Toward a Shared Praxis of Assessment: Performance Assessment, Alignment, and Student Learning in the Secondary Social Studies Classroom

---

A Dissertation

Presented to

The Faculty of the School of Education and Human Development

University of Virginia

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

---

by

Michael Gurlea

March 2022

Abstract

This dissertation consists of three manuscripts that explore student learning, assessment practice, and assessment policy in secondary social studies classrooms using qualitative research design. The first manuscript presents a case study of a unit on market structures in an Advanced Placement (AP) Microeconomics classroom. This manuscript investigates student thinking about their own learning on a teacher-constructed end-of-unit test (intended to mirror the AP test) using an analysis of classroom observations and interviews with the teacher and four focal students. The second manuscript focuses on a state-level policy in Virginia which sought to reduce the number of traditional multiple-choice tests students were required to take in middle school social studies and replace them with "local alternative assessments" designed at the division-level. Using document analysis and survey and interview data from division-level coordinators this manuscript explores the macro-level implementation of the policy. Finally, the third manuscript presents a systematic review of the recent empirical literature on large-scale and classroom-based assessment in secondary social studies. The three manuscripts are united by a focus on assessment's role in the social studies classroom and the ways that teachers, teacher leaders, and students experience large-scale and classroom-based assessment.

Curriculum, Instruction, and Special Education

School of Education and Human Development

University of Virginia

Charlottesville, Virginia

APPROVAL OF THE DISSERTATION

This dissertation, "Toward a Shared Praxis of Assessment: Performance Assessment, Alignment, and Student Learning in the Secondary Social Studies Classroom," has been approved by the Graduate Faculty of the School of Education and Human Development in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Stephanie van Hover, Advisor

_____
Peter Youngs

_____
Jennifer Chiu

_____
Jennifer Pease

March 31, 2022

Dedication

To the students I have had the privilege of teaching.

To the athletes I have had the privilege of coaching.

To Paul and Gina Gurlea—This was only possible with your love.

Acknowledgements

I would first like to acknowledge the extreme circumstances that my peers and I lived through during the pursuit of our PhDs and the completion of our dissertations. This task should not have been possible considering the state of the world, but it was made possible through the unprecedented aid and care of the community that I would like to thank below.

I would like to acknowledge and express gratitude to my dissertation committee that has guided me through this process with wisdom, kindness, and understanding. Thank you to Dr. Jen Pease for your invaluable guidance and example both in research and in teacher education. Thank you to Dr. Jennifer Chiu for your clear and kind support, all in the service of pushing my work to a higher standard. Thank you to Dr. Peter Youngs for being a steadfast supporter of me and all PhD students across the School of Education and Human Development. Finally, thank you to my advisor, Dr. Stephanie van Hover, for your mentorship through both the most challenging periods in my life and, maybe, the most challenging periods of modern human history. Your respect for me and belief in me when I had trouble finding those things myself helped me grow into a better scholar and educator during this program.

I also want to thank the community of colleagues that have remained supportive of me since my time as a public school teacher. Thank you to my mentor from my first-year teaching and friend, Kim Steiner, for your endless support. Thank you to Ryan Girard for your inspiration as an excellent teacher of social studies and for your kindness. Thank you to Kaitlin and Mark Murray for helping me learn how important schools are as communities and for your friendship through all of the highs and lows of the last few years. Thank you to Nick and Katie DelDotto for your unwavering belief in me, your friendship, and your guidance in my decision to come to the

University of Virginia. You all are the finest educators I know, and it is because of all of you that I think public schools are good and can be better

I would also like to acknowledge several peers from the School of Education and Human Development, who have turned from classmates and colleagues to friends. Thank you to Dr. Ariel Cornett for making a concerted effort to be a mentor and friend to me when I first arrived in Charlottesville and every day after. Thank you to Dr. Alicen Brown for your support, camaraderie, and kindness. Thank you to Dr. Tyler Woodward, my closest collaborator in my time here- all of those hours we spent working together were in service of the research, but also in service of forging a lifelong friendship. Finally, it is with massive gratitude that I acknowledge the people who became my "bubble" and safe space during the global pandemic. Thank you to Hunter Holt for being a steadfast and true friend no matter the circumstances. Thank you to Amy Laboe for both literally and metaphorically giving me a place to 'be' when I did not have anywhere else. Thank you to Kristan McCullum for your shining example and for innumerable conversations filled with love and support – I simply could not have done this without you. Without all of your love the experience of a PhD program during a global pandemic would have been nigh impossible.

