a division is further hampered by the differing definitions of constructs such as higher-order learning, authentic, or real-world used in the tool and what constitutes student demonstration of these skills (Darling-Hammond & Ancess, 1996; Goldberg & Roswell, 2000; Khattri et al., 1995). The VQCT does not have documentation to describe and define these constructs or the intent of the 17 sub-criteria, nor to provide clear definitions of the rankings of little, partial or full evidence. Although 10 divisions are using the VQCT there is a possibility that divisions may evaluate the same performance assessment differently depending on their interpretation of the tool and the wording of the 17 sub-criteria. With this variation of understanding and use of the VQCT, combined with the fact that six of the divisions allow at least some teacher-developed and chosen assessments for LAAs, all of the assessments used for state accountability may not deliberately or accurately evaluated against the VQCT. When teachers have the level of autonomy in selecting assessments seen in six of the divisions in this study, lack of a shared understanding and application of the VQCT across all teachers and staff make it difficult to ensure that the use of quality criteria is being implemented with integrity. While the VDOE requires school divisions to use the VQCT to promote comparability in the quality of LAAs, the familiarity with and consistent use of the VQCT as a quality check is an area of potential for growth for divisions across Virginia.

Another structure to ensure quality performance assessments is analyzing the alignment of the performance assessment to the intended learning outcomes and the standards. The literature on creating quality performance assessments and the VQCT stresses the importance of identifying desired student learning outcomes, then constructing tasks that require students to demonstrate those student responses (Brookhart, 2015; L. Grant & Gareis, 2015; Lane, 2014; Stosich et al., 2018; Wiggins & McTighe, 2005). Only four divisions specifically reported they unpacked the standards in the process of developing their performance assessments. Although unpacking these standards may be one step in the larger training by the VDOE or outside consultants that divisions attended, it is not specifically listed by division leaders as a separate or distinct step during the interview process. Yet the clear identification of the standards to be measured was the first step for both WASL and College Board with AP exams (College Board, 2023; Stecher, 2000). In addition, the fact that most of the divisions use a template and a set of criteria, like the VQCT, to evaluate their LAAs, implies an unpacking of the standards to ensure alignment in sub-criteria 1A and 5A. However, six of the assessments still omitted standards, indicating that a greater emphasis on the unpacking of the standards when developing LAAs may contribute to greater alignment between the intended learning outcomes and the quality of the resulting task.

These successful divisions are using several research-based structures to promote quality performance assessments, but in a loosely coupled system with individual teachers developing assessments there may not be uniformity and regularity in the use of these strategies. One division leader explained the challenges in achieving their performance assessment goals, explaining they were “a fairly small office” with limited staff to provide training. Another division explained that they had no division-level position specifically dedicated to social studies

to facilitate the performance assessment initiative. Divisions in the study strove to use research-based strategies, but six specifically report issues of time shortages and/or limited staffing that prevented them from being able to fully implement these strategies. These findings align with Goldberg and Roswell (2000) whose research asserts teachers were overwhelmed by the time and resources required to adequately devise performance assessments. Given that half of the divisions allowed some teacher selection of the LAAs, it is possible that while division leaders may promote the use of research-based strategies for developing quality performance assessments, individual teachers may not be sufficiently trained to implement these strategies consistently on all of the LAAs being used for state accountability. To ensure comparable rigor and quality of performance assessments across the state, more support should be provided to teachers and divisions on common understandings of the seven criteria and 17 criteria on the VQCT; additionally, divisions need to provide sufficient training on their templates for teachers and administrators to uniformly and appropriately use them for meaningful teaching and learning.

The Role of Feedback and Revision. Once a performance assessment is developed additional reviews of the assessment should take place before using the assessment to evaluate student progress. The existing literature recommends the use of outside experts, a data-driven pilot, and review of student products to evaluate the alignment of the assessment to the intended content and skills and to identify potential biases in the assessment. (Lane, 2014; Moon et al., 2005; Wren & Gareis, 2019). Only two divisions in the study reported utilizing outside expert review; one division reported having the division assessments being reviewed by an educational consultant and another division reported having both educational consultants and local historians and museums review the assessments for historical accuracy. Although more tightly coupled

programs such as WASL and AP exams used review by content and assessment experts, this process would require divisions to have the resources and access to experts, which might not be feasible for all 132 divisions in Virginia. Another form of feedback recommended by the literature is the piloting of performance assessments and the gathering of student work samples to analyze whether the assessment will produce the desired outcomes by students before widespread implementation (Brookhart, 2015; Khattri et al., 1998; Lane, 2014; Pecheone & Kahl, 2014). The field testing or piloting should include gathering data on the length of time it took students to complete the assessment, the appropriateness of the vocabulary, the fairness of the prompt and materials, and the feasibility of the assessment (L. Grant & Gareis, 2015). While four divisions reported field testing their assessments, none described such extensive data-gathering or analytical processes as the research suggests. Extensive data-collection pilots, the deliberate review of student work, and expert reviews would require considerable time and additional resources; the lack of these steps aligns with the research asserting teachers and school leaders were overwhelmed by the demands of creating and field-testing complex performance assessments (Goldberg & Roswell, 2000). Since the option for outside reviews and piloting of performance assessments, as the literature supports, are strategies that many divisions lack the resources to engage in, it is even more important that developers critically analyze the performance assessments that they develop for alignment and bias.

Although divisions might not have used the research-supported strategies of extensive data collection pilots or expert reviews, all but one division reports that they gathered teacher feedback after implementing LAAs and used that feedback to revise and make adjustments before the next school year or assessment cycle. Some divisions reported they have conducted annual reviews and have set structures for the revision process, such as a Google Form to collect

feedback or annual professional development sessions focused on revising the assessments. Other divisions reported less frequent and less structured processes such as periodic solicitation of teacher opinions or revising the assessment when the standards change. The soliciting and incorporation of teacher feedback in the revision of the LAAs, while not as data-driven as a true pilot, does align with the research focused on increasing teacher involvement in the development process to promote teacher appropriation of performance assessments (Khattri et al., 1998). Given the lack of formal pilots, teacher feedback may provide data concerning clarity, feasibility and duration, which can improve the quality of division LAAs.

Divisions have been employing templates, teacher feedback, and increasingly using the VQCT tool to develop quality LAAs. Even with best intentions, however, divisions face several challenges in ensuring that the LAAs meet the criteria for quality performance assessments. The loosely coupled, distributed nature of assessment development across divisions and across teachers within divisions presents challenges in ensuring that all developers have common understandings of the constructs of quality. Additionally, in six of the 12 divisions, at least some, if not all, LAAs were teacher-created and varied across the division. Although division-wide assessments might have undergone the processes described by the division leaders, this did not mean that every assessment developed and used by teachers is adhering to the research-based processes that ensure quality, nor does it mean that each individual teacher is sufficiently trained in employing these strategies with integrity. Thus, the implementation of the VDOE performance assessment policy may be uneven (Fusarelli, 2002). Finally, the heavy reliance on teachers to develop LAAs on top of other classroom responsibilities, combined with the limited social studies administrative staff, can make the use of time and resource-heavy research-based strategies for quality performance assessments overwhelming for divisions (Goldberg &

Roswell, 2000). The challenges in developing quality LAAs in a loosely coupled system support the need for extensive and on-going professional development that includes training on the constructs and strategies for developing quality performance assessments and the need for the VDOE to continue to provide training to divisions to continue to move 132 divisions toward a common direction on performance assessments and the VQCT (Weick, 1982).

Discussion of Research Question 2 Findings: Number and Types of Performance Assessments

When the Code of Virginia removed state-developed, multiple-choice SOL tests in five courses, the VDOE chose to replace the tightly aligned state accountability system with a more loosely coupled approach where "each school shall annually certify that it has provided instruction and administered an alternative assessment" (VDOE, 2014, p.1). The use of the singular noun clearly indicates a single alternative assessment, but then the subsequent text of the code reads school divisions "will incorporate options for age-appropriate authentic assessments," the plural noun implies the use of more than one performance assessment, creating an apparent contradiction within the code (VDOE, 2014, p.1). Thus, the initial instruction to school divisions in 2014 provided no clear guidelines on the number of assessments to be used, implying one or more would be sufficient means of assessing student learning and allowing divisions to define and develop their own assessments. In 2019 the VDOE added to the LAA policy an expectation that divisions would develop Balanced Assessment Plans "to include a variety of assessment types including performance assessments," again using the plural noun to suggest more than one (VDOE, 2019c, p. 1). Unlike the more tightly coupled WASL where the state created the assessments the language of the Virginia Code and the VDOE guidelines granted divisions considerable autonomy in interpreting the meaning of the policy and deciding both the number and types of assessments to be used to meet the reform. As a result, division Balanced

Assessment Plans consist of varying numbers of assessments which encompass a range of performance assessment types, but most divisions have chosen to focus entirely on performance assessments being used in a summative nature.

The Balanced Assessment Plans. Given the ambiguous language of the Virginia code and the VDOE with the use of both singular and plural nouns, divisions reported using anywhere from one to six performance assessments in their Balanced Assessment Plan, with the most common responses being three or four assessments. These limited numbers of assessments being used by divisions contradicts the literature which showed that the number and type of assessments affect the conclusions about student learning that can be drawn from the LAAs. McBee and Barnes (1998) in a study of four performance tasks, Gao et al. (1994) in a study of five tasks, and Stecher and Klein (1997) in a study with two tasks and another study with four tasks all found that similar tasks do not produce consistent student scores. Webb et al. (2000) in a study of two tasks found that student scores vary by occasion as well. All four research teams concluded that the variability of student scores on performance tasks prevents score generalizability. This research concludes that multiple performance tasks are required to produce reliable, generalizable data on student performance. The divisions in this study were using one to six assessments, numbers comparable to the studies in the research, and thus, according to the research, the numbers were too low to provide reliable data on student progress. Given that the LAAs were to replace the SOL tests to demonstrate student achievement and competency in social studies content and skills on the Virginia SOLs, more performance assessments would be needed to provide reliable, generalizable scores. Unfortunately, many of the assessments given by the divisions in the study take two to four class periods for the IDMs and DBQs, and even longer for research projects, which makes adding more performance assessments challenging

within the available instructional time. The challenges of using a sufficient number of performance assessments aligns with the extant literature which acknowledged that the length of complexity in performance assessments makes it difficult to construct large enough numbers of performance assessments to establish reliable scores on student achievement (Stecher & Klein, 1997). The limitations of time and the ambiguity of the VDOE policy regarding LAAs may have contributed to divisions not employing a sufficient number of performance assessments to draw reliable, generalizable conclusions about student progress. The reliance of divisions and the VDOE on these limited assessments to ensure student academic progress for state accountability underscores the importance of the LAAs being high-quality performance assessments.

The lack of generalizable scores on assessments being used for state accountability is further complicated as nine divisions used only performance assessments in their Balanced Assessment Plans; the other three divisions used a combination of multiple choice and performance assessments. The divisions' reliance on performance assessments mirrors the language of the Virginia Code and the VDOE clearly emphasize alternative, performance, and authentic assessments; however, the VDOE statement on Balanced Assessment Plans specifically stated an expectation of "a variety of assessment types" (VDOE, 2019c, p. 1). The literature supports the varied assessments used by the three divisions and recommended by the VDOE as educational researchers agree that no single type of assessment can provide all the data that students and schools need; therefore, a purposeful selection of a variety of assessment types should be employed to measure student learning (Chappuis et al., 2017; Haney & Maddaus, 1989; Wren & Gareis, 2019). In addition, the reliance of the majority of the divisions solely on performance assessments for state accountability contradicts the work of McBee and Barnes (1994) who warned that the use of performance assessments in high stakes situations is not

appropriate. Part of the disconnect between the VDOE balanced assessment policy and division practices of solely performance assessments may arise from the delay between the implementation of the removal of the SOLs in 2014 and the announcement of the Balanced Assessment Plan policy in 2019. The divisions in this study moved to performance assessments in 2014 when the SOLs were initially removed, and the emphasis was placed on performance or authentic assessments. Given the lack of a clearly articulated policy and expectation by the VDOE in this loosely coupled system, the intent of the performance assessment reform was implemented unevenly with divisions meeting the state's stated expectation of one or more performance assessments, but not a balanced assessment plan that provided reliable data of student learning (Fusarelli, 2002). Five years later the VDOE introduced guidelines for Balanced Assessment Plans, still giving divisions considerable autonomy, but adding more detail to the expectations with language that would suggest one or two performance assessments for state accountability is insufficient as it specified that divisions should have a variety of assessments to include performance assessments for state accountability. The interviews in this study were conducted in the fall of 2021, 2 years after the Balanced Assessment Plan policy and following the disruption of the COVID year, which may have delayed division adjustments to the Balanced Assessment Plan model. Although performance assessments are desired to promote higher-order skills, to ensure that all students are making progress on the standards an adherence to the VDOE policy of a Balanced Assessment Plan would provide more accurate data on student growth than the current small set of performance assessments being used by the divisions in this study.

Although division Balanced Assessment Plans may be limited on what student progress is measured, the LAAs are not the only measure of student progress, nor the only way students and teachers receive feedback on student understanding and growth. The assessments specified in

this study only reflected the assessments used as LAAs to replace the SOLs. The division leaders reported that teachers used additional performance and multiple-choice assessments in the classroom that are not part of the division's Balanced Assessment Plan. In addition, to address concerns about what data can be gained from the performance assessments and to provide additional data that students are meeting the SOLs and growing academically, one division created short multiple-choice quizzes for teachers to administer on content, and another division created skills-based multiple-choice pre- and post-assessments. These new developments in these divisions being implemented in the 2021-2022 school year better align with the extant literature and the VDOE balanced assessment policy. Although school and teacher practices of using a variety of assessments to measure student learning better align with the extant research, the set of solely performance assessments that divisions are currently using as LAAs for state accountability measures contradict the best practices outlined in the literature. The limited number and types of the LAAs do not provide reliable, generalizable scores that would meet the state expectation of ensuring that all students in Virginia are making adequate academic progress.

Since the initial introduction of the Balanced Assessment Plan policy in 2019, the VDOE has continued to emphasize a system of balanced assessments which is more in accordance with existing educational research. In August 2021, the VDOE issued implementation support for Balanced Assessment Plans that stressed a “move toward balanced assessment” which “must include performance assessments” (VDOE, 2021b). In the document the VDOE provided an example where a performance assessment might serve as the LAA for one unit while other skills “are best measured through local alternative assessments composed of multiple-choice and short answer questions,” specifically indicating a mix of both performance assessments and multiple

choice and short responses (VDOE, 2021b, p. 1). While many divisions quickly moved from a single multiple-choice SOL test to a small number of solely performance assessments following the 2014 policy, as the VDOE policy continues to evolve and the VDOE continues to further define the Balanced Assessment Plan expectations divisions may develop more research-based, balanced assessment systems that can be more reliable, generalizable measure that meet the state objectives of ensuring that all students across Virginia "are making adequate academic progress in the subject area and that the Standards of Learning are being taught" (VDOE, 2014, p. 1).

Currently, the divisions in Virginia that have been successful in implementing the use of performance assessments to replace the SOL test, are using a small number of almost entirely performance assessments for state accountability. While the current practice of using from one to six performance assessments fulfills the VDOE policy as initially written, the limited number of assessments contradict the existing literature concerning the reliability and appropriate use of performance assessments. As the 2019 VDOE policy concerning Balanced Assessment plans continues to evolve and better articulate the goals and intended direction of the reform, more divisions may adjust to more balanced systems of assessment which better align with the current literature, as two in the study have already done.

Types of Performance Assessments in Use. Within the one to six assessments being used, divisions reported using a variety of performance assessments. Of the assessments submitted to this study five are IDMs, three are modified or simplified IDMs, five are DBQs, and the rest consist of journal or letter writing as a historical figure, research to create a brochure, PowerPoint or other report of information, or creative writings of imagined scenarios. Besides the types of assessments shared with the study or reported in an interview, 6 of the 12 divisions

also allowed teachers to develop at least some of their own assessments, and thus could not specify all of the types of performance assessments being implemented for the LAAs.

The variety of assessments is appropriate and aligned with the literature as social studies require a wide range of skills, such as analyzing graphs and data, interpreting images, contextualizing and analyzing the perspective of documents, researching, communicating both verbally and in writing, and constructing arguments based on evidence (S. G. Grant et al., 2004; O'Brien, 1997; Suh & Grant, 2014). This range of skills to be assessed in the social studies is evident in the Virginia Standards of Learning as Standard 1 for each social studies course identifies seven distinct skills students should learn. Thus, the use of a variety of types of performance assessments in the LAAs allows divisions and teachers to use the appropriate type of task to assess the identified skills and allows students the means to demonstrate the variety of skills that make up social studies instruction. Although using a variety of assessments allows students to demonstrate a variety of different skills or demonstrate those skills in diverse ways, this practice further reduces the reliability and generalizability of the scores. When using one to six assessments with different skills measured on each one, the reliability of the data on student achievement of each skill or on any one assessment is further limited. Research has shown it takes a large number of tasks to obtain generalizable scores as previously discussed; the limited assessments do not offer enough information to draw reliable conclusions (Gao et al., 1994; McBee & Barnes, 1998; Stecter & Klein, 1997; Webb et al., 2000).

While divisions used different types of assessments, these assessments focus on similar skills with 11 of the assessments submitted to the study requiring students to analyze documents or images, and 13 requiring students to construct an argumentative essay in response to a prompt. Additionally, in several divisions the LAAs are the same format with different content; three

divisions adopted all IDMs or modified IDMs and another division uses DBQs. This means all of the performance assessments follow a similar format and require the same skills of students in each assessment. While having assessments of a similar format provides a larger number of tasks per student to build reliability, the same type of task may repeatedly disadvantage students who do not perform well on that type of task, and thus cannot accurately reflect student achievement. The research of Shavelson et al. (1993) shows that student scores may vary depending on the nature of the task, with some students performing stronger or worse on particular types of tasks than other types. The use of a limited type of assessments that focus on the same skills also contradicted the literature that argues that no single type of assessment can provide all the data students and schools needed to improve teaching and learning and promoted a balanced assessment system with a variety of assessment measures (Chappuis et al., 2017; Wren & Gareis, 2019). This creates a challenge for school divisions in the construction of their LAAs because they must create assessments that are both reliable and generalizable. Division leaders need to carefully analyze the intended goals of the Balanced Assessment Plan and how the types of assessments in the plan align with those instructional goals.