Lastly, I would like to thank my family. To Andrew and Tristyn Holtzhauer, thank you for being my family and home in all but name. To my father, Paul Gurlea, there has never been a moment where I felt like you doubted that I could do anything I put my mind to, thank you for tireless confidence. To my mother, Gina Gurlea, thank you for everything: I am who I am because of you. I love you all.

Table of Contents

Introduction and Overview of Manuscripts

Over the course of my time at the University of Virginia I became interested in assessment in the secondary social studies classroom. Assessment, both classroom-based and large-scale, is often ignored in the literature on teaching and learning social studies. An empirical understanding of assessment, however, represents an empirical understanding of what should be at the core of every teacher's practice: student learning. This dissertation proposal consists of three manuscripts that trace my research trajectory from an interest in student learning in context toward assessment policy and practice in the secondary social studies classroom. Taken as a whole, this dissertation presents a broader argument that the field of secondary social studies should focus deeply on assessment, both in how teachers use assessment to help students reach ambitious disciplinary goals and in how assessment policy creates contexts in which teachers and students must make decisions about what is important to learn.

Specifically, I argue that the three papers of this dissertation and the current research base in assessment in secondary social studies suggest an expansion of Stephen Thorton's seminal framework of teacher as curricular-instructional gatekeeper (1989, 2006). Thornton's framework is effective in helping us understand the classroom teacher in context, as they make sense of instructional approaches, curriculum, and their own philosophies and goals and filter these in service of a particular classroom experience for their students. This framework is highly interactive and recognizes the complexity of a teacher's decision-making process (See Figure 1). The work I have done calls for an expansion of Thornton's framework to include assessment, both at the macro level (i.e., the assessment policy context that a teacher inhabits) and at the micro level (i.e., the assessment tools a teacher can access given their classroom, school, and division context). Assessment policy contexts and assessment tools matter a great deal to both

teacher education and research on education as they influence how we think about, not just student classroom experiences, but about student learning (See Figure 2).

Figure 1

Teacher as curricular-instructional gatekeeper

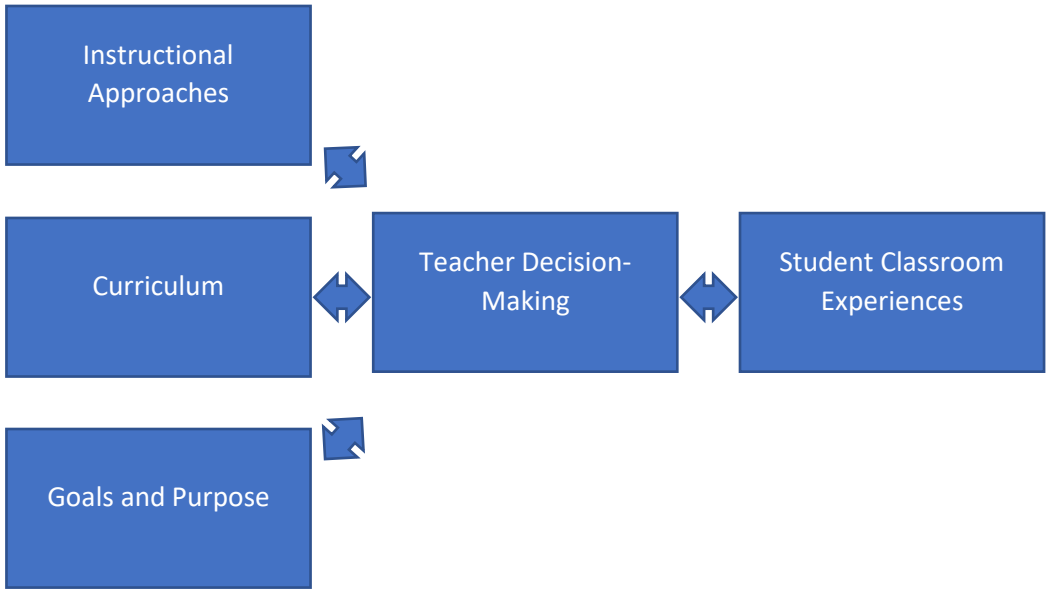
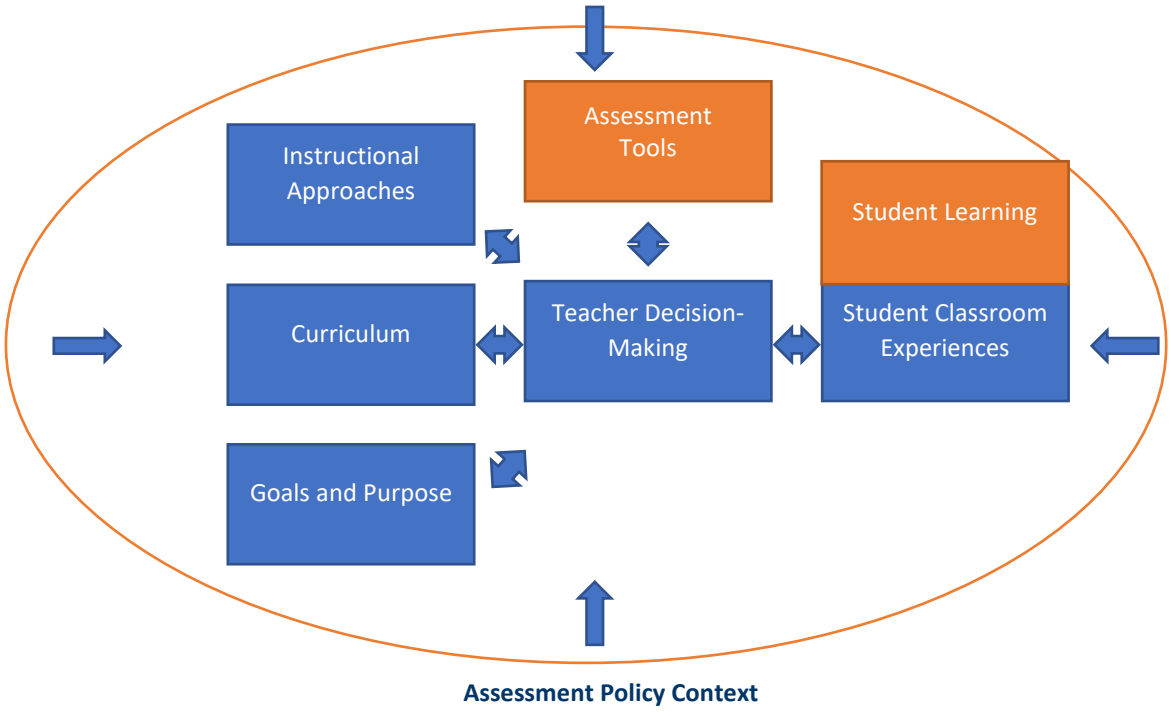


Figure 2

Teacher as curricular-instructional-assessment gatekeeper



Taken together the manuscripts of this dissertation also present an argument that the policies of the last few decades have influenced teachers, researchers, and teacher educators to think less about assessment as a measure of student learning and more as an external accountability tool. Therefore, in the field of social studies education a refocus on classroom-based assessment as a formative tool to understand student learning is warranted. Further, it is important to continue to examine the literature surrounding our conception of best practice classroom-based assessment and how that influences our shared language and praxis.