Based on the existing literature, the number and type of assessments being used by divisions in the study is not sufficient to produce reliable, generalizable data on student achievement or progress on the SOLs. The reliability of the LAA data was not only a concern for demonstrating student progress on the SOLs for state accountability purposes, but also because divisions are using performance assessment data to make other decisions. Six of the 12 divisions reported using student outcomes on the performance assessments to draw conclusions about student growth and/or to provide feedback to students on their progress. Five divisions reported using student data on performance assessments to improve and inform instruction. Given the

limited reliability of student response data, divisions needed to reflect on what conclusions could be drawn about student progress with such a small number of performance assessments and the nature of the tasks before making decisions on student achievement or classroom instruction. Since divisions report that teachers are using other assessments of a variety of types in the classroom, the data from the LAAs may need to be combined with data from other classroom assessments outside the formal Balanced Assessment Plan to provide a more complete picture of student progress and to make decisions about teaching and learning.

Divisions have responded to the autonomy to devise their own assessment plans and develop performance assessments by replacing the single multiple-choice SOL tests with a small number of performance assessments of a variety of types. Given the limited generalizability and reliability of performance assessments, the primary reliance on a limited number of performance assessments contradicts the proper use of performance assessments as discussed in the extant literature. The evolving emphasis of the VDOE on Balanced Assessment Plans could address these limitations and result in bringing division practice more in accordance with educational research. The current reliance on performance assessments further underscores the need to ensure that the LAAs being used by divisions meet the standards of quality performance assessments and that teachers and division personnel are well-trained in the development and implementation of performance assessments. Given the limitations of the current set of assessments, the existing LAAs need to be high quality to provide any meaningful data, thus divisions need to continue to invest the time and resources in performance assessment development and implementation to ensure that this performance assessment initiative meets the state goal of ensuring all students in Virginia are meeting the state standards and achieving academic progress.

Discussion of Research Question 3 Findings: The Quality of LAAs

The VDOE movement to LAAs started in 2014 as a grassroots implementation with divisions given the autonomy to develop assessments, assessment plans, and scoring systems. As the VDOE continued to communicate the direction of the reform through workshops and trainings that began in 2016, division understandings of performance assessments continued to evolve, as one division leader stated, "our teachers were kind of getting focused that everything had to be this big project, and so I was trying to get this message across that constructive response can be a one paragraph kind of thing" and another division leader reporting that over time "we were first shifting away from the sole-PBL model to a broader use of variety of performance assessments." To further articulate the direction of the reform, the VDOE clarified its definition of quality performance assessments and thus their expectation for performance assessments through the VQCT in 2019. Divisions in this study already had performance assessments in use for state accountability prior to the communication of these quality criteria and were working to build teacher understanding of the tool and using the tool to review both their existing and newly developed assessments. Given the evolution of the VDOE policy, this study sought to explore the extent to which the assessments currently in use aligned with the VDOE standards and expectations.

Eleven of the divisions in the study submitted two LAAs, one from USI and one from USII to be reviewed against the seven criteria and 17 sub-criteria of the VQCT. The assessment scores ranged from 18 to 46 out of a possible 51 points with a mean of 29.64, median of 30.5, and a mode of 31. Most of the assessments submitted to the study are authentic tasks involving higher-order thinking, with students communicating their ideas in writing and/or verbally and tasks and resources well-aligned to the interned learning outcomes scoring higher on Criteria 1,

2, 3, and 5. While there are some issues concerning cultural sensitivity or bias (Sub-Criterion 5C), the major weaknesses in the performance assessments are more logistical rather than the structure of the assessment. The scores are lower on Criteria 4, 6, and 7, in terms of identifying and specifying instructions, schedules, acceptable modifications for accessibility, and rubric usefulness or generalizability. These scores matched the findings of Brookhart (2005), who found that most teacher-developed assessments are of sufficient quality and are successful in matching assessments to the learning standards but that the assessments needed better rubrics.

Performance on Criterion 1, Criterion 2, and Criterion 3: Intended Learning Outcomes, Authenticity, and Language Use. Criteria 1, 2 and 3 focus on the nature of the task students are being asked to complete as well as whether the task itself meets the goals of performance assessments to promote authentic, higher-order thinking skills. In social studies, the authentic skills (Criterion 2) of analyzing sources and constructing arguments based on evidence align with the requirement of disciplinary practices "such as analysis, evaluation, synthesis or original creation" in Sub-Criterion 1B and the deeper learning competencies of 'Learning how to think critically…communicating effectively…and developing an academic mindset" specified in Sub-Criterion 1C (VDOE, 2019d, p. 2). All of these skills require the use of diverse forms of language when engaging with data and expressing a response as specified in Criterion 3. Overall, the assessments submitted to this study scored well on Criterion 2, Criterion 3, and Sub-Criteria 1B and 1C with 68.2% of the assessments scoring Full or Partial Evidence on Sub-Criteria 1B and 1C, 72.7% scoring Full or Partial Evidence of Criterion 2, 77.3% scoring Full or Partial Evidence on Sub-Criterion 3A, and all 22 assessments scoring Full or Partial Evidence on Sub-Criterion 3B. The high scores on these three criteria reflect the success of these divisions in

constructing performance assessments designed to meet the goals of performance assessments with authentic, higher-order and deeper-learning competencies.

Criterion 1, which measures the identification of and alignment to the standards along with the use of complex thinking and deeper-learning competencies, scored the fourth highest mean of the seven criteria. Sub-Criterion 1A, alignment to the standards, scored slightly lower than the other two sub-criteria since six of the assessments did not include the standards on the materials provided to the researcher. For the assessments that did provide the standards, 15 scored full or partial evidence of alignment between the standards and the task and resources provided. Although 1A focused on a more logistical concern, the identification of standards, divisions needed to clearly identify the standards to ensure that when individual teachers use the performance assessment, teacher and classroom practice focus on students engaging in the intended skills. The assessments scored higher on 1B and 1C, with means of 2.18 and 2.10 out of 3 respectively. Eleven assessments involved analyzing documents, images, and maps, and 13 required students to construct an argument based on documents or researched information. These higher scores on Criterion 1 matched the research that asserts performance assessments that ask students to use a variety of skills and knowledge to evaluate sources and construct and defend an argument are better methods for students to display higher-order understanding and deeper learning (Baron, 1996; Darling-Hammond & Adamson, 2014; Foote, 2005; Wiggins, 1998; Wren & Gareis, 2019). Those assessments that scored lower on these criteria were the brochures, letters, news articles, or other writing tasks that could be completed based on repeating or reframing learned content, and thus did not align with Virginia Standard 1, which requires "analyzing and interpreting artifacts and primary and secondary sources," "interpreting" visual images, or "using evidence to draw conclusions and make generalizations" (VDOE, 2015, p. 1).

The majority of the LAAs submitted to this study were intended to promote complex thinking and deeper-learning competencies with 15 of the assessments, 68.2%, scoring full or partial evidence of these sub-criteria, thus aligning with the literature's goals for performance assessments and the expectations of the VDOE.

Criterion 2, the third highest mean, measures the authenticity of the performance assessments; this corresponds with the research that argued performance assessments should engage students in meaningful tasks which resemble professional practice and prepare students for the real-world (Darling-Hammond, 2014; Gulikers et al., 2004; Wiggins, 1998; Wren & Gareis, 2019). Authenticity in social studies assessments can translate into a variety of tasks, as the social studies professions require a wide range of skills such as analyzing graphs and data, interpreting images, contextualizing and analyzing the perspective of documents, conducting research, constructing arguments based on evidence, and communicating both verbally and in writing (S. G. Grant et al., 2004; O'Brien, 1997; Suh & Grant, 2014). The assessments in the study mainly focus on the skills of the social studies profession as 72.7% (16 assessments) of the performance assessments showed full or partial evidence of authenticity. These assessments asked students to analyze images, documents, and/or maps, and write analytical responses to prompts or complete a real-world task like advocating for particular policies or responses to an event or issue. Like Criterion 1, the lower scores on Criterion 2 were the brochures, research tasks, writing of letters, or news articles that just repeated learned material without requiring the analysis of sources or constructing an argument from the information, since those are not authentic skills of the social studies disciplines.

Criterion 3, use of language, scored the highest in terms of students using a variety of sources of information, specifically documents, images, graphs, and maps to obtain information

and then express their ideas and products in writing and/or verbally. Seventeen of the assessments required students to encounter a variety of information that scored Full or Partial evidence on 3A, while 19 had full evidence requiring students to use one or more means of communication to convey their learning and products. Similar to Criteria 1 and 2, while students are reading and viewing a variety of sources and writing responses, it is possible, as found by Khattri et al. (1998), that students are not engaging in higher-order thinking. When these students engage in descriptive writing that does not require critical thinking, it allows for lower quality writing. Criterion 3 scored higher than Criteria 1 and 2, supporting the research that asserts that even when students are using language in a variety of forms, the tasks are not necessarily constructed in such a way as to require higher-order thinking or authentic use of skills in that language use (Khattri et al., 1998).

While the divisions developed their LAAs with the intent of being authentic and promoting deeper learning and complex thinking, there is the possibility that even though a performance assessment may ask a student to read a prompt, draw information from sources, and write a response, students still may not engage in the intended authentic, higher-order thinking or analysis (Baker, 1994; Cumming & Maxwell, 1999; Linn & Baker, 1996; O'Brien, 1997). Of the assessments evaluated for this study, 11 of the 22 reflected the concerns of the literature as students were asked to either research and write or to view documents and answer a prompt, but the assessment could have been completed by summarizing secondary sources or regurgitating taught material without engaging in the analysis of documents, images, or maps or constructing and supporting an argument. Three of the assessments had students produce brochures, seven asked students to write news articles or descriptive letters from historical perspectives or similar tasks, and one was a writing task that summarized data from graphs and charts. Each of these

tasks, with enough structure and careful classroom delivery, can require students to authentically use document analysis and higher-order thinking skills; however, with the instructions provided, the assessments in the study made it possible for students to complete the writing task based on reframing content in the SOL curriculum framework and/or provided during classroom instruction. Thus, the assessments, while intended to meet Criterion 2 and Sub-Criteria 1B and 1C, in practice did not demonstrate the higher-order thinking skills authentic to the social studies as listed in the Virginia Standards of Learning, nor did they match the intentions of performance assessments. These shortcomings underscored the need to clearly identify the standards, as measured in Sub-Criterion 1A, and carefully evaluate the structure of the performance assessment to ensure that the assessment required students to engage in the designated skill.

The omission of standards and the subsequent potential for misalignment of tasks or classroom instruction combined with the potential for seemingly authentic, higher-order tasks to devolve into the lower-order reframing of learned material demonstrates the need for divisions and teachers to engage in sufficient professional development and to deliberately use research-based strategies to ensure quality performance assessments. The alignment of the tasks to the intended learning outcomes and to the goals and purpose of performance assessments can be improved through a careful analysis of the LAA, including the nature of the task, the instructions provided to teachers and students, and the rubric to ensure that the implementation of the task requires students to engage in the intended authentic, higher-order thinking skills to complete the task and to score well on the assessment. To construct meaningful performance assessments and to engage in careful evaluation of the instructions, resources, and completed assessment, teachers and division personnel must be well-versed in the constructs and components of quality performance assessments.

Performance on Criterion 4: Success Criteria. While the first three criteria focused on the nature of the task students are asked to complete, the last four criteria focused on more logistical elements such as rubric, resources, accessibility, and schedule. Each of these criteria affects the integrity of the performance assessment and thus the quality and meaningfulness of the performance assessment and resulting student outcomes. Criterion 4 focuses on student success criteria in terms of the existence of a rubric, and the alignment, clarity, and transferability of that rubric. Criterion 4 scored on the lower end of the 7 criteria, with a mean of 4.59 out of 9 partly due to the omission of a rubric in the materials submitted to the study. The more significant concern in Criterion 4 was the conflicting requirements on the VQCT to create rubrics that are specific enough to clearly communicate expectations and "to provide useful feedback to students," while also being generic enough "to be used across performance assessments within the course" (VDOE, 2019d, p. 4).

The conflict between tightly aligned rubrics that provide useful feedback and the need for rubrics that can be used across all assessments in the course is further complicated by the introduction of the VDOE common rubric. In 2020, 6 years after the initial reform and one year after the VQCT, the VDOE introduced common rubrics for history and social science; the most recent VDOE guidelines state that schools are to use the VDOE common rubrics to evaluate all LAAs (VDOE, 2021a). At the time of the interviews, 10 of the 12 divisions had or were moving to the common rubric; therefore, many of the assessments in the study used the common rubric. The VDOE rubric is written to fit any performance assessment and is intended to be broad enough to encompass a variety of performance assessments. However, this does not align with Criterion 4A, which states that the assessment must "include[s] a rubric that is tightly aligned to the performance expectations of the intended learning outcomes within the performance

assessment" (VDOE, 2019d, p. 4). The existing literature underscores the importance of tightly aligned rubrics as clear success criteria force teachers and test developers to clarify specific and desirable levels of performance and structure instruction to achieve those levels, supporting the language of Sub-Criterion 4A (Shiel, 2017; Wiggins, 1998). While the VDOE's generic rubric creates familiarity to improve teaching and learning, provides students with consistent expectations, and allows students to demonstrate growth over time, this contradicts the work of Wiggins (1998) who argues that a common rubric does not fit every type of assessment given nor does it provide specific feedback to students. Performance assessments can take on a variety of forms, from on-demand tasks that are shorter in time such as constructed-response or stand-alone tasks to extended, long-term performance assessments such as curriculum-embedded tasks or complex projects but the common rubric, as one division leader expressed, seems to "be designed for free response questions and document-based questions" and does not match as well with other types of performance assessments "like multiday projects," making it challenging for teachers to utilize the rubric for various types of assessments (Brookhart, 2015; Khattri et al., 1998; Wren & Gareis, 2019). As the evaluation team compared the intended learning outcomes and assessment tasks to the common rubric, evaluators felt the rubric was "too vague" and did "not align well with the performance assessment"; thus, the team gave limited evidence, a score of 1, to the 13 assessments using the common rubric and scores of 2 and 3 to divisions that provided more task-specific, teacher or division created rubrics. The evaluation team's ratings of the rubrics matched with concerns voiced by the divisions in the interview process about the alignment of the generic rubric to project-specific intended learning outcomes. One division leader expressed that "it's a generalized rubric and its necessarily ambiguous about things because it has to apply to all kinds of different work, but we are applying it to a specific task."

The conflicting aims of the VDOE in requiring the use of a generic, common rubric while also evaluating assessments for having tightly aligned rubrics lowered the scores of the LAAs on the VQCT.

The other complication of the common rubric related to the contradiction between the language of the VDOE common rubric and the requirements of Sub-Criterion 4B. This sub-criterion evaluates the language of the rubric and the ability of the rubric to provide useful feedback to students about their work. The existing literature stated that the clearly communicated success criteria, as measured by Sub-Criterion 4B, were important as they enabled students to know they have achieved their learning goals (Shiel, 2017; Wiggins, 1998). Both the review team and division leaders reported that the current language of the common rubric is not friendly for students, especially middle school who are the subject of this study, or for parents who are not familiar with educational terms and language; thus, this rubric has limited usefulness for students to understand the feedback and use that information to improve performance going forward. Assessments using the common rubric scored low on these criteria, as the review team agreed that the rubric language was "not very student friendly" and "does not allow for effective feedback as it is tough for students and teachers to understand it." The review teams' comments echoed teacher complaints that the common rubric was, as one division leader shared, "hard to make sense of and hard to share with students." One division created its own rubric that, while using similar, student-friendly language, allowed the altering of specific references to content or unique skills from one assessment to the next. Thus, the division-created common rubric was consistent across assessments and created consistent expectations, but minor word substitutions tailored the content of each box on the rubric to a specific assessment,

allowing teachers and students to better see the connection between the rubric and the task being assessed.

The tension between common rubrics and the need for tightly aligned, clear rubrics created disparities between scores on Sub-Criterion 4C compared to 4A and 4B. Throughout the evaluation process, performance assessments using the state common rubric routinely scored low on Sub-Criteria 4A and 4B for failing to be tightly aligned and not in audience-friendly language but scored high on 4C which measures the feasibility of using a rubric across performance assessments. The 10 assessments that used the common rubric scored a 1 on both 4A and 4B, but a 3 on 4C, while the three tightly aligned, clearly-written rubrics scored a 3 on both 4A and 4B, but a 0 on 4C. The evaluation team was frustrated by Sub-Criterion 4C because the rubric that the team felt was the clearest, most tightly aligned to the task, and written in language that could be used by students to improve scored a 0 in 4C because the rubric was so task-specific. One evaluator noted, "Sub-criterion 4C punishes good rubrics that are tied to the task simply because they are not generic enough." The tensions between being aligned to specific learning objectives and being generic enough to be applied across different types of assessments led to disparate scores on the three sub-criteria of Criterion 4.

The literature stresses the need for tightly aligned rubrics that can provide clear feedback for students to measure their growth and know how to improve, which aligns with the stated requirements of the VQCT Sub-Criteria 4A and 4B (Shiel, 2017; VDOE, 2014; Wiggins, 1998). This creates a disconnect between the literature-based sub-criteria 4A and 4B and sub-criterion 4C that calls for a rubric which can be used across all the diverse types of performance assessments in a course. The VDOE created common rubrics that divisions are required to use with their LAAs contradicts both the literature and sub-criteria 4A and 4B with its generic

content and difficult language; in contrast, well-written rubrics that students could use to identify areas of growth are penalized on the VQCT for their lack of transferability. While the VDOE goals of consistent expectations and the existence of a means to measure growth over time are important, promoting those goals through the current common rubric and Sub-Criterion 4C may not be appropriate, especially in a field like social studies where performance assessments can take a variety of forms and when the use of a common rubric conflicts with the requirements of Sub-Criteria 4A and 4B.

Performance on Sub-Criterion 5: Directions and Bias. The sub-criteria of Criterion 5 focused on the directions, prompt, and resources of the assessment and evaluate two different elements of performance assessments. Sub-Criteria 5A and 5B focus on the existence, alignment and clarity of the instructions and resources, logistical factors that ensure the assessment is structured in such a way that students can demonstrate their ability to perform the intended learning in the assessment. These two sub-criteria are easier to observe and evaluate, but Sub-Criterion 5C requires a deeper analysis of the nature and design of the task, prompt, and materials to identify any potential biases. This may affect a student's ability to demonstrate the identified skills, but also requires developers to be alert to their own biases and be sensitive to how other groups may respond to elements of the assessment.

Sub-Criteria 5A and 5B, measure the alignment, presence, and clarity of instructions and resources. These scored higher, aligning with the scores on Criteria 1, 2, and 3, reflecting the overall quality of the performance assessments created. While Criterion 5 scored the second highest mean of 5.82 out of 9, 5B scored higher than 5A and 5C. Sub-Criterion 5B, like Sub-Criteria 1A and 1B and Criterion 2, reflected that divisions have done well in constructing tasks with clear prompts and directions and assembling resources that are aligned and appropriate for

the task and skill. Thirteen of the 17 (77.3%) of assessments had full or partial evidence of 5B with clear, complete, and accessible prompts and resources for students. Sub-Criterion 5A, similar to Sub-Criterion 1A, scored lower because six assessments did not provide standards to measure the alignment of the resources. The team could only compare the alignment of the resources to the prompt or task. In addition, there were three assessments that scored low on Sub-Criterion 5A for poor alignment of the resources to the skills as students were given general URLs to large depositories of resources that the team was concerned students would have trouble navigating. The concerns described in Criteria 1, 2, and 3 over whether a task was in practice authentic and higher-order thinking or a reframing of learned materials relates directly to the elements evaluated by Sub-Criteria 5A and 5B. The existence of clear and complete instructions for both teacher and student and appropriate resources, as measured by 5A and 5B, are the structures necessary to maintain the intended authenticity and higher-order, real-world competencies measured in Criteria 1 and 2 and the meaningful interaction with language measured in Criterion 3. Fully documented, clearly communicated instructions and well-chosen resources are tools developers can use to structure classroom activities to ensure quality performance assessments in practice.

Sub-Criterion 5C, while related to 5A and 5B in allowing students to adequately display their progress on the intended skills, addresses the bias and cultural sensitivity of the performance assessment. The assessments scored lowest on 5C, with 15 assessments raising concerns about cultural sensitivity or bias by the evaluation team. Only four assessments received specific comments of “no bias” or “bias free” from the review team. One concern of the review team that resulted in lower scores on the bias measurement was the number of prompts and/or resource sets that were leading students toward a particular perspective or interpretation

of the past. Biased tasks and resources run counter to the goals of social studies education and are inherently not authentic to the discipline. The National Council for the Social Studies (1994) states that the goal of the social studies is to "help young people make informed and reasoned decisions" and the literature argues that the authentic tasks of historians are to use historical reasoning and evidence to formulate decisions (S. G. Grant et al., 2004; Maddox & Saye, 2017). Thus, these performance assessments go against the literature in the field because they lead students to particular answers rather than allow students to authentically use higher-order thinking around an open-ended question to evaluate documents and then devise and justify their own stance (S. G. Grant et al., 2004). Besides the lack of authenticity and higher-order thinking, leading questions rest on the views of the developer and exclude the multiple perspectives that exist in the culturally diverse society that social studies education seeks to prepare students for (National Council for the Social Studies, 1994). Not only does this prevent students from engaging with multiple perspectives, but a task that forces a student to a particular interpretation may not allow a student to demonstrate the intended skills given the diversity of perspectives that student hold and may have been exposed to (Baker, 1994). Limiting or leading student responses also could be potentially problematic for students or parents who may hold different perspectives and lessens the authenticity of the discipline. Thus, developers need to critically analyze all elements of a performance assessment including the nature of the task, the prompt, and the provided resources; this will ensure the assessment allows the student to engage in the authentic task of the discipline and define their own evidence-based stance on the topic. In addition, bias-free assessments will allow and encourage representation of the multiplicity of viewpoints that exist in the real world.

The review team also spent considerable time discussing performance assessments that asked students to take on the role of historical figures or groups, especially marginalized groups, and write from or about those perspectives. Even assessments that asked students to write from the perspective of majority groups in the past often lacked sufficient resources for students to gain a deep understanding of those experiences. The review team was concerned these tasks risked products or student work that may perpetuation misconceptions or stereotypes about people in the past or be offensive to particular groups, rather than align with the standards and promote learning; thus, these tasks received lower scores on 5C. The difficulty for developers and evaluators is identifying areas of potential cultural insensitivity. The members of the review team for this study had all spent years as classroom teachers, but while three of the review team members immediately identified all the ways that students could respond in culturally insensitive ways and gave the assessments a 0 on 5C, two other team members read the instructions as they were intended by the developers and scored 5C a 2 or 3. During the team discussions the higher-scoring team members, after hearing lower-scoring members' rationales, admitted to never having thought of the worst-case outcomes but did concede that student products could be inappropriate, stereotyped, or offensive and agreed with the consensus score of a 0 or 1. This points to the challenge for performance assessment developers, rooted in their own cultures and experiences, to be alert to the potential for cultural insensitivity in a task. While it is possible that developers may not be aware of what may be perceived as inappropriate by other individuals, the current social and political climate necessitates a greater sensitivity to how performance assessments may be perceived. The other challenge is that often developers, like members of this study's review team, only consider the desired student outcomes and responses without thinking about the multiple different ways students may respond to a prompt or task.

The divisions in this study had clear and complete instructions, prompts, and resources, which are important for ensuring the quality of the performance assessments and addressing the potential concerns with Criterion 1, 2, and 3. The written instructions for both student and teacher provide the structure for the assessment to be implemented consistently and as intended by the developers across different teachers and classrooms. With clear instructions and resources, developers communicate the steps to be taken by students as they progress through the assessment; this will ensure that students engage in the desired authentic, higher-order skills rather than resort to a repetition of learned material. Coupled with a clear, tightly aligned rubric from Criterion 4 that holds students accountable for engaging in the desired skills, developers promote quality performance assessments through clear instructions and resources. This shows developers need to critically analyze all elements of the LAAs, not just the nature of the task, to create a structure that results in the desired learning outcomes. In addition, developers need to critically analyze all elements of a performance assessment for potential areas of cultural insensitivity and utilize research-based strategies such as an external review for cultural biases in order to prevent a performance assessment or student response that may promote stereotypes, cultural misconceptions, or offensive material rather than the desired deeper-learning competencies.

Performance on Criterion 6 and Criterion 7: Accessibility and Feasibility. Criteria 6 and 7, when combined with Criteria 4 and 5, provide additional structures to ensure that performance assessments are implemented across different classrooms in a way that requires all students to engage in the intended authentic, higher-order, deeper-learning competencies. Despite the importance to consistent implementation, Criterion 6: Accessibility and Criterion 7: Feasibility are the two lowest scoring criteria of the seven. While the scores on Criteria 1, 2, and

3 demonstrate that divisions have invested time in developing the actual tasks, there seems to be less emphasis on the logistical characteristics of plans for scaffolding, differentiation, time schedules, and duration.

Admittedly, the lower scores result from the lack of this information in most of the assessments. One of the assessments had a 0, "No evidence," and 10 had limited evidence of 6A; 14 had no evidence of 6B; eight had no evidence of 7B; and eight had no evidence of 7C. It is possible that this information was not on the material provided to me and might be documented elsewhere, or this may be information that was provided to teachers in less official means through division PD, school-level PLCs, or teacher collaboration. Even if this information is shared in less formal means, specific documentation of these elements is important. Clarity on the appropriate scaffolding and accommodations (Criterion 6) and the scheduling of the assessment (Criterion 7) create consistency in the student experience and provide for students of varying abilities to best demonstrate their progress on the intended learning outcomes.

While the elements of Criteria 6 and 7 may be things experienced teachers know and do not need explicit instructions to enact, less experienced teachers or teachers with less understanding of performance assessments may need these instructions. In addition, most divisions have more than one school and within each school there may be more than one teacher per course, creating a need to ensure consistency in the implementation of the assessment across all classrooms and students in the division. Without clear indications of where the assessment corresponds with prior learning and the proper sequence in instruction and instructional time dedicated to the LAA (Criterion 7), teachers may not embed the performance assessment in instruction as intended. One division leader stated that teachers seemed to think, "okay we're at the end of the unit, time to give the assessment," but the leader was frustrated that the LAA given

by the teacher “wasn’t even in the unit that…was just taught so the alternative assessment wasn’t anything [the students] had just learned.” Additionally, without a clear indication of the duration and schedule for the assessments teachers may allow too little time for students to adequately engage with the evidence and be able to sufficiently demonstrate their learning on complex skills. Since the goal of performance assessments is to shape classroom instruction, clear expectations of how the assessment connects to prior learning and a schedule for implementation are necessary to integrate the skills into all classrooms. Similar to Criterion 7, the scaffolding and differentiation in Criterion 6 may be assumed by experienced teachers, but this may not be as apparent for teachers who are less experienced with performance assessments and teaching higher-order thinking and argumentative writing. Regardless of experience, without clear instructions teachers may choose scaffolding or accommodations that could change the nature of the student outcomes and thus reduce the authentic, higher-order nature of the task. Specifications of acceptable and appropriate modifications for different types of learners and a schedule for the assessment provides the structure to promote consistency across teachers and schools to maintain the intended learning outcomes of the LAA for all learners in a particular course.

Especially for assessments that are being used across a school or a division, clearly communicated, detailed descriptions of these logistical elements are needed not only for less experienced teachers but also to ensure consistency of the assessment. In a loosely coupled system where each teacher makes decisions in their own classrooms, the lack of specific guidelines on acceptable accommodations, scaffolding, and scheduling, may result in different classroom implementation and affect the ways students engage with the assessment (Fusarelli, 2002). This would allow for a loss of the intended learning outcomes and, ultimately, alter the

conclusions about student progress and classroom instruction that could be drawn from the student products. Developers need to clearly document and define appropriate accommodations, scaffoldings, and scheduling for an assessment that would still maintain the intended learning outcomes so that the performance assessment is implemented consistently for all students to best demonstrate their progress on the intended skills in any classroom.

Summary

The findings of this study corresponded with the literature asserting teachers and division leaders needed extensive, on-going and repeated professional development to successfully develop and incorporate performance assessments and that teachers should be included in the process of performance assessment development. As a result of this investment in professional development, divisions have developed assessments focused on authentic, deeper-learning skills that involved the use of a variety of resources and student communication of their findings. Still, given the challenges of time and of the lack of a deeper understanding of performance assessments, these assessments reflect the concerns in the research that seemingly higher-order performance assessments may not actually require authentic, deeper-learning on the part of students. Despite these challenges, the divisions in the study remained appreciative of the autonomy to develop LAAs and committed to continuing to improve their LAAs and to build division capacity to implement performance assessments for the benefit of teaching and learning thus demonstrating the increased efficacy that Weick (1976) argues is a benefit of a loosely coupled system.

The findings in this study also highlight the challenges of a grassroots strategy of allowing individual divisions to determine how to meet a state-level initiative. Given the limited number of LAAs, the loosely coupled system of individual teachers administering the LAAs in

their own classrooms, and the autonomy granted to the teachers in half of the divisions in this study to choose their own assessments, the potential exists that all LAAs may not match the intended outcomes of division leaders, the VDOE or the developers. The lack of detailed supporting documentation concerning the logistical elements of the LAAs contributes to the possibility that LAAs created do not match the intent of the developers. The VDOE has developed the VQCT to support division work in improving and developing quality performance assessment, and most divisions were using the tool to evaluate and revise their assessments, however, the brevity and lack of documentation of the VQCT allows for different interpretations and applications of the tool has resulted in variation in the quality of LAAs across the state.

Implications for Policy and Practice

The VDOE is currently expanding the grassroots local alternative assessment initiative to high school social studies courses and potentially other courses; it is also continuing to emphasize the role of the VQCT to support consistency of LAAs across the state in meeting state standards. Weick (1982) wrote for a large-scale change, such as a shift to performance assessments, to be successful the leaders of the change must talk frequently about the goals of the initiative and clearly articulate a general direction for the reform with persistence and detail.

For this performance assessment policy to be successful and beneficial to teaching and learning the VDOE needs to continue to articulate the direction of the reform to promote even implementation across the state and for all students in Virginia (Fusarelli, 2002; Weick, 1982). To help achieve the goals of this policy, the VDOE can use the successful strategies of the divisions in this study which have embraced the autonomy granted to them and employed research-based strategies to develop quality performance assessments and invested in on-going professional development for division educators. The challenges faced by these divisions can

also inform the evolving LAA policy and practice in Virginia. Given the loosely coupled system of individual teachers administering the LAAs in their classrooms and the goal that performance assessments are more than a single event but are to improve teaching and learning, this study recommends that divisions engaging in a performance assessment initiative invest in high-quality, on-going professional development for division leaders and educators. This study also recommends that assessment developers undertake a careful, analytical review of the assessments to ensure the tasks align with the intended learning goals and includes thorough documentation of all the elements of the assessment to support proper implementation. Finally, since the VDOE and Virginia school divisions emphasize the use of the VQCT to ensure quality performance assessments, this study recommends more complete documentation of the VQCT to promote more consistent and effective usage of the tool. A summary findings, related recommendations, and associated literature is presented in Table 42 and followed by explanation and discussion.

Table 42

Summary of Findings and Recommendations for Policy and Practice

Findings	Related Recommendations	Supporting Literature
Divisions used teachers to develop and revise LAAs	Recommendation 1. School divisions should invest in quality, ongoing PD for teachers and staff surrounding performance assessments, their implementation, and their role in instruction and provide sufficient time to develop and plan quality assessments Recommendation 2. Teachers and division leaders need to be trained to thoughtfully and deliberately analyze the tasks in the performance assessments.	Bandalos, 2004 Goldberg & Roswell, 2000 Khattri et al., 1998 Koretz, Barron et al., 1996 Koretz, Mitchell et al., 1996 Learning Forward, 2022 Marion & Leather, 2015 Messick, 1994 O'Brien, 1997 Pfeifer, 2002 Sivalingam-Nethi, 1997 Stosich, et al., 2018 Stecher & Mitchell, 1995 Wiggins, 1998
Although some division LAAs were authentic and higher-order, the lack of documentation and structure for student tasks and teacher practice could result in student products that might not require deeper-learning or might be biased or culturally insensitive	Recommendation 1. Teachers and division leaders need to thoughtfully and deliberately analyze the tasks in the performance assessments.	Baker, 1994 Biemer, 1993 Brookhart, 2015 Cumming & Maxwell, 1999 Darling-Hammond & Adamson, 2010 Darling-Hammond et al., 2014 Goldberg & Roswell, 2001 S. G. Grant et al., 2004 Lane, 2014 Linn & Baker, 1996 McCann & McCann, 1992 Moon et al., 2005 O'Brien, 1997 Parke & Lane, 2008 Pecheone, & Kahl, 2014 Suh & Grant, 2014 Wiggins, 1998 Wren & Gareis, 2019
The most common division strategies for quality performance assessments were teacher feedback and the VQCT, but divisions and teachers lack comfort with the VQCT	Recommendation 1. School divisions should invest in quality, ongoing PD for teachers and staff surrounding performance assessments, their implementation, and their role in instruction and provide sufficient time to develop and plan quality assessments Recommendation 3. The VDOE should provide more extensive documentation and explanation of the VQCT to ensure both effective use and quality performance assessments across the Commonwealth.	Darling-Hammond, 2017 Darling-Hammond & Adamson, 2010 Darling-Hammond & Ancess, 1996 Goldberg & Roswell, 2000 Frey & Schmidt, 2007 Khattri et al., 1995 Kroesch, 2015 Messick, 1994 Meyer, 1992 Parke et al., 2006 Spillane & Zeuli, 1999 Stecher, 2010 Wiggins, 1989

Note. LAA = Local Alternative Assessment, PD = Professional Development, VQCT = Virginia Quality Criteria Tool, VDOE = Virginia Department of Education

Recommendation 1: High-Quality, On-Going PD

To effectively achieve large-scale change in this loosely coupled policy with individual teachers and divisions developing, administering, scoring and using performance assessments to improve classroom instruction, articulating and building common understandings and practices around performance assessments is necessary for the success of the policy (Weick, 1982). School divisions should invest in quality, ongoing professional development for teachers and staff surrounding performance assessments, their implementation, and their role in instruction and provide sufficient time to develop and plan quality assessments (Goldberg & Roswell, 2000; Khattri et al., 1995).

The divisions in this study, identified as successful in implementing performance assessments by educational leaders in Virginia, support the conclusions of the educational research concerning the need for focused, ongoing, and consistent professional development. Educational researchers argued for the need for sufficient professional development to ensure the successful implementation of performance assessments to improve teaching and learning. To promote the successful implementation of quality performance assessments to improve teaching and learning, professional development needs to be meaningful and of high quality. Eleven standards of quality performance assessments are defined in Learning Forward's (2022) "Standards for Professional Learning" focused on Conditions for Success that provide the basic framework for high quality learning, Rigorous Content for school staff, and Transformational Practices that change educator skills, practice, and mindset. As divisions plan professional development surrounding performance assessments, the Learning Forward standards provide guidelines for promoting meaningful, high-quality professional development.

The four standards under the category of "Conditions for Success" outline the need for a deliberate, organized system of professional development provided through a variety of channels. The standards of "Resources" and "Leadership" state that division leaders need to have a clear vision for professional development that sustains coherent support for educators as well as advocates for and allocate resources for this training. All of the divisions in this study reported professional development and training for their teachers over multiple years. Five divisions specifically stated they started the process of training teachers around performance assessments in or prior to 2014, when the state initiative to replace SOLs with LAAs was first introduced. Divisions in the study reported ongoing training, such as a four-year program of teacher training with a consultant in one division, a 3-year training cycle with one consultant and training with a second consultant in another division.

In addition to planning additional trainings, the third standard, "Culture of Collaborative Inquiry," emphasizes the need for these plans to allow educators to engage in continuous improvement and build their understanding (Leaning Forward, 2022). This need for on-going, repeated training was seen in the study. Divisions felt that this long-term process was producing results. One division leader shared, "Seven years ago the term performance assessment was spooky and scary and the understanding of what a performance task at that point wasn't completely ingrained into how we do things like it is now." Another said, "In 2014 performance assessments felt more disconnected to instruction and people did not find them meaningful, but now we have moved in our capacity with performance assessments."

Finally, the fourth standard, "Equity Foundations," supports the need for structures of individualized supports of PLCs and instructional coaches seen in the study, as the standard measures equitable access to learning and a culture of support for all staff. According to the

"Conditions of Success" and the experience of the divisions in this study, division leaders need to have a clear vision for the implementation of a new performance assessment initiative, thoughtfully design a long-term plan for providing professional development on performance assessments to educators and plan a system of on-going support and capacity-building individualized to specific teacher needs. A coherent plan of the training on the topics and skills of performance assessments, the identification of the most effective resources and means for delivering that training, and the designation of focused time throughout the year to engage in professional learning provides the necessary framework for high quality professional development and, ultimately, performance assessment success (Learning Forward, 2022).