The first manuscript is a single-teacher, single-unit case study that took place in an Advanced Placement (AP) Microeconomics classroom. This case study highlighted a key tension between the instructional decision-making of the teacher, the experiences of students during the unit of study, and students' poor performance on their end-of-unit assessment constructed of released AP test items. The teacher felt pressure to include the AP-style end-of-unit assessment, but her instruction during the unit did not align with the skills necessary for students to perform well on this test.

The tension between assessment and instruction in this case study teacher's classroom led me to a broader interest in the impact of assessment policy on divisions, schools, and teachers. A unique opportunity to explore the relationship between policy and practice presented itself in the state of Virginia as the General Assembly removed two secondary social studies end-of-course standardized tests in 2014. How might teachers and schools respond in a shifting assessment context without the pressure of an externally designed and administered test? The second manuscript traces the implementation of this new assessment policy in the state of Virginia through an analysis of policy documents, surveys, and interviews with division-level social

studies coordinators. This study was impacted by the onset of the COVID-19 pandemic and led to limitations in access to schools and interview subjects.

In working with the assessment literature in both the first and second manuscripts it became clear that student learning and the way that it is measured is often not the focus of empirical research in secondary social studies classrooms. While some scholars have explored assessment more recently in history education (see Ercikan & Seixas, 2015) there has not been a systematic review of the literature on assessment and student learning in secondary social studies since 2008 (Grant & Salinas, 2008). For the third manuscript, I undertake this systematic review of the literature to more deeply understand what form best practice assessment in social studies should take, especially as new policy contexts allow divisions, schools, and teachers more freedom to measure student progress in ways other than an end-of-unit or end-of-course multiple choice tests.

While the research design and focus of each of these manuscripts are different, there is a united narrative centered on how assessment manifests in the secondary social studies classroom and the ways in which policy impacts this manifestation. In this introduction to my dissertation proposal, I will provide an overview of each manuscript and outline a timetable for the completion and publication of each project.

Overview of Manuscripts

Manuscript 1

This manuscript examines how a secondary social studies teacher planned and implemented a unit on market structures in an Advanced Placement (AP) Microeconomics course, how students experienced that instruction, and how students thought about their learning on an end-of-unit assessment constructed by the teacher using AP-released test items. This study

explores the alignment between instructional planning, teaching, student experiences, and student learning answering the call for research that investigates these dynamic instructional connections (Ball & Forzani, 2007).While the field of social studies education has a clear vision of best practice instructional moves based in inquiry, project- and problem-based learning, discussion, and source analysis (Barton & Avery, 2016) this portrait of instruction is often disconnected from student learning and the pressures that teachers feel to prepare students for high-stakes tests (Au 2007; 2009).

This study investigates the dynamic connections between instruction and learning using a qualitative case study design (Yin, 2018) and a research protocol adapted from Nuthall and Alton-Lee (1995). The methodological protocol employed by Nuthall and colleagues involved 1) the development and implementation of a pretest aligned with a teacher's articulated learning outcomes for a unit of study; 2) daily observations of instruction; 3) identification of 4-6 focal students to be continuously observed and recorded; and 4) posttests and interviews of students. We adapted his protocol in order to capture, qualitatively, the interactional complexity of how students describe and experience classroom instruction (Nuthall, 1999), a focus largely missing from research on teaching and learning in social studies classrooms. The study was guided by the following research questions:

1. To what extent were students able to learn the content defined by the teacher's instructional aims and measured by the teacher-constructed assessment?
2. How does the teacher-constructed assessment align with the ways in which students experienced instruction?

It is important to note that I did not engage in data collection for this manuscript. However, I did take the lead role on data interpretation and analysis. Data collection for this

manuscript took place as part of a larger project on student learning in context. During the process of data analysis, I reviewed and coded artifacts of student work, video-recorded classroom observations, and interviews with the instructor and four focal students. This coding took place over multiple, iterative rounds and is described in more depth in the manuscript. These data contributed to the construction of the findings section. I have the role of lead author for this manuscript.