Once the vision and structure for a system of high-quality professional development is in place, division leaders need to ensure that in each of those learning sessions all educators are engaging in "Rigorous Content" to improve student outcomes (Learning Forward, 2022). High-quality professional development on performance assessments should focus on training educators on "Professional Expertise" on performance assessments in terms of what constitutes a quality performance assessment and how to apply those standards in teacher practice. Divisions reported they needed to start with building common understandings of what a performance assessment looks like. One division leader explained, "Getting teachers to wrap their heads around the fact that we're no longer looking at memorization, we're really looking at application of what the students can do." The division experiences are in line with the literature which asserts teachers and administrators lack common understanding of performance assessments, deeper learning, higher-order thinking skills, and authenticity and constructing common understandings; thus, the subsequent shift to performance assessments to benefit pedagogy required quality and sufficient professional development (Khattri et al., 1998; Koretz, Barron et al., 1996; Koretz, Mitchell et

al., 1996; O'Brien, 1997; Sivalingam-Nethi, 1997; Stosich et al., 2018; Stecher & Mitchell, 1995). In Virginia one means of furthering teacher understanding and use of quality performance assessments would be training teachers in the VQCT and how to evaluate their own assessments against the tool as seen in the divisions in this study. In tandem with the standards for quality performance assessments, educators should be trained in Learning Forward's second standard of "Equity Practices" to identify biases or potential areas of cultural insensitivity and how to apply more inclusive practices when developing performance assessments and revising their performance assessments to promote equity (Learning Forward, 2022). Finally, professional development should promote the alignment of "Curriculum, Assessment, and Instruction" by training teachers on the purpose and skills of performance assessments and how to embed those skills and the assessment into instruction to improve teaching and learning. The Learning Forward standards outline the rigorous content for performance assessment professional development that provide educators with a deeper understanding of quality performance assessments; thus, promoting the development of carefully constructed assessments with clear instructions and free of bias. This, ultimately, will improve teaching and learning and "ensure the comparability of rigor and quality across the state" (VDOE, 2019d, p. 1).

The third category of Learning Forward's (2022) Standards for Professional Learning, "Transformational Processes," emphasizes the need for long-term, on-going professional development based on student and educator progress data that continues to expand educator capacity and results in significant changes in educator mindsets, skills, and practices around performance assessments. The divisions in this study demonstrated the need for continuous and ongoing professional development even after the LAAs are developed and in use. One division spent years doing a training on authenticity and performance assessments and still noticed last

year that numerous teachers did not grasp that performance assessments are not just end-of-year or end-of-unit assessments, but rather "that the performance assessments are for learning and not just of learning." This division is planning training this year to "go back and fill in the gaps we may have had on understanding," with PD sessions tailored to areas where teachers need additional support. Another division felt that while it had a strong foundation in performance assessments at the start of the initiative in 2014, they still spent last year "reframing" teacher understanding of performance assessments with a year-long professional learning program to further build teacher capacity. A third division reported they conducted training with teacher leaders every two years, and even with this repeated training, is currently seeing increases in teacher buy-in around performance assessments. Even in divisions that have "leaned in on teacher training" for years, teacher understanding continues to deepen and the divisions in this study have planned or are planning to continue to embed performance assessment training throughout the year, underscoring the need for on-going professional development. These experiences corresponded with the research that asserting initial training efforts, divisions need to continue to regularly provide additional professional development on performance assessments as teachers build their understanding of the role and use of performance assessments (Bandalos, 2005; Marion & Leather, 2015; Wiggins, 1998).

Building on the foundational condition of success "Leadership," "Transformational Processes" stress the need for a compelling vision for the performance assessment reform committed to a comprehensive professional development plan focused on supporting all educators (Learning Forward, 2022). Once there is a vision and clearly delineated plan, division leaders need to provide sustained, clear support for all educators on performance assessments, not simply present a workshop and assume that all teachers are now prepared to implement

performance assessments. Divisions in this study have followed this standard by planning division PD days around performance assessments supported by topics for PLC meetings or instructional coaches, including additional support through Canvas courses or division resource banks. To be effective these carefully planned learning sessions must have realistic goals and be centered around evidence-based "Learning Designs" (Learning Forward, 2022). For each professional development session, division leaders need to identify and set achievable learning goals centered on research-based strategies for performance assessments. Once the learning goals are set, each professional development session should be planned to best engage educators in the learning process. Educators should be given time to experience performance assessments in a way that mirrors the student experience may help "educators understand the rigors and requirements" of performance assessments and then have the opportunity to practice, then reflect, and refine. These quality professional learning designs can improve educators' use of performance assessments (Learning Forward, 2022). Similarly, the "Implementation" standard stresses the importance of designing this professional development over time and through a variety of modalities such as the PLCs and instructional coaches used by the divisions in this study. The repeated cycles of professional development seen in the study match with the Learning Forward standards by allowing educators multiple experiences of practice, reflections, and adjustments. Repeated support during the PD sessions to promote capacity-building and will improve teaching and learning. Finally, "Transformational Processes" promote "Equity Drivers" by requiring teachers to continue to identify their own biases and beliefs and learn about other perspectives (Learning Forward, 2022). This will allow teachers to better develop performance assessments that are authentic and without bias, avoid culturally insensitive tasks or student products, and implement performance assessments in ways that are accessible to all students.

Performance assessment training cannot be a singular experience, nor can division leaders assume that teachers have mastered the concepts and practice of performance assessments and deprioritize the initiative in division professional development plans, especially given the importance of performance assessments in improving teaching and learning. Both the literature and the progress division leaders reported after years of focused professional development on performance assessments demonstrate the need for and importance of extended, focused professional development on performance assessments when implementing similar initiatives. The Learning Forward standards and the experiences of the successful divisions in this study demonstrate that school leaders seeking to successfully implement a performance assessment reform should be prepared to engage in thoughtful, deliberate, long-term professional development in order for teachers and administrators to achieve the desired effects on teaching and learning.

Recommendation 2: Deliberate Review and Documentation of Performance Assessments

While other states have implemented performance assessment accountability reforms, such as Maryland, Kentucky and Washington, those states created the assessments for accountability and used expert reviews and formal pilots in the development process. Virginia, instead of a tightly coupled system where the state mandated the assessments, chose a grassroots approach allowing individual divisions to define and develop their own assessment plans for state accountability. To implement performance assessments that are equitable and comparable in rigor and quality for all students in this loosely coupled system, teachers and division leaders need to analyze the tasks in the performance assessments thoughtfully and deliberately. They need to include thorough, detailed instructions for implementation for both teachers and students to ensure the quality of the tasks and to protect against bias and/or cultural insensitivity.

The goal of performance assessments is to have the ability to promote the development of desperately needed higher-order thinking skills and the transfer of knowledge in students (Biemer, 1993; Darling-Hammond & Adamson, 2010; Darling-Hammond et al., 2014). While assessments may appear complex and authentic, closer examination may reveal that the task could be completed by memorization, or that the real-world elements were tacked on to a mundane task (Cumming & Maxwell, 1999; Linn & Baker, 1996). Teachers and division leaders need to critically analyze the tasks, the structure of the task, and the resources provided to ensure that students are required to engage in the higher-order, authentic tasks desired and intended by the performance assessment. As O'Brien (1997) asserts, students can be asked to read a prompt, draw information from sources, and write a response without engaging in higher-order thinking; and, likewise, writing can be descriptive without making an argument or incorporating sources critically. As the review team discussed each assessment in the study, at least one team member, if not more, raised the questions about whether students had to engage in higher-order thinking to complete the task in 15 of the assessments; and for 14 of the assessments, one or more team member questioned whether the task was authentic to the discipline. For example, historians do not usually write letters or journals as if they are a figure in the past. Thus, a careful analysis of the task, the teacher- and student-facing resources and instructions, and the rubrics are needed to ensure that a student must demonstrate the intended skills to complete the task and score well on the assessment.

Beyond just the task or prompt, divisions need to critically examine how the documents and visual resources are being used in the task. O'Brien (1997) voices a concern that students could view documents and not incorporate them critically. Suh and Grant (2014) had similar findings when they evaluated the use of images in NAEP questions and found that visuals

resources are often used in assessments for observation and summary, not a higher-order evaluation of the reliability and validity of these images. In nine of the assessments that included documents and images, the review team agreed that the student could complete the task by drawing on taught information or prior knowledge and not use the resources provided, as there was no mechanism in the task nor the rubric to ensure that students used or critically analyzed the sources or images. In three additional assessments, students were asked to conduct research, but were given broad URLs that took students to sites that had a wealth of information rather than given specific documents or images to use. Students were not given structures or guidance on how to navigate these sites, how to conduct quality research, or how to critically analyze the material they found. Thus, depending on the teacher-led instruction in the classroom, students could again, as described by O'Brien (1997), produce a written product without being critical or engaging in higher-order processes.

Even well-designed tasks need thorough, detailed instructions to ensure that when implemented, the performance assessments require students to engage in the skills intended in the assessments. In eight of the assessments, review team members expressed concern that the task could be higher-order or authentic, or it could be a low-level task depending on how the task was implemented in the classroom. If a teacher was intentional in the implementation, the task could require document analysis or proper research skills, but the tasks could also be completed based on prior knowledge, regurgitation of learned material, or based on student perceptions. Parke and Lane (2008) in a study of teacher-submitted performance tasks reported that the assessments provided could have been implemented at a lower cognitive level than indicated by the submitter depending on the teacher delivery and encouragement of higher-order thinking. Teacher knowledge, experience, and skills vary; thus, to ensure that performance assessments are

being implemented as designed and require students to engage in higher-order skills, detailed instructions for implementation need to be provided. Students also need clear, specific instructions with structures, such as rubrics, which specify the skills to be demonstrated and require students to engage in higher-order skills. Task design, instructions, and rubrics need to be tightly constructed with clear statements of purpose and clear procedures to ensure quality performance assessments that necessitate the higher-order thinking skills of the social studies and enable the benefits of performance assessments on teaching and learning (Chappuis et al., 2017; Khattri et al., 1998; Wiggins, 1998).

Once constructed, tasks, resources, and directions for students need to be critically analyzed to ensure that the wording and format do not reduce the cognitive level of the task. Teacher instructions need to provide guidance so that individual teachers do not present the task in such a way as to reduce the cognitive level of the task (S. G. Grant et al., 2004; McCann & McCann, 1992). Implementing research-based strategies for quality performance assessments, while time and resource-consuming, is a strategy divisions and teachers can draw on to assist in reviewing performance assessments. Divisions and teachers should engage in an honest and deliberate process of identifying the skills to be measured and analysis of whether the tasks designed require those skills. The use of expert review, and more deliberate field tests, in which samples of student work are analyzed to evaluate whether students are demonstrating higher-order thinking, should be used in this analysis (Brookhart, 2015; Khattri et al., 1998; Lane, 2014; Moon et al., 2005; Pecheone & Kahl 2014; Wren & Gareis, 2019).

As teachers and division leaders engage in a critical analysis of the performance assessments, another factor to attend to is the presence of bias in the wording or structure of the task. As Goldberg and Roswell (2001) assert, all performance assessments need close and critical

analysis of the effect of item language and format on student performance. Baker (1994) supported this assertion, arguing that students should not be scored lower due to a student response not matching either a widely accepted historical argument or the perspective of a particular teacher. Bias toward a particular response by students could be either in the structure of the prompt or in the selection of the documents provided to the student to answer the prompt. Three of the assessments had prompts that the review team agreed were leading students toward a particular direction for their argument, including IDMs that otherwise promoted higher-order skills such as document analysis and supporting an argument. Additionally, six assessments had a prompt that promoted document analysis and construction of an argument and higher-order skills but provided a set of documents or resources for students that only presented one perspective on the prompt and pushed students toward a particular response. Two other assessments lacked sufficient resources for students to adequately respond to the prompt, such as asking students about a minority population experience without including any sources produced by the group referenced. When the task is constructed in such a way as to not allow a student to construct their own argument based on diverse facts, it reduces the authenticity of the task. Historians, when answering historical questions, determine their own arguments and choose their own sources; thus, a task that on the surface seems to be an authentic, critical construction of an argument based on an analysis of the source may not actually require a student to employ and demonstrate those skills due to its construction and bias (S. G. Grant et al., 2004). Thus, divisions and teachers need to carefully analyze both questions and resources provided to students to ensure that students have the ability and comfort to answer the question in diverse ways.

Careful analysis of the task should also consider the ways in which the prompt or student products could lead to culturally insensitive outcomes or products that are offensive to some members of the community. Social studies teachers have repeatedly come under fire in the media for activities that teachers see as fun and hands-on or experiential but are perceived by others as offensive or inappropriate. This concern is heightened by current events and the national societal climate. Eight performance assessments in the study asked students to take on the role of historical figures or groups and write from or about those perspectives. The evaluation team argued the sources provided led students to write on behalf of or, in some cases, from the perspective of people from a different time and/or culture with superficial knowledge. The evaluation team was concerned that these assessments risked products or student work that may at worst be offensive to particular groups or may perpetuate misconceptions or stereotypes about people in the past, rather than promote learning and align with the standards. One of the more successful assessments had students use primary sources to write a speech to persuade people in a historic time period to pursue a particular course of action. This approach allowed students to analyze and use documents, engage in an authentic task of arguing for policy, and demonstrate knowledge of the time period without asking students to take on a persona that could be potentially insensitive. Division leaders and teachers need to carefully examine the prompts and resources provided to students for political and cultural biases. Expert or outside reviews by individuals with different perspectives may be a strategy that divisions could employ for additional support.

Constructing quality performance assessments takes time and training and, as seen in the interviews, school divisions have devoted considerable time and resources to developing the assessments shared in this study. The assessments in the study are generally well-constructed,

requiring authentic, higher-order thinking skills with deeper learning competencies. A careful review of the tasks and the instructions to ensure that students must engage in the intended skills to complete the assessment and to remove the potential for stereotyping or culturally insensitive responses will strengthen the quality of the assessments.

Recommendation 3: Documentation of the VQCT

In the summer of 2017, three years after the replacement of the SOL tests with LAAs, the VDOE provided divisions with the VQCT for Performance Assessments and began training divisions on the tool during the school year 2017-2018, revising the tool in 2019 (VDOE, 2017). The current VDOE guidelines on LAAs state that "the continued use of the Virginia Quality Criteria Tool during the development of performance assessments and/or revision of existing tasks is expected in order to ensure that all students have access to quality assessments aligned to the SOL" (VDOE, 2021a, p. 2). Thus, the VQCT is an important element of the VDOE's articulation of the direction of the policy reform to promote even implementation (Fusarelli, 2002; Weick, 1982). As an important means of guiding policy in a loosely coupled system to ensure common understandings and comparable use of performance assessments across 132 school divisions the VQCT needs to be articulated with detail (Weick, 1982). Currently, the VQCT document consists of a cover page that lists the seven criteria and one short paragraph about division use of the tool. The other five pages are filled with a four-column chart with the 17 sub-criteria in one column, and two empty columns for a score and a rationale for evaluators using the tool to make notes on an assessment. Given that the VQCT consists of seven criteria and 17 sub-criteria, several of which contain complex constructs that require definitions, five pages does not provide much room for explanations of the criteria to achieve consistent application of the criteria to support "comparability in rigor and quality across the state" (VDOE,

2019d, p. 1). This study recommends that the VDOE provide more extensive documentation and explanation of the VQCT to ensure both effective use of the tool and quality performance assessments across the Commonwealth.

Communicating the Role and Importance of the VQCT. The legislative intent of the replacement of the SOL with performance assessments was to "encourage the greater use of assessments, such as performance assessments, that may be used by teachers to improve instruction" which corresponds with the purpose of performance assessments to improve teaching and learning as described in the educational literature (VDOE, 2014, p. 2). Thus, the state LAA initiative is more than an assessment initiative: the goal is to use assessments to improve instruction, which depends on the effective implementation and incorporation of quality performance assessments. The VDOE relies on the VQCT to ensure that these important LAAs are quality assessments as the instructions for the "Implementation of Performance Assessments" in the "Guidelines for Local Alternative Assessments," is a three-sentence paragraph, consisting primarily of a sentence stating that divisions are required to use the VQCT to develop and revise performance assessments (VDOE, 2021a). Given the importance of the VQCT to the process of ensuring that "all students have access to quality assessments," the VDOE might need to more clearly document the role of each of the criteria in achieving that goal (VDOE, 2021a, p. 2).

Currently the cover page of the VQCT provides no explanation of the interconnectedness of the 17 sub-criteria nor the importance of each sub-criteria to a quality performance assessment. The paragraph on the cover page states, "the criteria may be considered in any order that suits the division's needs and purpose" (VDOE, 2019d, p. 1). As a result, a teacher or administrator could read that statement as using the criteria as suits their needs and view the attached chart containing the 17 sub-criteria as a checklist, prioritizing certain criteria over others

without understanding the necessity of each criterion in regard to the others. As previously discussed, clearly-written teacher and student instructions, appropriate resources (Criterion 5), and a clear, tightly-aligned rubric (Criterion 4) are essential to a performance assessment being implemented as authentic and deeper-learning based (Criterion 2 and Criterion 1). In addition, the specifications for scaffolding and accommodations (Criterion 6) and schedule (Criterion 7) also provide a consistent structure to ensure all students are engaging with quality assessments. Without this explanation to teachers and leaders across Virginia, divisions may, as seen in this study, omit the standards, accessibility plans, schedule or clear instructions and rubrics as they focus on the first three criteria and the nature of the task. These omissions reduce the overall quality of the assessments and decrease the likelihood of a "comparability of rigor and quality across the state" (VDOE, 2019d, p. 1).

The examples from this study have focused their attention on Criteria 1, 2, and 3, developing tasks that, if implemented as intended, would be authentic and higher-order skills based, but with less attention to Criteria 4, 5, 6, and 7. These assessments lack the structure needed to achieve the intended learning goals and the goal of performance assessments. Given the importance of the VQCT a better communication of the role and importance of each of the 17 sub-criteria in constructing a quality performance assessment may better support the VDOE's goal of all students having access to quality assessments.

Ambiguity of Language. Currently, the VQCT is available on the VDOE website with descriptions of each sub-criteria that are brief enough to fit within the cells of the table, but there is no detailed explanation of the meanings or intentions of those descriptions. As a result, the brevity allows for ambiguity and different interpretations. The review team, three of whom had attended various VDOE workshops involving the VQCT, debated various elements of the sub-

criteria as the team worked to achieve interrater reliability and a consistent application of the tool, an issue that is compounded when considering the number of teachers in the 132 school divisions in Virginia with varying degrees of familiarity with the VQCT and performance assessments. Additional documentation to supplement the VQCT that defined terms and explained the intent of each sub-criteria could create common meanings and promote more consistent use of the VQCT and thus quality performance assessments.

The first concern was the ambiguity about what evidence was sufficient or required for each sub-criterion. Sub-criterion 1C and Criterion 2 each have two bullets, and the question for the review team was whether the expectation was for assessments to meet both bullets or meet one bullet or the other. 1C reads that students should have the opportunity "to develop and demonstrate deeper-learning competencies" in the first bullet and goes on to read "the performance assessment may also provide opportunities for…life-ready competencies" or "technology-ready competencies" (VDOE, 2019d, p. 2). The wording suggests that deeper-learning opportunities must be present, but the VQCT did not clarify whether an assessment could get full credit for having life-ready skills, such as technology-related competencies, even if deeper learning was lacking. Criterion 2 is more ambiguous, stating that the assessment is authentic if it was "relevant to the real-world," and/or if it was asking students "to do work authentic to the discipline" (VDOE, 2019d, p. 3). The review team debated whether assessments needed to be both real-world and authentic to the discipline to receive full evidence, or if meeting one of the two categories was sufficient. The team decided that an assessment that met one of the two bullets would receive full credit since the description lacked an "and" to suggest both. Sub-Criteria 7C was more problematic in that two different measures were connected in one paragraph with no bullets or use of the words "and" or "or." The first sentence asks if the

assessment covers multiple days and if so, is there a schedule for instruction for each day, while the second sentence measures the existence of information about student prior learning and where the assessment connects to the prior learning. The combination of these two factors in one paragraph created discrepancy in the review team scores with the team in full agreement or one scorer off by only one point for only 68.2% of the scores and scores varying by more than one point in 18.1% of the scores. The review team members struggled to determine if the assessment had a schedule for multiple days but no connection to prior learning (or vice versa) should the assessment be awarded Partial, Little, or No Evidence since a central component, identification of prior learning, was missing. Similarly, 3B includes the phrase "one or more forms of language"; the team deliberated whether one form, written language, was sufficient to be "Full Evidence," since the criteria read "one or more" and only requires "may allow" various forms. The team went with the wording in the rubric of one being the minimum to receive full credit, meaning that the team decided that an argumentative essay in which students explained their thinking met the criteria. Sub-Criterion 6B created similar discussion because, while it specifically identifies the use of Universal Design for Learning (UDL), the wording in the VQCT is "such as through the application" of UDL. Since none of the assessments clearly demonstrated UDL and the wording was "such as," the evaluation team decided that the use of UDL was not required but provided as an example and did not reduce the rating based on the absence of UDL. For these criteria, the VQCT was ambiguous in terms of the quantity required to meet the expectations or requirements of the state, thus creating challenges for consistent use of the tool to ensure quality performance assessments across the state.

The review team also struggled with the language in the two sub-criteria in Criterion 6 and what each sub-criterion was intended to evaluate. Sub-Criterion 6A begins with the word

"accommodation" which in many K-12 settings implies special education accommodations, but then the text goes on to describe "appropriate supports to facilitate accessibility" which seems to describe a process of scaffolding for any learning of varying abilities (VDOE, 2019d, p. 5). Clarifying the meaning of 6A was further complicated by the wording of sub-criterion 6B. Both sub-criteria use the term "accessibility," but 6B goes on to specifically indicate "differentiating" the student process which more closely aligns with the differentiation for particular groups of students such as gifted students, students receiving special education services, or multiple-language learners. The review team decided, after comparing the wording of the two sub-criteria and trying to distinguish between them, that 6A meant that the assessment indicated appropriate and acceptable scaffolds for any struggling learner and 6B addressed specific differentiation strategies for designated students. While it is not clear from the VQCT that this is a correct assumption by the review team, the team was unclear how else to differentiate the intent of 6A from 6B and award separate scores. A clearer explanation of the meaning and intent of the two sub-criteria and how the two differ from one another is necessary to ensure that the structures VDOE expects to see in LAAs are being implemented because, as previously discussed, without specific instructions for scaffolding and differentiation LAAs cannot provide comparable rigor and quality for all learners of all abilities across the state.

The variation in the interpretation of the VQCT and application by users can also be seen in the review team scores for sub-criteria 3A and 6A. Similar to the debate over 3B with the phrase "one or more," 3A had less agreement among the review team due to the phrase "multiple means of accessing…academic language" (VDOE, 2019d, p. 2). The review team was in almost complete agreement 68.2% of the time but had scores varying by more than one 22.7% of the time. Some reviewers felt multiple types of written documents such as primary and secondary

sources constituted multiple means and academic language while other review team members felt that the intent of 3A was multiple forms such as graphs, maps, video, or websites. While all team members felt several written texts constituted Partial or Little Evidence, resulting in no zero scores, the scores varied from one to three 22.7% of the time. 6A had even less agreement as the review team interpreted the meaning and expectations for "appropriate supports or alternatives to facilitate accessibility" differently (VDOE, 2019d, p. 5). The team was only in complete agreement 22.7% of the time and this sub-criterion had the highest percent of scores varying by more than one at 36.4%. For some review team members, the existence of graphic organizers, guiding questions, or student choice in the means of conveying their learning offered to all students constituted scaffolding strategies while other review teams read the VQCT to mean a list of additional scaffolds beyond the structure of the assessment to be provided only to struggling learners. With this level of ambiguity and the variation in performance assessment expertise across the state, some teachers or divisions may award higher scores than warranted to assessments and lead to a reduction in the rigor and quality of assessments across the state.

The review team also debated the extent of evidence needed to achieve Full Evidence, a 3, versus Partial Evidence, a 2. Different teachers and school leaders may hold different standards for full or partial evidence depending on their experience, expertise, and perspectives. The literature shows that teachers may overstate the degree of evidence without clear guidelines. Both Spillane and Zeuli (1999) and Parke et al. (2006) compared teacher surveys to classroom practice and found that teachers reported using more higher-order thinking skills than researchers observed in the classroom; thus, teachers and divisions self-evaluating their own performance assessments may feel they are more authentic or higher-order than the assessments actually are without specific definitions and guidelines provided by the VDOE. Examples, exemplars, or a

written description of what constitutes full evidence and partial evidence for each sub-criteria would allow greater consistency of the application. Without consistency of definitions of Full and Partial Evidence and the reliance on each division to evaluate their own assessments, the VDOE cannot ensure comparable rigor and quality of the LAAs across the state.

The other source of ambiguity was in language and definitions of concepts. Some of the terms used in the criteria are defined differently within education, creating opportunities for different application of the VQCT. The research reveals that terms such as "real-world," as seen in Sub-Criterion 1C, or "authentic," as seen in Criterion 2, hold diverse meanings. Archbald and Newmann (1988) first described authentic assessments as meeting three criteria: disciplined inquiry, integration of knowledge, and value beyond evaluation or, as stated by Newmann, the "construction of knowledge through disciplined inquiry to produce discourse or performances that have meaning or value beyond success in school" (Newmann et al., 1998, p. 19). Wiggins (1989) then defined authentic assessment as tasks that replicated real-world challenges and performances of professional adults that required the posing of questions, solving of problems, and explanation of responses. These advocates used the term authentic assessment to stress the real-world, beyond the classroom focus of the assessment. Many researchers and proponents of performance assessments use the terms "performance assessment" and "authentic assessment" interchangeably, but Meyer (1992) argued that the two are different in that authentic assessments use a real-world context while performance assessments do not. Therefore, some educational researchers, and possibly teachers and administrators, perceived performance assessments as including authentic assessments as one category while others saw all performance assessments as inherently authentic (Frey & Schmitt, 2007). Darling-Hammond (2017) and others have defined performance assessments as those that require students to construct answers or produce products

in a context that emulates the conditions of real life and require students to apply knowledge and reasoning (Darling-Hammond, 2017; Darling-Hammond & Adamson, 2010; Stecher, 2010). A guide for California social studies teachers of suggested performance activities defined performance assessments as any "substantial activity in which student work is measured with a clear rubric" that showed the "level of student mastery of all the key elements established as the learning goals of the unit" (Kroesch, 2015, p. 1). With these different definitions of what performance assessments should require of students, teachers may interpret substantial activity in terms of time, length of the product, or other measures that do not require critical thinking. As Messick (1994) stated, "in the rush to move performance assessments forward one gets the impression that any product, performance or student-constructed answer serves the cause" (p. 14). The lack of common definitions and understanding of the characteristics of performance assessments complicate the work of teachers and researchers in developing, implementing, and evaluating these assessments (Stecher, 2010). Additionally, according to the research, creating understandings of what constitutes acceptable student demonstrations of these skills is challenging for teachers who lack experience in this work (Darling-Hammond & Ancess, 1996; Goldberg & Roswell, 2000; Khattri et al., 1995). Although the VDOE does provide brief definitions in the "Guideline for Implementation," the brevity of the definitions might allow teachers or administrators to interpret the definitions to include tasks that are not higher-order thinking. The VDOE defines performance assessments as those that "require students to perform a task or create a product that is typically scored using a rubric" (VDOE, 2021a, p. 2). Under this definition a factual poster project scored by a rubric is a performance assessment but does not require higher-order thinking or deeper learning competencies. The VDOE defines authentic assessments as tasks that "mirror those that might occur in a 'real-life' situation and/or are

authentic to the academic discipline" using the ambiguous terms of "real-life" and "authentic" in the definition (VDOE, 2021a, p. 2). In order to ensure that the criteria are employed consistently, the VDOE should clarify their definitions of the various constructs used in the VQCT and potentially provide exemplars to help build common understandings across the commonwealth to ensure that the desired goals and student outcomes are achieved (Messick, 1994).

While the literature reflected the debate over the definitions of "authentic" and "performance assessments," other terms in the VQCT may also be defined differently by different users such as "bias" and "cultural sensitivity" in Sub-Criterion 5C, "academic language" in sub-criterion 3A, "realistic" in 7A and 7B, and "accessible language" in 5B. As seen in Chapter 4 these four sub-criteria had lower levels of agreement among the review team in part because of how the reviewers defined and perceived the language of the sub-criteria. 7A and 7B both measure the realistic nature of the assessment with 7A assessing whether the resources and materials are realistic and accessible by teachers and 7B focused on the realistic nature of the duration of the assessment. The review team was in almost complete agreement 68.2% of the time for both sub-criteria and disagreed by more than one 22.7% of the time, only four sub-criteria had less agreement. Similarly, 5B measured if the student task, prompt directions and resources were written in "accessible language appropriate to the grade level" and the review teams was only in almost complete agreement for 54.5% of the scores and had 22.7% of the scores varying by more than one. Sub-Criteria 5C presented a different challenge as previously discussed in Recommendation 2 due to the potential for bias and culturally-insensitive responses. The review team was only in almost complete agreement for 59.1% of the scores and had the second highest percentage of scores varying by more than one at 31.8%. Members of the review team less experienced with middle school students often read the assessment as intended and

assumed optimal student outcomes, scoring the assessments Full or Partial Evidence, while those reviewers with extensive middle school experience immediately thought of the myriad of ways students could answer inappropriately and scored the assessment Little or No Evidence. Given the distributed nature of this policy across 132 school divisions with teachers and division leaders of varying experiences, backgrounds and perspectives, users of the VQCT could interpret the language of the tool and its application differently resulting in varied quality of assessments across the state running counter to VDOE's objectives. Providing elaboration on the intent and purpose for these sub-criteria and the pitfalls the VDOE seeks to prevent, as well as providing examples and exemplars of what is realistic but rigorous, could help promote common understandings and use of the VQCT across the state to ensure all students have access to quality performance assessments.

As seen in this study, school divisions that have invested in performance assessment development and are seen as successful, still have varying degrees of comfort and familiarity with the VQCT, with one division reporting using a simplified evaluation tool and two divisions still working to build capacity on the VQCT. These responses from the study participants indicate that, as written, the VQCT is not easily incorporated into practice by teachers and division leaders, yet it is an important piece of VDOE's means to articulate and shape the implementation of the performance assessment policy to provide quality performance assessments to all students (Weick, 1982). Additional documentation to include definitions, explanations, and exemplars of the sub-criteria that divisions can refer to in order to expand their understanding of the application of the VQCT may contribute to increased and more reliable usage of the tool in the development and revision of LAAs. With a clearer understanding of the 17 sub-criteria and the role and significance of each, divisions may be able to better use the

VQCT to develop quality performance assessments that can benefit all students across Virginia and better align with the intended goal of the VDOE policy.

Summary of the Recommendations for Policy and Practice

VDOE's grassroots policy of allowing school divisions to develop LAAs to meet state accountability reflects the emphasis by ESSA and Virginia on 21st Century skills and the resulting demand for better assessments that more effectively reflect student abilities to think critically, communicate, and engage in real-world competencies (Darling-Hammond & Adamson, 2010; Darling-Hammond et al., 2014) The resulting VDOE emphasis on performance assessments allows students to engage in and demonstrate higher-order thinking and deeper-learning competencies (Baron, 1996; Darling-Hammond & Adamson, 2016; Foote, 2015). In addition, the use of performance assessments contributes to improved teaching and learning with more student time engaging in problem-solving, writing, and deeper learning skills (Darling-Hammond & Adamson, 2016; Khattri et al., 1998; Parke & Lane, 2008; Stosich et al., 2018). To meet these goals of improved teaching and learning, school divisions and the VDOE need to ensure that, as stated in the VQCT, "all students have access to quality assessments" that are comparable in rigor across the state (VDOE, 2019d, p. 1). To ensure high quality performance assessments which benefit learners across 132 school divisions in Virginia, divisions and the VDOE need to engage in high-quality, on-going professional development so all teachers understand the purpose and nature of high-quality performance assessments and how to develop and use them. With this professional development, assessment developers need to carefully analyze and review all performance assessments to ensure the task requires students to engage in the intended authentic, deeper-learning competencies and is structured to be consistently implemented without bias across any classroom and teacher. The VDOE needs to aid in this

process by clearly documenting the VQCT with definitions and exemplars to ensure proper and consistent application of the tool. These three practices in tandem could promote high-quality performance assessments across Virginia that will achieve the Virginia General Assembly and VDOE goals of ensuring all students have access to high quality assessments that ensure all students are making adequate academic progress (VDOE, 2019c; VDOE, 2021a). In a loosely coupled system where individual teachers and divisions are responsible for developing and implementing this policy with little oversight or accountability measure from the state, the VDOE needs to frequently and clearly articulate the goals and direction of the reform, providing details to provide common understandings and practice to promote even implementation across the state (Fusarelli, 2002; Weick, 1976, 1982).

Recommendations for Future Research

This study identifies the strategies of successful divisions in responding to a grassroots implementation of a state policy using locally developed performance assessments and then evaluates the quality of the resulting assessments. While providing insights into the processes divisions used and the extent to which the policy objectives of quality assessments were being met, the study also gained insights into strategies that the state and other divisions could employ to increase capacity around performance assessments and promote quality assessments across the state. As an exploratory study with a small, non-representative sample, the findings from this study are limited, but the finding do suggest possibilities for future research. A more extensive study of this type, an examination of student products, and an examination of teacher perceptions all could provide more thorough data on the degree to which state policy objectives are being met. An extensive study of the processes that divisions engage in could reveal a correlation between division practice and the results of quality assessments. A reliability and validity study

of the VQCT could provide guidance on the construction of those quality performance assessments.

Recommendation 1: More Extensive Review of Division LAAs

The goal of this study is to identify strategies successful divisions have utilized in developing LAAs and to provide insights for other divisions, but it does not accurately portray the implementation of the grassroots LAA policy across Virginia. Since the VDOE intends to expand the use of LAAs to more courses and subjects, a more thorough investigation of the outcomes and success of the initial policy would be beneficial before expanding the initiative. Therefore, the first recommendation for future research is a more extensive study of the types and quality of LAAs being used by Virginia school divisions to provide a more accurate depiction of how and how well divisions are responding to the autonomy given to them by the state.

The findings of this exploratory study are limited due to the small size and the nature of the sample of the study, including only 12 divisions out of the 132 (9.1%) of the divisions in Virginia. However, the sample did represent seven of the eight geographically based superintendent's regions and a mix of divisions of different size and per pupil spending. In addition, these divisions were recommended by organizations and individuals involved in the Virginia performance assessment initiative, thus these are divisions known for being actively engaged in training events and demonstrating a commitment to this initiative. These 12 divisions have invested considerable time in professional development and continue to structure professional development and PLC time around further growing capacity, have incorporated the VQCT into their practice, and have started moving toward the common rubric, the limitations of

the sample in this study means these findings may not be indicative of the degree to which all divisions in Virginia respond to the state performance assessment initiative.

The sample of this study is further limited in terms of evaluating the types and quality of performance assessments being used across Virginia for state accountability. The study asked each participating division to share one assessment per course, USI and USII, but each division uses from one to six assessments per course. In six divisions schools or teachers developed their own assessments, resulting in an unknown number of different assessments being used as LAAs. One division did not share any assessments. As a result of these factors, the assessments submitted to the study represent only a fraction of the assessments in a single course. The 22 assessments reviewed by the team represent a tiny fraction of the assessments being used by these divisions as LAAs. Still, these 22 assessments by successful divisions who have invested considerable time and resources in developing, evaluating, and revising the assessments, scored from 18-46 out of 51 possible points on the VQCT with a mean of 29.64, which is 58% of the total points, suggesting the need for improvements in the quality of many of the existing LAAs. Five of these divisions submitted assessments that consisted of students writing brochures, letters, or other similar products that were predominantly the reframing of taught content or relaying of researched information rather than the deeper-learning competencies and authentic tasks measured in the VQCT and desired by the proponents of performance assessments. The range of scores on the VQCT and the inclusion of lower-order skill assessments within the study raises questions of the types and quality of assessments being used in other divisions not recommended for this study.

As the VDOE expands this policy to high school social studies courses, and possibly other courses, more research into the type and quality of the current LAAs in use across the state

and the division preparations to meet the expectation of quality performance assessments would provide more information about how well divisions are meeting the intent of the VDOE policy and are ensuring students are meeting state standards.

Recommendation 2: Correlation Between Practices and the Quality of Performance Assessments

Given that the goal of the study is to provide insights and suggestions for the VDOE and other divisions on how to navigate the autonomy given to divisions, further research could be conducted on the correlation between the strategies used by divisions to ensure quality performance assessments and the actual quality of the assessments the divisions developed.

While the interview protocol for this study asked divisions leaders to describe the research-based strategies used in developing their LAAs, the data were not sufficient to establish correlations. First, the interview protocol did not ask nor require that respondents carefully review their records of past meetings or workshops to specifically list out their experiences, instead relying on participant memory of the steps taken to develop LAAs. Besides the lapse of time, gathering these details was further hampered by the turnover in the social studies coordinator position, 7 of the 12 participants had come to this role after the performance assessment process had begun, five of those had been in the position for one to three years, and one person said, “this is my 6th year in this role, so I am still new.” Because of this turnover, interviewees did not necessarily know all of the steps their predecessor had engaged in prior to their arrival. Even interviewees who had been in their position since 2014 could not remember specifics from when the initiative began 7 years prior. A study focused on getting more descriptive, researched, and accurate details on the exact steps and timelines of assessment development could strengthen the findings of this research. An examination of documentation of

past meetings, committee proceedings, and workshop content could provide enough data to conduct a correlation between which strategies have the strongest correlation with high quality performance assessments.