Manuscript 2

The second manuscript explores Virginia as a key case in the implementation of a new state-level assessment policy. In 2014 the Virginia General Assembly moved to replace end-of-course multiple-choice tests in 6th and 7th grade history courses with 'local alternative assessments' that were intended to take the form of performance-based assessments. Virginia's policy, which leaves the design and implementation of these assessments to each division within the state, represents an example of what Linda Darling-Hammond (1994) described as "top-down support for bottom-up reform." While the literature is clear that large-scale assessment policy can act as a lever for instruction (Au 2007; 2009) the trajectory of implementation from the state to the division level has not been explored in secondary social studies. This manuscript was guided by the following research questions:

1. To what extent is the adopted state policy of local alternative assessments being implemented across Virginia school divisions?
2. How do different stakeholders (state-and division-level social studies personnel) make sense of this assessment policy shift?

Data collection included three phases: 1) document analysis to construct a framework for the inception, implementation, and dissemination of the policy; 2) distribution of a survey to all

state-level and division-level social studies administrators and college educators; 3) semi-structured follow-up interviews with select survey respondents. We collected policy documents and news articles to first trace the inception of this shift in assessment policy. Then, we made the decision to create and administer the survey to social studies coordinators, as these were the individuals that were most likely leading the effort to implement the new policies within their school divisions. Lastly, we identified specific coordinators to interview to obtain more in-depth views about what was happening in their respective school divisions and to triangulate our data sources.

My role for this manuscript involved all data collection and analysis. I systematically searched, sorted, and analyzed key documents from news sources and the Virginia Department of Education (VDOE), constructed the survey (with the second author), reached out to division-level social studies coordinators, and conducted follow-up interviews. The second author and I spent significant time in data analysis through multiple rounds of iterative coding cataloguing emergent themes and findings. I have taken on the role of lead author for this manuscript.

Manuscript 3

This manuscript engages in a systematic review of the literature regarding large-scale and classroom-based assessment in secondary social studies. Some scholars, especially in the discipline of history education, have made efforts to present visions of best practice in the realm of assessment (see Ercikan & Seixas, 2015). However, in social studies education more broadly, a review of the empirical evidence regarding assessment has not been conducted since 2008 (Grant & Salinas, 2008). As federal and state policy contexts remain in flux in the transition between *No Child Left Behind* (*NCLB*) and the *Every Student Succeeds Acts* (*ESSA*), there is a renewed necessity for both a collection of empirical evidence regarding best practice in

assessment in social studies education and also an exploration of directions for future research. My role in the composition of this third manuscript is as sole author.

Status of Each Manuscript and Timetable

The first manuscript was submitted to the peer-reviewed journal *Theory and Research in Social Education* (*TRSE*) in December of 2020 and was rejected in early 2021. Using feedback from this committee and from *TRSE* reviewers this manuscript has been edited and submitted to *The Journal of Social Studies Research* (*JSSR*) as of March of this year (See Table 1 for a breakdown of the timeline for each manuscript). The second manuscript has not been submitted to a journal but, rather, will be submitted for inclusion in an edited book about assessment in secondary social studies. The third manuscript is complete and undergoing current edits with feedback from this dissertation committee. I am considering submission to *Review of Educational Research* (*RER*) or *Teachers' College Record* (*TCR*). Submission for manuscript three is planned for the summer of 2022.

Table 1

Manuscript Status and Timetable

Manuscript	Title	Status	Timetable
Manuscript 1	*"I just kind of guessed": Student Constructions of Knowledge in AP Microeconomics*	Under Review	Under Review at the Journal of Social Studies Research (JSSR)
Manuscript 2	*Top Down, Bottom Up…and Then What Happened? Assessment Policy Change in Middle School History Classrooms*	Will submit for inclusion in an edited book in collaboration with Stephanie van Hover and Gabriel Reich	Book currently in proposal phase
Manuscript 3 (Proposal)	*Toward A Shared Praxis: Best Practice Assessment in Secondary Social Studies*	Will submit to Review of Education Research (RER)	Submit to RER summer of 2022

References

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258-267. https://doi.org/10.3102/0013189X07306523

Au, W. (2009). Social studies, social justice: W(h)ither the social studies in high-stakes testing?. *Teacher Education Quarterly*, *36*(1), 43-58. https://www.jstor.org/stable/23479200

Ball, D. L., & Forzani, F. M. (2007). What makes education research" Educational?" *Educational Researcher*, *36*(9), 529.