The results of a correlation study also could provide more specific direction and course of action to other divisions engaging in a performance assessment reform. Given the significant investment of time and resources this performance assessment initiative has required of the successful divisions, being able to identify the more effective strategies would allow other divisions to better focus their efforts and maximize their resource use.

Recommendation 3: Examination of Student Products from the LAAs

As discussed in the findings of this study, performance assessments can appear to be authentic tasks requiring higher-order thinking and deeper-learning competencies but may not result in students engaging in those desired skills (Cumming & Maxwell, 1999; Linn & Baker, 1996). To obtain a better measure of the quality of performance assessments being developed by divisions, and thus the degree to which students are meeting state standards, a third recommendation for future research is to evaluate student responses or products to LAAs and teacher scores of those products.

An evaluation of student responses and teacher scores would reveal whether the assessment as constructed and implemented is a quality assessment meeting Criteria 1 and 2 of the VQCT. While the literature recommends a review of student work samples when piloting or field testing a performance assessment, none of the divisions in the study described such a deliberate, data-driven pilot. A future study could use the pilot methodology and examine student work to analyze whether or not the assessment s and rubrics were producing the desired processes by students (Brookhart, 2015; Khattri et al., 1998; Lane, 2014; Pecheone & Kahl,

2014). There is a precedent for such an evaluation as O'Brien (1997) evaluated a Kansas state-wide performance assessment initiative by collecting student work samples and found that students failed to actually engage in the skills desired by the assessment, instead generating factual reports of information. A study could go further and analyze teacher scores of student work to determine whether teachers are requiring students to demonstrate the authentic, higher-order thinking skills to score well on the assessment or are allowing students to respond with a reframing of learned content and still score well. This would provide better analysis of the extent teachers are upholding the intent of the performance assessments. Since divisions are required to retain samples of student responses and a record of student scores for a potential VDOE desk review of the division Balanced Assessment Plans and LAAs, these materials should be accessible for future research (VDOE, 2021a).

While divisions may be constructing quality performance assessments, or assessments that appear to be quality assessments, the implementation of those assessments may result in a loss of quality and the possibility of students not demonstrating the intended skills of the assessment. Thus, a study of the quality of student responses and teacher scores of those responses would be another indicator of the quality of LAAs being developed by divisions for state accountability of student learning.

Recommendation 4: Examination of Teacher Perceptions and Experiences with LAAs

This study relied on the perspectives of the division-level administrator supervising social studies instruction. The extent to which the intent and goals of the division-level social studies supervisor shaped and informed teacher practice was not within the scope of the study. Thus, the fourth recommendation for future research would be a study of teacher experiences and perceptions of the LAAs. This would provide more insight into the steps divisions have taken to

prepare for performance assessments and the success of those efforts. While a division-level social studies specialist may feel very invested in the reform and professional developments, teachers who have a variety of responsibilities are less focused on the reform and may not feel as prepared or as well-versed in performance assessments as divisions leaders perceive.

References in numerous interviews suggest that the level of understanding of and emphasis on performance assessments conveyed by the division leader was not shared throughout the division. Half of the division leaders alluded to experienced teachers who still had not embraced or not fully grasped the role and purpose of performance assessments, and newer teachers who had less familiarity with performance assessments. Prior research on state performance initiatives further supports the need for research into teacher perceptions. Stecher and Mitchell (1995) in Vermont; Koretz, Barron et al. (1996) in Kentucky; and Koretz, Mitchell, et al. (1996) in Maryland surveyed teachers implementing state performance assessment initiatives. Each study demonstrates that even after training and implementation, teachers still lacked common understanding of the higher-order constructs being measured in the assessments. Khattri et al. (1998) interviewed teachers and gathered samples of teacher-developed performance assessments used for the Maryland MSPAP initiative and also found a lack of a common understanding of performance assessments both in teacher interviews and in the assessments teachers used in the classroom. Parke et al. (2004) and Parke and Lane (2006) also surveyed Maryland teachers and analyzed classroom materials; this research shows teacher perception of the quality of their activities outpaced the actual quality of the performance assessments teachers were using. Thus, a similar study to these in Virginia could investigate how evenly and effectively the VDOE emphasis on the quality and the role of performance assessments in instruction is filtering to the teachers in the classroom.

Identifying a potential disconnect between division-leader interest and teacher engagement could affect the implementation of the LAAs since the research argues teachers must understand the purpose and usage of performance assessments to integrate the desired skills and understanding into curriculum and instruction (Khattri et al., 1998; Wiggins, 1998; Wren & Gareis, 2019). Without this understanding many teacher-created assessments become hands-on activities inserted into existing instruction rather than effective evaluation of student skills (Firestone et al., 1998; Goldberg & Roswell, 2000; Gong & Reichy, 1996; Messick, 1994). One challenge may be that teachers often lack clear understandings of performance assessments given the range of forms performance assessments can take from on-demand tasks such as constructed-response or stand-alone tasks to extended long-term assessments such as curriculum-embedded tasks or complex projects (Brookhart, 2015; Khattri et al., 1998; Wren & Gareis, 2019). Within those categories there is, as seen in this study, tremendous variety of assessments from DBQs and IDMs to research presentations and policy arguments as Wren and Gareis (2019) list 21 types of performance assessments. Since performance assessments are an evolving field, teachers and administrators may benefit from a taxonomy of performance assessments with categories and examples to deepen their understanding of what constitutes ad performance assessment, thus contributing to more meaningful incorporation of performance assessments in classroom instruction.

While the lack of teacher understanding and engagement with the reform has more effect on the quality of performance assessments in divisions where schools or teachers develop their own LAAs, even divisions with common LAAs still leave individual teachers to administer the assessments to their students. Individual teachers may deviate from, supplement, or overly direct students through portions of the performance assessment which would affect student

performance and the quality of the performance assessment. Thus, a study of teacher perceptions, understanding, and practices surrounding performance assessments may provide more insight into the actual structure of the assessments as given to students, the quality of the performance assessments in practice, and the success of the divisions in meeting the state policy expectations.

Recommendation 5: Reliability and Validity Study of the VQCT

Since the VDOE is requiring divisions to use the VQCT in the development and revisions of LAAs and views this tool as a means to "support comparability in rigor and quality" in performance assessments across the state, the VQCT needs to be a reliable, valid assessment in order to ensure that these goals are being met (VDOE, 2019d, p. 1). There is no documentation, research, or formal evaluation of the VQCT. The concerns of both division leaders and the review team about the ambiguity of the language, the challenges of using the VQCT, and the difficulty of obtaining consistent results across all users, demonstrate the need for an evaluation of the VQCT. Thus, another possibility for future research would be a reliability and validity study of the VQCT.

As discussed previously in this chapter, the lack of documentation and the brief statement of each of the 17 sub-criteria allows for different interpretations of the criteria, the concepts embedded in the criteria and what constitutes full or partial evidence affects the reliability and validity of the VQCT. The brevity and possibility for different interpretations of the criteria may affect the validity and reliability of the tool when evaluating any given assessment. With 132 school divisions and multiple teachers per division using the VQCT to ensure "all students have access to quality performance assessments," the VQCT's ambiguous language may allow for considerable variation of assessment scores across time and scorers (VDOE, 2019d, p. 1). The review team for this study had varying levels of agreement over the 17 sub-criteria as reviewers

differed in their interpretations of the VQCT and how to apply it to an assessment given their perceptions of authentic, realistic, bias, as well as how to apply the sub-criteria that include a variety of factors. Without clear descriptions and exemplars of the constructs in the criteria, what constitutes full or partial evidence, the intent of multi-part criteria, and the difference between 6A and 6B, teachers and administrators may score assessments differently depending on the day, familiarity with the particular content of the assessment, or type of task in the LAAs. With 132 divisions using the tool to certify comparable rigor and quality, there needs to be confidence in a given VQCT score on a particular LAA.

The lack of documentation and explanation of the role and importance of each sub-criteria might also affect the validity, because depending on the user's understanding of the criteria the application of the VQCT may not accurately measure the quality of the assessment. One concern of the review team is that an assessment could, "check all the boxes and score high but there are fundamental flaws in the assessment." An assessment could identify standards (1A), ask students to analyze sources (1B), have students use resources and write (3), contain a rubric (4), directions and resources (5A and 5C), a list of accommodations and scaffolds (6), and a schedule (7), but still be a weak or poor assessment. The team looked at several assessments that earned high scores in all of these categories and scored high but contained a flawed prompt. Examples of flaws include prompts that are too esoteric and complex for middle school students to adequately answer, are too biased, ask students to respond from the perspective of a group that the student could never adequately understand, or is too broad and all-encompassing. Thus, the team agreed that the overall task is flawed even though the assessment scored high on the VQCT by meeting the individual components. As previously discussed, this is supported by literature analyzing tasks which appear higher order but actually are a low-level writing task.

Fundamentally flawed assessments scoring high because so many separate, distinct criteria were present while the main purpose of performance assessments, authentic deeper learning (Criterion 2 and Sub-Criterion 1C) is not met, raises a question about the usefulness of the VQCT to accurately demonstrate the quality of the LAAs.

Another potential area of concern would be the overlap of sub-criteria content and the contradiction of one sub-criterion by another. As noted in the findings, there is considerable overlap between assessment scores on sub-criteria 1B & 1C and Criterion 2. The authentic skills of the social studies (Criterion 2) include complex thinking, such as analysis, evaluation, or synthesis (Sub-Criterion 1B) and deeper learning, such as critical thinking (Sub-Criterion 1C). Thus, assessments are repeatedly awarded high scores for the same elements of the assessment or repeatedly penalized with low scores for the same flaw in the assessment. Admittedly, in light of the review team's frustration at poor assessments scoring overall high scores, added weight and emphasis to what is the foundation and purpose of performance assessments may be important. This can be done by measuring key components such as authentic, higher-order, deeper-learning competencies under separate sub-criteria. The findings in this study also suggest a possible contradiction between Sub-Criteria 4A and 4B with Sub-Criterion 4C. As previously discussed, assessments that scored high on 4A and 4B for clear, tightly aligned rubrics score poorly on 4C for a lack of transferability and vice versa. Further research into the overlap and contradictions of sub-criteria identified in this study may identify concerns about the VQCT's validity or provide opportunities to provide greater clarity and accuracy in the use of the VQCT in the development and revision of quality performance assessments.

The Code of Virginia states that the LAAs should ensure students are making adequate academic progress; the VDOE requires individual school divisions to independently use the

VQCT to evaluate those LAAs to ensure all students have access to quality performance assessments. Given the reliance of the VDOE and school divisions on the VQCT to develop, evaluate, and ensure quality performance assessments to improve teaching and learning, research into the reliability and validity of the tool would provide confidence in the use of this tool for the intended purposes.

Conclusion

In 2014 the Virginia General Assembly removed a single, end-of-course, multiple-choice assessment from 5 elementary and middle school courses to be replaced with local alternative assessments, shifting from a more tightly coupled to a more loosely coupled system for accountability of student learning (Fusarelli, 2002). The VDOE then implemented this legislative mandate through a grassroots policy that allowed school divisions to develop the new alternative assessments locally. A sample of divisions responded to this level of autonomy by engaging in on-going professional development for division leaders and teachers provided by both the VDOE and by outside consultants. Based on these trainings, most division assessment plans consist solely of a set of performance assessments developed predominately by teachers with the use of a template and the VDOE's VQCT for Performance Assessments. The majority of the LAAs divisions developed require students to engage with a set of sources and content and construct a written argument based on the evidence. While the divisions have focused on writing authentic tasks with deeper-learning skills and the use of language by students, the documentation for the implementation of the assessments in terms of schedule, accommodations, and tightly aligned rubrics is less consistent across the divisions, and some contain potential for biased or culturally insensitive responses.

Given the limited sample of this study, the experiences of these divisions may not be reflective of all divisions and, especially with the grassroots nature of the reform, there may be uneven implementation of the policy across Virginia. With the stated goal of the VDOE initiative to "ensure that all students have access to quality assessments aligned to the SOL," a more extensive review of division LAAs should be conducted as well as a review of student products from these LAAs to determine if state skill standards are being adequately met across all 132 divisions (VDOE, 2021a, p. 2). Although the VDOE intends for the VQCT to ensure comparable quality assessments across the divisions, the Tool lacks documentation or a reliability and validity study which may hinder its usefulness in meeting this goal. Additionally, since this study relied on division leaders self-reporting on their practice, an examination of teacher perceptions and experiences as well as a more thorough accounting of division professional development offerings may provide a more accurate measure of how well the state goals of divisions developing and using performance assessments to improve teaching and learning are being met.

Despite the limitations of the study, the experiences of these successful divisions provide guidance as the VDOE continues, and expands, this grassroots initiative which seeks to measure student progress toward state standards and promote improved teaching and learning. Based on the experiences of these divisions, for a grassroots policy to be successful, all professionals involved in its implementation should engage in quality, on-going, state-supported professional development. Additionally, given that the state is not reviewing the LAAs, local development teams need to analyze their LAAs thoughtfully and deliberately to ensure that the tasks require students to demonstrate the state standards of authentic, deeper-learning competencies and avoid bias or cultural insensitivities in the prompt, resources, or student responses. To assist divisions in developing quality performance assessments that meet state standards, the VDOE needs to

continue to clearly articulate the initiative through increased documentation of the VQCT to ensure equitable and comparable performance assessments across the state despite the level of autonomy granted to local divisions.

The division leaders in this study responded to this level of autonomy with enthusiasm and embraced the performance assessment initiative. In the interviews division leaders were passionate about the transition to performance assessments and being able to promote more meaningful learning outcomes for students. The enthusiasm of the division leaders was apparent as the leaders eagerly shared their process of developing their LAAs and the ways the division was supporting teachers through the process. The interviewees appreciated the autonomy given by the VDOE as it allowed them to draw on existing strengths from previous division initiatives and professional relationships to develop assessments that met the needs of their teachers and students. Additionally, the division leaders felt teachers had greater ownership over the LAAs as teachers contributed to the development and revision process. Even with their commitment of time, resources, and energy in successfully developing LAAs, each division leader eagerly shared their current and future plans to continue to increase teacher capacity and to improve their existing LAAs to promote more meaningful teaching and learning. Just as the work of Khattri et al. (1998) asserts that teacher appropriation of performance assessments is necessary for an assessment reform initiative to be successful, maybe the success of a grassroots state initiative requires the appropriation of local division leaders to be successful.

With the autonomy of a grassroots policy these 12 divisions have developed quality performance assessments where students engage in deeper-learning skills authentic to the social studies, but this success requires quality, on-going professional development and a deliberate analysis of the performance assessments to ensure unbiased authentic tasks.

REFERENCES

Abbott, A. L. (2016). Locally developed performance assessments: One state's decision to supplant standardized tests with alternative measures. *Journal of Organizational and Educational Leadership, 2*(1), Article 5. https://digitalcommons.gardner-webb.edu/joel/vol2/iss1/5

Abbott, A. L., & Wren, D. G. (2016). Using performance task data to improve instruction. *The Clearing House, 89*(1), 38-45. https//doi.org/10.1080/00098655.2016.1138924

Abrams, L. M., Madaus, G. F, & Pedulla, J. J. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice, 42*, 18-29. https//doi.org/10.1207/s15430421tip4201_4

Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school.* National Association of Secondary School Principals.

Baker, E. L. (1994). Learning-based assessments of history understanding. *Educational Psychologist, 29,* 97-106. https//doi.org/10.1207/s15326985ep2902_5

Bandalos, D. L. (2004). Can a teacher-led state assessment system work? *Educational Measurement: Issues and Practice,* 33-40. https//doi.org/10.1111/j.1745-3992.2004.tb00157.x

Baron, J. B. (1996). Developing performance-based student assessments: The Connecticut experience. In Baron, J. B. & Wolf, D. P. (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 166-191). The University of Chicago Press.

Baxter, G., & Glaser, N. (1998, Fall). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, 37-45. https//doi.org/10.1111/j.1745-3992.1998.tb00627.x

Biemer, L. (1993). Authentic assessment. *Educational Leadership, 50*(8), 81-82. https://www.ascd.org/el/articles/-authentic-assessment

Brookhart, S. M. (2005). District assessments used in Nebraska's school-based, teacher-led assessment and reporting system (STARS). *Educational Measurement: Issues and Practice, 24*(2), 14-21. https//doi.org/10.1111/j.1745-3992.2005.00007.x

Brookhart, S. M. (2015). *Performance assessment: Showing what students know and can do.* Learning Sciences International.

Brown-Kovacic, C. (1998). *Construct validity of a math performance assessment project* (Publication No. 9911740) [Doctoral dissertation, University of New Mexico]. ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/construct-validity-math-performance-assessment/docview/304440261/se-

CAST. (2018). *Universal design for learning guidelines version 2.2.* http://udlguidelines.cast.org

Center for Collaborative Education. (2013). *Quality performance assessment: A guide for schools and districts.* Center for Collaborative Education.

Chappuis, S., Commodore, C., & Stiggins, R. (2017). *Balanced assessment systems: Leadership quality and the role of classroom assessment.* Corwin.

College Board. (2023). *Exam Development.* https://apcentral.collegeboard.org/courses/how-ap-develops-courses-and-exams/exam-development

Conley, D. T. (2015). A new era for educational assessment. *Education Policy Analysis Archives, 23*(8), 1-36. https://doi.org/10.14507/epaa.v23.1983.2-36

Crehan, K. D. (2001). An investigation of the validity of scores on locally developed performance measures in a school assessment program. *Educational and Psychological Measurement 61*(5), 841-848. https://doi.org/10.1177/00131640121971554

Creswell, J. W., & Guetterman, T. C. (2019). *Educational research: Planning, conducting, and evaluating qualitative and quantitative research* (6th ed.). Pearson.

Cumming, J. J., & Maxwell, G. S. (1999). Contextualising authentic assessment. *Assessment in Education: Principles, Policy and Practice, 6,* 177-194. https://doi.org/10.1080/09695949992865

Daniel, J. (2012). *Sampling essentials: Practical guidelines for making sampling choices.* Sage.

Dappen, L., & Isernhagen, J. C. (2005). Nebraska STARS: Assessment for learning. *Planning and Changing 36*(3/4), 147-156.

Darling-Hammond, L. (2017). *Developing and measuring higher order skills: Models for state performance assessment systems*. Learning Policy Institute and Council of Chief State School Officers.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning.* Stanford University, Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., & Ancess, J. (1996). Authentic assessment and school development. In Baron, J. B. & Wolf, D. P. (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 52-83). The University of Chicago Press.

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for college and career readiness: Developing a new paradigm. *Education Policy Analysis Archives*, *22*(86), 1-34. https://doi.org/10.14507/epaa.v22n86.2014

DeWitt, S. W., Patterson, N., Blankenship, W., Blevins, B., DiCamillo, L., Gerwin, D., Gradwell, J. M., Gunn, J., Maddox, L., Salinas, C., Save, J., Stoddard, J., & Sullivan, C. C. (2013). The lower-order expectations of high-stakes tests: A four-state analysis of social studies standards and test alignment. *Theory and Research in Social Education, 41,* 382-427. https://doi.org/10.1080/00933104.2013.787031

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*, 289-303. https://doi.org/10.1207/s15324818ame0404_3

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20,* 95-113. https://doi.org/10.3102/01623737020002095

Foote, M. (2005). *The New York performance standards consortium: College performance study*. New York Performance Standards Consortium

Frey, B. B., & Schmitt, V. L. (2007). Coming to terms with classroom assessment. *Journal of Advanced Academia, 18,* 402-423. https://doi.org/10.4219/jaa-2007-495

Frey, B. B., Schmitt, V. L., & Allen, J. P. (2012). Defining classroom authentic classroom assessment, *Practical Assessment Research and Evaluation, 17*(2), 1-17. https://files.eric.ed.gov/fulltext/EJ977576.pdf

Fusarelli, L. D. (2002). Tightly coupled policy in loosely coupled systems: Institutional capacity and organizational change. *Journal of Educational Administration, 40*(6), 561-575. https://doi.org/10.1108/09578230210446045

Gao, X., Shavelson, R. J., & Baxter, G. P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. *Applied Measurement in Education, 74*(4), 323-342. https://doi.org/10.1207/s15324818ame0704_4

Gareis, C. R. (2017, January 11-12). *A crosswalk of Virginia-specific quality criteria.* Performance-based assessments: A hands-on workshop, Richmond, VA.

Gareis, C. R. & Grant, L. W. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning.* Routledge.

Gerwin, D., & Visone, F. (2006). The freedom to teach: Contrasting history teaching in elective and state-tested courses. *Theory and Research in Social Education, 3*(2), 259-282. https://doi.org/10.1080/00933104.2006.10473307

Goldberg, G. L., & Roswell, B. S. (2000). From perception to practice: The impact of teachers' scoring experience on performance-based instruction and classroom assessment. *Educational Assessment, 6*(4), 257-290. https://doi.org/10.1207/S15326977EA0604_3

Goldberg, G. L., & Roswell, B. S. (2001). Are multiple measures meaningful? Lessons from a statewide performance assessment. *Applied Measures in Education, 14*(2), 125-150. https://doi.org/10.1207/S15324818AME1402_2

Gong, B., & Reidy, E. F. (1996). Assessment and accountability in Kentucky's school reform. In Baron, J. B. & Wolf, D. P. (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 215-233). The University of Chicago Press.

Grant, L., & Gareis, C. (2015). *Teacher-made assessments: How to connect curriculum, instruction, and student learning.* Routledge.

Grant, S. G., Gradwell, J. M., & Cimbricz, S. K. (2004). A question of authenticity: The document-based questions as an assessment of students. *Journal of Curriculum and Supervision, 19*(4), 309-337. https://eric.ed.gov/?id=EJ732631

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P.A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67-86. https://doi.org/10.1007/BF02504676

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Why, whats and whithers. *The Phi Delta Kappan, 70*(9), 683-687. https://www.jstor.org/stable/20404000

Hackmann, D. G., Malin, J. R., & Bragg, D. D. (2019). An analysis of college and career readiness emphasis in ESSA state accountability. *Education Policy Analysis Archives, 27*(160), 1-27. https//doi.org/10.14507/epaa.27.4441

Heckathorn, D. D. (2002). Respondent driving sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems 49*(1), 11-34. https://doi.org/10.1525/sp.2002.49.1.11

Hewlett Foundation. (2013). Deeper learning competencies. https://hewlett.org/wp-content/uploads/2016/08/Deeper_Learning_Defined__April_2013.pdf

Hong, H., & Hamot, G. E. (2015). The associations of teacher professional characteristics, school environmental factors, and state testing policy on social studies educators' instructional authority. *Journal of Social Studies Research, 39*(4), 225-241. https://doi.org/10.1016/j.jssr.2015.06.009

Honig, B., & Alexander, F. (1996). Rewriting the tests: Lessons from the California state assessment system. In Baron, J. B. & Wolf, D. P. (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 143-165). The University of Chicago Press.

Kan, A., & Bulut, O. (2014). Crossed random-effect modeling: examining the effects of teacher experience and rubric use in performance assessments. *Eurasian Journal of Educational Research*, 57, 1-28. https://doi.org/10.14689/ejer.2014.57.4

Khattri, N., Kane, M. B., & Reeve, A. L. (1995). How performance assessments affect teaching and learning, *Educational Learning, 53*(3), 80-83. https://www.ascd.org/el/articles/-how-performance-assessments-affect-teaching-and-learning

Khattri, N., Reeve, A. L., & Kane, M. B. (1998). *Principles and practices of performance assessment.* Routledge.

Koretz, D., & Barron, S. (1998). The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS). RAND Corporation. https://www.rand.org/pubs/monograph_reports/MR1014.html

Koretz, D., Barron, S., Mitchell, M., & Stecher, B. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KIRIS).* RAND Corporation.

Koretz, D., Mitchell, K., Barron, S. & Keith, S. (1996). *Final Report: Perceived effects of the Maryland performance assessment program*. National Center for Research on Evaluation, Standards and Student Testing (CRESST)/RAND Institute on Education and Training.

Kroesch, G. (2015). Social studies curriculum units and measurable performance activities for high school. *Social Studies Review 54*, 74-91.

https://casocialstudies.org/resources/Documents/CCSS_Social%20Studies%20Review_201516.pdf

Lane, S. (2014). Performance assessments: The state of the art. In Darling-Hammond, L., & Adamson, F. (Eds.) *Beyond the bubble test: How performance assessments support 21st century learning.* Jossey-Bass.

Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment, 8*(4), 279-315. https://doi.org/10.1207/S15326977EA0804_1

Learning Forward (2022). *Standards for professional learning*. https://standards.learningforward.org/

Linn, R. C., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In Baron, J. B. & Wolf, D. P. (Eds.), *Performance-based student assessment: Challenges and possibilities* (pp. 84-103). The University of Chicago Press.

Maddox, L. E., & Saye, J. W. (2017). Using hybrid assessments to develop civic competency in history. *The Social Studies, 108*(2), 55-71. https://doi.org/10.1080/00377996.2017.1283288

Marion, S., & Leather, P. (2015). Assessment and accountability to support meaningful learning. *Education Policy Analysis Archives, 23*(9), 1-13. https://doi.org/10.14507/epaa.v23.1984

McBee, M. M., & Barnes, L. L. (1998). Generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education, 11*(2), 179-194. https://doi.org/10.1207/s15324818ame1102_4

McCann, F. C. & McCann, C. J. (1992). Authentic evaluation in history. *OAH Magazine of History, 6*(4), 6-9. http://www.jstor.org/stable/25154078

McTighe, J. (2016). *Performance task review criteria*. https://jaymctighe.com/downloads/Task-Review-Criteria-2.pdf

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

Meyer, C. (1992). What's the difference between authentic and performance assessment? *Educational Leadership, 49,* 41-42. https://files.ascd.org/staticfiles/ascd/pdf/journals/ed_lead/el_199205_meyer.pdf

Moon, T. R., Brighton, C. M., Callahan, C. M., & Robinson, A. (2005). Development of authentic assessments for the middle school classroom. *The Journal of Secondary Gifted Education, 16*(2/3), 119-133. https://doi.org/10.4219/jsge-2005-477

National Council for the Social Studies. (1994). *Expectations of excellence: Curriculum standards for social studies.* NCSS.

Newmann, F. M., Secada, W. G., & Whelage, G. G. (1995). *A guide to authentic instruction and assessment: Vision, standards, and scoring.* Wisconsin Center for Educational Research.

Newmann, F. M., Brandt, R., & Wiggins, G. (1998). An exchange of views on "semantics, psychometrics, and assessment reform": A close look at 'authentic' assessments. *Educational Researcher, 27*(6). 19-22. https://doi.org/10.2307/1176091

O'Brien, J. (1997). Statewide social studies performance assessments: Threat or treat? *Social Studies, 88*(2), 53-59. https://doi.org/10.1080/00377999709603747

Obama, B. (2009). *Obama speaks to the U. S. Hispanic chamber of commerce*. http://www.washingtonpost.com/wp-srv/politics/documents/Obama_Hispanic_Chamber_Commerce.html

Parke, C. S., & Lane, S. (2007). Students' perceptions of a Maryland state performance assessment. *The Elementary School Journal, 107*(3), 305-324. https://doi.org/10.1086/511709

Parke, C. S., & Lane, S. (2008). Examining alignment between state performance assessments and mathematics classroom activities. *Journal of Educational Research, 101*(3), 132-146. https://doi.org/10.3200/JOER.101.3.132-147

Parke, C. S., Lane, S., & Stone, C.A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation 12*(3), 239-269. https://doi.org/10.1080/13803610600696957

Pecheone, R., & Kahl, S. (2014). Where are we now? In Darling-Hammond, L., & Adamson, F. (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning.* Jossey-Bass.

Pfeifer, G. R. (2002). *The influence of authentic assessment tasks and authentic instruction on lutheran elementary school fifth- and sixth-grade students' attitudes toward social studies and authentic projects* (Publication No. 3056348) [Doctoral dissertation, University of Minnesota]. ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/influence-authentic-assessment-tasks-instruction/docview/275677989/se-2

Power, M., Schulkin, J., Loft, J., & Hogan, S. (2009). Referral sampling: Using physicians to recruit patients. *Survey Practice 2*(9). https://doi.org/10.29115/SP-2009-0038

Reed, L. C. (1993). Achieving the aims and purposes of schooling through authentic assessment. *Middle School Journal, 25*(2), 11-13. https://doi.org/10.1080/00940771.1993.11495198

Roschewski, P. (2004). History and background of Nebraska's school-based teacher-led assessment and reporting system (STARS). *Educational Measurement: Issues and Practice, 23*(2), 9-11. https://doi.org/10.1111/j.1745-3992.2004.tb00153.x

Rothbart, G. S., Fine, M., & Sudman, S. (1982). On finding and interviewing the needles in the haystack: The use of multiplicity sampling. *Public Opinion Quarterly, 46*(3), 408-421. https://doi.org/10.1086/268737

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232. http://www.jstor.org/stable/1435044

Shiel, T. K. (2017). *Designing and using performance tasks: Enhancing student learning and assessment.* Corwin.

Sivalingam-Nethi, V. (1997). *Examining claims made for performance assessments using a high school science context.* (Publication No. 9714915) [Doctoral Dissertation, Cornell University] ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/examining-claims-made-performance-assessments/docview/304344664/se-2

Spillane, J. P., & Zeuli, J.S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation, 21,* 1-27.

Stanford Center for Assessment, Learning and Equity. (2014). *Performance assessment quality rubric*. SCALE: Stanford University.

Stecher, B. (2010). *Performance assessment in an era of standards-based educational accountability*. Stanford University, Stanford Center for Opportunity Policy in Education.

Stecher, B. M., Barron, S. L., Chun, T. & Ross, K. (2000). *The effects of the Washington state education reform on schools and classrooms*. Center for the Study of Evaluation.

Stecher, B. M. & Klein, S. P. (1997). The cost of science performance assessments in large scale testing programs. *Educational Evaluation and Policy Analysis, 19*(1), 1-14. https://doi.org/10.3102/01623737019001001

Stecher, B. M., & Mitchell, K. J. (1995). *Portfolio driven reform: Vermont teachers' understanding of mathematical problem solving*. National Center for Research on Evaluation, Standards, and Student Testing. https://cresst.org/wp-content/uploads/TECH400.pdf

Stone, C. A. & Lane, S. (2003). Consequences of a state accountability program: Examining relationships between school performance gains and teacher, student, and school variables. *Applied Measurement in Education, 16*(1), 1-26. https://doi.org/10.1207/S15324818AME1601_1

Stosich, E. L., Snyder, J., & Wilczak, K. (2018). How do states integrate performance assessment in their systems of assessment? *Education Policy Analysis Archives, 26*(13) 1-26. https://doi.org/10.14507/epaa.26.2906

Stotsky, S. (2016). Testing limits. *Academic Questions, 29,* 285-298. https://doi.org/10.1007/s12129-016-9578-4

Strong, S., & Sexton, L.C. (2000). A validity study of Kentucky's performance based assessment system with national merit scholars and national merit commended. *Journal of Instructional Psychology, 27*(3), 202-206.

https://link.gale.com/apps/doc/A66355142/HRCA?u=anon~588bf5c8&sid=googleScholar&xid=21d22df3

Suh, V., & Grant, L. W. (2014). Assessing ways of seeing the past: Analysis of the use of historical images and student performance in the NAEP U.S. history assessment. *The History Teacher, 48*(1), 71-90. http://www.jstor.org/stable/43264380

Taylor, C. S. (1998). An investigation of scoring methods for mathematics performance-based assessments. *Educational Assessment, 5*(3), 195-224. https://doi.org/10.1207/s15326977ea0503_3

University of California Regents. (2018). What is smarter balanced? http://www.smarterbalanced.org/about/

U.S. Department of Education. (2020). College and career readiness standards. https://www.ed.gov/k-12reforms/standards

Van Duinen, D.V. (2006). Authentic assessment: Praxis with power. *International Journal of Learning, 12*(6), 141-148. https://doi.org/10.18848/1447-9494/CGP/v12i06/45150

VanHover, S., Hicks, D., Stoddard, J., & Lisanti, M. (2010). From a roar to a murmur: Virginia's history and social science standards 1995-2009. *Theory and Research in Social Education, 38*(1), 80-113. https://doi.org/10.1080/00933104.2010.10473417

Van Hover, S., Hicks, D., & Washington, E. (2011). Multiple paths to testable content? Differentiation in a high-stakes testing context. *Social Studies Research and Practice, 6*(3), 34-51. https://doi.org/10.1108/ssrp-03-2011-b0003

Virginia Board of Education. (2021). *Virginia board of education July 2021 meeting update (Item I).* Virginia School Boards Association. https://www.vsba.org/wp-content/uploads/2021/07/July-2021-Meeting-Report.pdf

Virginia Department of Education. (2014). *Guidelines for local alternative assessments for 2014-2015 Developed in Response to 2014 Acts of Assembly.* (Superintendents Memo 292-14 Attachment B). Virginia Department of Education.

Virginia Department of Education. (2015). *History and social science standards of learning for Virginia public schools: United States history to 1865.* Virginia Department of Education. https://www.doe.virginia.gov/teaching-learning-assessment/k-12-standards-instruction/history-and-social-science/standards-of-learning

Virginia Department of Education. (2017). *Update on the implementation of local alternative assessments.* (Superintendent's Memo 135-17). Virginia Department of Education.

Virginia Department of Education. (2019a). *Assessment Literacy Glossary.* Virginia Department of Education.

Virginia Department of Education. (2019b). *Balanced assessment plans for 2019-2020.* (Superintendent's Memo 181-19). Virginia Department of Education.

Virginia Department of Education. (2019c). *Guidelines for local alternative assessments for 2018-2019 through 2019-2020.* (Superintendent's Memo 025-19 Attachment A). Virginia Department of Education.

Virginia Department of Education. (2019d). *Virginia quality criteria tool for performance assessments.* Virginia Department of Education.

Virginia Department of Education. (2020). *History and social science common rubric middle school.* Virginia Department of Education. https://www.doe.virginia.gov/teaching-learning-assessment/k-12-standards-instruction/history-and-social-science/assessment-resources

Virginia Department of Education. (2021a). *Guidelines for local alternative assessments: 2021-2022 and beyond.* (Superintendent's Memo 214-21 Attachment A). Virginia Department of Education.

Virginia Department of Education. (2021b). *Implementation support for balanced assessment plans.* (Superintendent's Memo 214-21 Attachment B). Virginia Department of Education.

Virginia Department of Education. (2022a). *Profile of a Virginia graduate.* Virginia Department of Education. https://www.doe.virginia.gov/parents-students/for-students/graduation/policy-initiatives/profile-of-a-virginia-graduate

Virginia Department of Education. (2022b). *Virginia SOL assessment program.* Virginia Department of Education. https://www.doe.virginia.gov/teaching-learning-assessment/student-assessment/virginia-sol-assessment-program

Virginia Department of Education. (2023a). *Standards of learning introduction to TestNav 8: Multiple-choice and technology-enhanced item tests.* Virginia Department of Education.

Virginia Department of Education. (2023b). *Virginia School Quality Profiles.* Virginia Department of Education. https://schoolquality.virginia.gov/

Visintainer, C. (2002). *The relationship between two state -mandated, standardized tests using the norm -referenced TerraNova and the criteria -referenced, performance assessment developed for the Maryland school performance assessment program (MSPAP).* (Publication No. 3046012) [Doctoral Dissertation, Wilmington College]. ProQuest Digital Dissertations & Theses global. https://www.proquest.com/dissertations-theses/relationship-between-two-state-mandated/docview/305475370/se-2

Webb, N. M., Schlackman, J., & Sugrue, B. (2000). The dependability and interchangeability of assessment methods in science. *Applied Measurement in Education, 13*(3), 277-301. https://doi.org/10.1207/S15324818AME1303_4

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrator Science Quarterly, 2*(1), 36-49. https://doi.org/10.2307/2391875

Weick, K. E. (1982). Administering education in loosely coupled systems. *Phi Delta Kappan, 63*(10), 673-676. https://www.jstor.org/stable/20386508

Welch, S. (1975). Sampling by referral in a disbursed population. *Public Opinion Quarterly, 39*(2), 237-245. https://www.jstor.org/stable/2748151

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *The Phi Delta Kappan, 70*(9), 703-713. https://grantwiggins.files.wordpress.com/2014/01/wiggins-atruetest-kappan89.pdf

Wiggins, G. (1993). Assessment to improve performance, not just monitor it: Assessment reform in the social studies. *Social Science Record, 30*(2), 5-12.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* Jossey-Bass.

Wiggins, G., & McTighe, J. (2004). *Understanding by design professional development workbook.* ASCD.

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). Pearson.

Wren, D. G., & Gareis, C.R. (2019). *Assessing deeper learning.* Rowman & Littlefield.

APPENDIX A

Virginia Quality Criteria Review Tool for Performance Assessments

Revised: June 2019

The rubric for each quality rating is as follows:

0-No Evidence; 1-Limited Evidence; 2-Partial Evidence; 3-Full Evidence.