Barton, K. C., & Avery, P. G. (2016). Research on social studies education: Diverse students, settings, and methods. In D. Gitomer & C. Bell (Eds,), *Handbook of research on teaching* (pp. 985-1038). American Educational Research Association.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, *64*(1), 5-31. https://doi.org/10.17763/haer.64.1.j57n353226536276

Ercikan, K., & Seixas, P. (2015). *New directions in assessing historical thinking*. Routledge.

Grant, S.G., & Salinas, C. (2008). Assessment and accountability in the social studies. In L.S. Levstik & C.A. Tyson (Eds.), *Handbook of research in social studies education* (pp. 219–238). Routledge.

Nuthall, G., & Alton-Lee, A. (1995). Assessing classroom learning: How students use their knowledge and experience to answer classroom achievement test questions in science and social studies. *American Educational Research Journal*, *32*(1), 185-223. https://doi.org/10.3102/00028312032001185

Nuthall, G. (1999). The way students learn: Acquiring knowledge from an integrated science and social studies unit. *The Elementary School Journal, 99*(4), 303-341. https://www.jstor.org/stable/1002174

Thornton, S.J. (1989, March 27-31). *Aspiration and practice. Teacher as curricular-instructional gatekeeper in social studies* [Paper Presentation]. AERA 1989 Conference, San Francisco, CA. https://files.eric.ed.gov/fulltext/ED315347.pdf

Thornton, S.J. (2006). What matters most for gatekeeping? A response to VanSledright. *Theory and Research in Social Education, 34*(4), 416-418. https://doi.org/10.1080/00933104.2006.10473316

Yin, R. K. (2018). *Case study research and applications, 6th Edition.* Thousand Oaks, CA: Sage.

"I just kind of guessed": Student Constructions of Knowledge in AP Microeconomics

Michael Gurlea[1], Colleen Fitzpatrick[2], Stephanie van Hover[1], Ariel Cornett[3]

[1]University of Virginia

[2]University of Toledo

[3]Georgia Southern University

Tests constructed of discrete multiple-choice items have long been the dominant means of summative assessment in secondary social studies courses (Goodlad, 1984; Martin et al., 2011). Their use and impact on instruction and student learning, however, remains contested. While defenders of multiple-choice tests point to their ability to efficiently measure students' learning of clearly defined content and skills standards (Haladyna 1999; 2004), critics highlight that the goals of the multiple-choice test do not align with the goals of the field of social studies education: disciplinary thinking, inquiry, and source analysis (Barton & Avery, 2016; Ercikan & Seixas, 2015). Further, studies on student thinking on multiple choice tests provide evidence that these exams often measure factors irrelevant to content knowledge and disciplinary skills such as literacy and *test-wiseness* (Reich, 2009, 2013; Smith, 2017; Smith et al., 2019). And, in settings with end-of course multiple-choice tests, teachers face a dilemma in the sheer breadth of the content they must cover in a year, often at the expense of depth of knowledge and understanding (Au, 2007, 2009). The research suggests that multiple-choice testing often leads to multiple-choice teaching, with an emphasis on didactic memorization of factual content (Koretz, 2008; Popham, 2003).