Criterion 1: Standards/Intended Learning Outcomes

#	Description	Quality Rating	Evidence or Rationale
1A	Virginia Standards of Learning selected for the performance assessment are clearly listed in a task template, developmentally appropriate for target students, and aligned to the grade-level scope and sequence or grade-level curriculum. Performance assessment components, resources/materials, and student products are aligned to the listed SOLs.		
1B	The performance assessment goes beyond simple recall, elicits evidence of complex student thinking, and requires application of disciplinary or cross-disciplinary concepts, practices, and/or transferable skills, such as application, analysis, evaluation, synthesis, or original creation.		
1C	The performance assessment provides an opportunity for students to develop and demonstrate (even if not explicitly assessed): • Deeper learning competencies, defined as mastering rigorous academic content; learning how to think critically and solve problems; working collaboratively; communicating effectively; directing one's own learning; and developing an academic mindset. The performance assessment may also provide opportunities for students to develop and demonstrate: • Life-Ready competencies defined by the Profile of a Virginia Graduate as content knowledge, career planning, workplace skills, and community and civic responsibility; • Technology-related competencies;		

#	Description	Quality Rating	Evidence or Rationale
	• Integration of intended learning outcomes from two or more subjects.		

Criterion 2: Authenticity

#	Description	Quality Rating	Evidence or Rationale
2	The performance assessment is authentic along the dimensions: • The performance assessment's topic, context (scenario), materials/resources, products, and purpose/audience (i.e., what students are asked to do and for whom) are relevant to the real-world, students' community, students' interests, future careers, or other meaningful context. • The performance assessment asks students to do work authentic to the discipline (i.e., what adult practitioners of the discipline do), such as science inquiry; math problem-solving; analyzing and critiquing a text; analyzing and evaluating historical sources.		

Criterion 3: Language Use for Expressing Reasoning

#	Description	Quality Rating	Evidence or Rationale
3A	The performance assessment supports language use and development by providing multiple means of accessing and using developmentally appropriate academic and disciplinary language for the students to express their reasoning.		
3B	The performance assessment should require students to use one or more forms of language to communicate their reasoning. The performance assessment may provide access to functional, academic, and disciplinary language in various forms of language media (text, video, audio, oral) OR provide opportunity to practice the use of language through multiple means of expression and language production (text, language media production, oral language, or conversation with peers).		

Criterion 4: Success Criteria for Students

The Virginia Department of Education's Common Rubrics, when available, should be used to evaluate and score student work.

#	Description	Quality Rating	Evidence or Rationale
4A	The performance assessment includes a rubric or other appropriate scoring tools (e.g., checklist, analytic rubric) with scoring dimensions that are tightly aligned to performance expectations of the intended learning outcomes targeted within the performance assessment. Criteria should include language objectives, if applicable.		
4B	The scoring tool is written clearly and concisely, with audience-friendly language, as appropriate. Language of the scoring tool should describe how a response demonstrates performance expectations so that the tool may be used to provide feedback to students about their work and how it can be improved.		
4C	The scoring tool or feedback methodology should be used across performance assessments within the course so that results on the performance assessment can be used to communicate a consistent set of expectations to students, monitor students' academic growth over time, inform instructional decisions, and communicate student proficiency to others (e.g., parents/guardians).		

Criterion 5: Student Directions, Prompt, and Resources/Materials

#	Description	Quality Rating	Evidence or Rationale
5A	The student-facing task prompt, directions, and resources/materials are aligned to the intended learning outcomes, task purpose, and the performance expectations being assessed (i.e., the student product will provide evidence of the performance expectations).		
5B	The student-facing task prompt, directions, and resources/materials are clear, complete, written in accessible language appropriate to the grade level, and organized for students in an accessible format.		

#	Description	Quality Rating	Evidence or Rationale
5C	The task prompt/directions, topic, context (scenario), and materials/resources are sensitive to the community and free of bias.		

Criterion 6: Accessibility

#	Description	Quality Rating	Evidence or Rationale
6A	The performance assessment is designed to accommodate the participation of all students. Directions for teachers for the performance assessment identify appropriate supports or alternatives to facilitate accessibility while maintaining the validity and reliability of the assessment.		
6B	The performance assessment is accessible and allows for differentiating the ways that students demonstrate their knowledge such as through the application of principles of Universal Design for Learning (UDL). Refer to the National Center on UDL at the Center for Applied Special Technology (CAST).		

Criterion 7: Feasibility

#	Description	Quality Rating	Evidence or Rationale
7A	Student-facing prompts, directions, resources/materials, and scoring tools are included. Resources and materials required by the performance assessment are realistic and easily accessible to teachers.		
7B	Duration of implementation of the performance assessment is indicated and is realistic for the complexity of the assessment and the scope of performance expectations being assessed.		
7C	If the performance assessment is implemented over multiple lessons, a schedule indicating how the performance assessment is implemented across the lessons is included. Information about students' prior learning and how the performance assessment fits within a learning sequence is included.		

(VDOE, 2019d)

APPENDIX B

Interview Protocol

1. With the removal of the SOL tests in US I and US II, school divisions were required to develop a local alternative assessment plan (or balanced assessment plan) which identified the set of assessments the division would use to measure student learning in place of the removed SOL tests. Please describe your division's local alternative assessment plan for US I and US II.

Possible follow-up questions:

a) Describe the level of uniformity or variation of your local alternative assessment plan and assessments across the division (are assessments common across the division, common within a school or unique to each teacher, is there a set or pool of assessments from which teachers choose from or do teachers develop their own? Is every school expected to have the same number and formats of assessments?).

b) Describe the number and format of assessments in your local alternative assessment plan for a given course (may need to specify for a particular school if each school has its own plan, depending on the response to Part a). (e.g., multiple-choice, short answer, performance assessments, namely constructed-response, stand-alone, unit-embedded, and/or project-based, and how many of each?)

c) Describe the types of local alternative assessments (e.g., formative, diagnostic, summative) and how they will be used to measure students' mastery of specific content and skills for each course.

For the following questions, please think about ONE performance assessment for EACH US I and US II that your division developed and is implementing as part of the local alternative assessment plan. Do you have one performance assessment for each course in mind?

2. Describe the process the developer/developers engaged in to develop the assessment. (unpacking of the curriculum, internal audit of existing performance assessments, review of best practices/external resources)

   Possible follow up questions:

   a) How did the division decide who to involve in the development of the assessment? (teacher/group of teachers, instructional coordinators/curriculum coordinators/curriculum specialists/testing specialists, combination of teachers and instructional coordinators, consortium of school divisions working together, consultant outside of the division, how many people were involved, their roles or nature of their involvement: developer, reviewer)
   b) Describe any models, templates or other guidelines used to help construct the assessment.
   c) Describe the role of the VDOE Quality Criteria Tool in the process of development. (used as a guideline in the process or after completion as a tool to evaluate the assessment)
   d) Describe any piloting or reviews of the assessment after completion but before implementation. (Piloted, student responses reviewed to look for bias, problems with the wording/students not understand what is being asked of them, external expert review)
   e) Describe any review or revising of the assessment that has occurred since the initial development and implementation of the assessment.

3. Describe or identify any training on performance assessments and/or the VDOE policy on local alternative assessments (LAAs) in which the individuals who developed the performance assessment were able to participate. (workshops on LAAs provided by the VDOE, VASSL, SURN, or another Virginia professional educational organization, presenters or consultants led division-level training, division conduct internally led professional development, duration, follow up sessions, recorded so participants could revisit the training)
    a) Describe how the information gained at these trainings were disseminated to other teachers implementing the LAAs.
    b) Describe the duration of the training. (Was the training a one-time event, or a series of workshops? Did the same people participate throughout the sequence of trainings? Was additional time provided for participants to collaborate outside of this 'formal' training?)
    c) Describe the availability of opportunities for training for division personnel and/or in your division.
4. Describe the process for scoring student products?

    Possible follow up questions:

    a) Could you elaborate on who is involved in scoring student products from this performance assessment? (each teacher scores their own students, building-level or division-level teams score student work and compare scores)
    b) Describe any training or protocols used by the division to establish inter-rater reliability between teachers scoring the performance assessments.
    c) Describe any opportunities for cross-scoring student responses amongst teachers in a school or across the division.

APPENDIX C

Request for Recommendations

Hello. My name is Molly Sandling and I am conducting research for my dissertation for the School of Education at The College of William and Mary. For my study I am focusing on the performance-assessments that were created by local school divisions as a replacement for the SOL tests in US History I and US History II. With divisions working autonomously to develop their own performance assessments, my study is exploring how school divisions responded to the mandate in terms of what types of assessments have been developed and how well do those assessments meet the VDOE definition of quality. I am calling you as an educational professional who has worked with divisions across Virginia in this process, for recommendations of school divisions for my study. Would you be willing to give me the names of school divisions who you feel have taken a conscientious approach to developing the performance assessments for the LAAs or divisions that have had some success in developing strong performance assessments for the LAAS, namely in US History I and US History II? I understand that you may not have reviewed nor seen the assessments, that your judgement may be based more on what you know of the division or have heard, and you do not have to share the basis for your judgement. I will be reaching out to other groups and individuals for recommendations of divisions and I will not inform the divisions I contact who recommended them, thus no one will know who you recommended or that you recommended them.

APPENDIX D

Initial Contact

Hello. My name is Molly Sandling and I am conducting research for my dissertation for the School of Education at The College of William and Mary. For my study I am focusing on the performance-assessments that were created by local school divisions as a replacement for the SOL tests in US History I and US History II. With divisions working autonomously to develop their own assessments, my study explores how school divisions responded to the mandate in terms of what types of assessments have been developed and how well do those assessments meet the VDOE definition of quality. Your division was recommended to me by educational professionals in Virginia who felt your division has had success in developing strong performance assessments. Since the work in your division is so highly regarded, would you please be willing to share with me the individual in your division who was or is responsible for developing, implementing, and/or supervising the implementation of your performance assessments for the division LAAs in US History I and US History II? If you are the person responsible can we arrange a time to briefly discuss your LAAs? I would also like to send you or the person involved with supervising the implementation of LAAs a Letter of Informed Consent to participate in the study, is there an email to which I can send a copy of that letter?

APPENDIX E

Request and Interview

Thank you for being willing to talk to me about the performance assessments that your division has developed as one of your LAAs that replace the SOLs in US History I and US History II. Also thank you for returning the Letter of Informed Consent OR Have you had a chance to read the Letter of Informed Consent and could you please return that to me signed? For my study I am exploring the development process, types of assessments and quality of assessments being created by divisions across the commonwealth. Would you be willing to share with me two of the performance-based assessments that your divisions has developed as LAAs, one from US History to 1865 and one from US History 1865 to the present? If so, would you be able to give me a copy of the assessment and any/all student-facing materials such as prompts, resources or rubrics as well as any/all teacher-facing materials such as instructions, guidelines, and rubrics? I can send you a self-addressed stamped envelope if you have hard copies or you can email them to me. I also have a set of interview questions concerning how your division approached the process of developing these assessments. I will be recording your responses to ensure that I have an accurate account of what your words. All recordings, transcripts, and assessments will be stored securely and no there will not be any identifying information stored with them.

As you respond to the questions please choose ONE performance-based Local Alternative Assessment (LAA) that your division has developed to meet the VDOE requirements in "US History to 1865" and ONE performance-based LAA for US History 1865 to the Present. Please respond to the following questions in regard to the TWO assessments you have selected. This will be followed by the questions in Appendix B.

APPENDIX F

Follow Up Email

Thank you for your time and willingness to participate in my study. I appreciate the time that you spent with me during the interview and for sharing with me the experience of your division.

EITHER: If you could please respond to this email with copies of your performance LAAs, one from US History I and one from US History II, as well as any/all student-facing and teacher-facing materials.

OR: Thank you for sharing with me two of your performance LAAs.

Please feel free to contact me with any questions you may have.

Thank you again for your time and assistance,

APPENDIX G

Letter of Informed Consent

Research Participation Informed Consent Form
Education Department
The College of William and Mary

Protocol
Title: An Investigation of the Quality of Performance Assessments and Implications of a Grassroots Approach to Accountability Reform

Principal Investigators: Molly Sandling
This is to certify that I, _____, have been given the following information with respect to my participation in this study:

Purpose of the study: This study is designed to explore the different ways that school divisions have chosen to meet the requirement of replacing state-mandated, state-wide, multiple-choice tests with their own alternative performance assessments and evaluate the quality of those assessments using the Virginia Quality Criteria Tool for Performance Assessments.

What you will be asked to do: As a participant in this study, you will be asked to participate in an interview with the researcher about your division's processes for developing the local alternative performance assessments and the division's Balanced Assessment Plan in US I and US II. You will also be asked to share a copy of one performance assessment developed by your division for the local alternative assessment for each course, US I and US II. You may be asked to participate in a brief follow up interview with the researcher. Participation in the different steps in the data collection is voluntary. Please sign indicating that you are willing to participate:

I agree to a semi-structured interview and follow up interviews with the researcher

_____ signature _____ date

I agree to share one performance assessment developed as a local alternative assessment for US I and one performance assessment developed as a local alternative assessment for US II

_____signature _____date

Discomforts and risks: There are no known risks associated with participating in the study and interview.

Duration of participation: Participation in this study will take approximately 1-1.5 hours.

Statement of confidentiality: Your participation is confidential. The data you contribute to this research will be identifiable only by a number assigned by the researcher. The key of numbers

assigned and the names of the divisions will be kept in a separate location from all research materials. All data and records will be stored on password-protected computers and in a locked cabinet.

Voluntary participation: Participation is voluntary. You are free to withdraw at any time. You may choose to skip any question.

Potential benefits: There are no known benefits of participating in the study. However, your participation in this research will contribute to the development of our understanding performance assessments being developed by different divisions and the implications of the VDOE granting greater autonomy to divisions.

Termination of participation: Participation may be terminated by the researcher if it is deemed that the participant is unable to perform the tasks presented.
Questions or concerns regarding participation in this research should be directed to: Molly Sandling 757-561-3313.

I am aware that I must be at least 18 years of age to participate in this project.
I am aware that I may report dissatisfactions with any aspect of this study to Dr. Jennifer Stevens, Ph.D., the Chair of the Protection of Human Subjects Committee by telephone (757-221-3862) or email (jastev@wm.edu).

I agree to participate in this study and have read all the information provided on this form. My signature below confirms that my participation in this project is voluntary, and that I have received a copy of this consent form.

_____ Signature _____date

_____ Witness _____date

THIS PROJECT WAS APPROVED BY THE COLLEGE OF WILLIAM AND MARY PROTECTION OF HUMAN SUBJECTS COMMITTEE (Phone: 757-221-3966) ON [AND EXPIRES ON []

Preferred method and phone number/email for the researcher to contact me to arrange an interview:
_____

APPENDIX H

Common Rubric for History and Social Science Performance Assessments/Tasks Middle School

	4	3	2	1	Not Observed
Core Expectations (.1a and .1c)					
Accuracy of Content Synthesizing information sources Explaining Evidence	• Identified, analyzed and interpreted information sources to demonstrate an in-depth understanding of content • Integrated evidence from multiple information sources to determine characteristics of people, places, events or concepts • Used information to consistently develop, support, or refine the explanation or statement	• Analyzed and interpreted information sources to understand specific content • Gathered and classified information to sequence events and separate fact from opinion • Used information to develop and support an explanation or statement	• Used information sources to understand of concepts, people, places, or events • Classified information, sequenced events, and separated fact from opinion • Used information to support an explanation	• Used information sources to understand content • Separated fact from opinion • Identified information to support an explanation	
Task Specific Concepts and Skills					
Geographic Patterns and Trends (.1b)	Used geographic information to analyze the impact of geographic features on a pattern or trend.	Used basic map skills and geographic information to identify a pattern or trend in data	Used basic map skills to identify data	Used basic map skills	
Evaluating Sources (.1d)	Used evidence to draw conclusions and make generalizations about points of view and historical perspective	Used evidence to summarize points of view or historical perspective	Used evidence identify points of view or historical perspective	Answered questions about points of view or historical perspective	
Explanation or Statement (.1d)	Responded to the task with a decisive explanation or statement beyond conventional conclusions	Responded to the task with a reasonable explanation or statement	Responded to the task with a partially developed explanation or statement	Attempted to present a central explanation or statement	

	4	3	2	1	Not Observed
Differing Perspectives (.1e)	Compared and contrasted ideas about historical, cultural and political perspectives in history	Compared and contrasted concepts, people, places, or events	Explained concepts, people, places, or events	Identified concepts, people, places, or events	
Determine causes or effects (.1f)	Determined and explained relationships with many causes or effects	Explained direct cause-and-effect relationships	Identified direct cause-and-effect relationships	Identified a cause-and-effect relationship	
Connections across time (.1g)	Explained connections across time and place	Made connections between past and present events	Made connections between past events	Identified past and present events	
Making decisions (.1h)	Used a decision-making model identify the costs and benefits of a specific choice made	Identified the costs and benefits of a specific choice	Identified the costs or benefits of a specific choice made	Identified that a specific choice was made	
Citizenship (.1i)	Used authentic, valid sources and gave credit when using outside ideas, opinions, or theories	Used sources and gave credit when using outside ideas, opinions, or theories.	Used sources and gave credit incorrectly when using another person's ideas, opinions, or theories	Used sources	
Developing Research Questions (.1j)	Identified a question and made a connection between the question and existing information or ideas about a topic	Identified a question and stated existing ideas or information about a topic	Restated existing ideas or information about a topic	Made up ideas or information about a topic	
Selecting Sources (.1j)	Selected relevant sources by accessing a variety of media, including online resources	Selected sources from a variety of media	Selected sources that represent two different types of media	Selected sources	

(VDOE, 2020)

Vita

Molly M. Sandling

Education/Professional Development

Doctor of Philosophy in K-12 Administration, 2023
College of William and Mary
Williamsburg, VA

Master of Arts in Education, 2000
College of William and Mary
Williamsburg, VA.

Master of Arts, Department of History, 1996
Yale University
New Haven, CT.

Bachelor of Arts, Department of History, 1995
College of William and Mary,
Williamsburg, VA.
Magna cum Laude, High Honors in History, Phi Beta Kappa

Professional Experience

August 2000-present	**Social Studies Teacher** Jamestown High School Williamsburg, VA
June 2016-present	**Test item developer, AP Human Geography Exam** College Board Princeton, NJ
August 2000-2015	**Consultant, The Center for Gifted Education** College of William and Mary. Williamsburg, VA

Publications

Sandling, M. (2022). Adapting social studies curricula for high-ability learners. In J. VanTassel-Baska & C. Little (Eds.), *Content-based curriculum for gifted learners* (4th ed.). Prufrock.

Sandling, M., & Chandler, K. L. (2014). *Exploring America in the 1950s: Beneath the Formica.* Prufrock Press. Winner 2014 Legacy Book Award, Texas Association of the Gifted and Talented.