Most studies on high-stakes testing contexts focus on instructional decision-making (see Abram et al., 2003; Grant & Salinas, 2008; Pedulla et al., 2003) or student think-alouds (Reich, 2009, 2013; Smith, 2017; Smith et al., 2019). But what happens when a teacher known for student-centered instruction teaches a class with a high-stakes end-of-course Advanced Placement (AP) test? Our study aims to investigate how a teacher, Ms. Walter, planned and implemented a unit on market structures in an AP Economics classroom with the twin goals of both fostering student-centered learning experiences and student success on the end-of-course AP exam. Further, we analyzed what students learned through their classroom experiences as

assessed by an end-of-unit test that was constructed of released items from the AP test. Connecting student learning to instructional decision-making is important in empirical studies because, as Ball and Forzani (2007) suggest, research to understand student learning should push beyond focusing on just one component of the classroom (i.e., just teachers or just students) and try to capture the instructional dynamic that connects them within classroom contexts. As social studies teachers continue to face mixed messages between the pressures of external assessment and best-practice instruction, it is useful and important to investigate how teachers navigate, and how students experience, these contexts.

This study explores the complex connections between instructional practices and learning outcomes in a high school AP Microeconomics classroom. The AP program is a popular option among high school students and passing the end-of-course AP examination represents a means to compete for college admission and earn college credit. Building on the work of Reich (2009, 2013) and Smith and colleagues (2019), this project focuses on students' thinking about their own learning on a teacher-constructed assessment of released AP test items. Through observing classroom instruction, collecting classroom artifacts, and interviewing the teacher and focal students, this study addresses the following research questions:

1. To what extent were students able to learn the content defined by the teacher's instructional aims and measured by the teacher-constructed assessment?
2. How does the teacher-constructed assessment align with the ways in which students experienced instruction?

This manuscript is organized as follows. We begin with a review of the literature situating this study in the broader context of research on the secondary economics classroom and the AP classroom. We then present key empirical evidence from the classroom-based assessment

literature to situate our findings about instructional decision-making and student thinking on an end-of-unit test. Next, we present the methods section which includes our adaption of Nuthall and Alton-Lee's (1995) framework for understanding student learning in context. Methods is followed by the findings section, divided into four parts: 1) An analysis of two test items that students performed well on; 2) An investigation of their classroom experiences surrounding this content; 3) An analysis of two test items that students performed poorly on; and 4) An investigation of their classroom experiences surrounding this content. We conclude with the discussion and implications for practicing teachers, teacher education and future research.

Literature Review

Context: Economics and Advanced Placement

A small body of empirical studies have worked to understand what best practice social studies instruction looks like in the context of an economics classroom (e.g., Ayers, 2018; VanSickle, 1992; Wentworth, 1987). Standards documents and practitioner articles provide insight into visions of what best practice in economics should look like and are congruent with notions of best practice outlined in the larger field of social studies education. The *C3 Framework* (NCSS, 2013), for example, calls for social studies education to teach using the inquiry arc, focusing on students "developing questions and planning inquiries," "applying disciplinary tools and concepts," "evaluating sources and using evidence," and "communicating conclusions and taking informed action" (p. 12). In economics, the inquiry arc promotes student understanding of big ideas in economic education such as "economic decision making, exchange and markets, the national economy, and global economy" (NCSS, 2013, p. 36-39). Despite the aforementioned vision of best practice in economics, the research base is thin and focuses more on specific teaching moves (Maxwell et al., 2005; Mergendoller et al., 2001; 2006), teacher

knowledge and instructional decision-making (Ayers, 2018), and quantitative measures of student knowledge (Butters & Asarta, 2011; Miller & VanFossen, 2008; Soper & Walstad, 1988; Walstad & Rebeck, 2001; Walstad et al., 2013). Few studies explore how students experience economics instruction.

Very little research has explored how teachers teach disciplinary economic reasoning and how students learn from this approach. Ayers's (2018) qualitative case study is an exception and explored what pedagogical content knowledge (PCK) looked like with three award-winning secondary economics teachers. She found that the teachers "used economics reasoning in ways geared toward helping students unpack often difficult economic content and apply economic reasoning tools to their personal lives" (Ayers, 2018, p. 69). These teachers engaged in multidisciplinary instruction, focused on real-world application of economic concepts with an emphasis on economic ways of thinking, and worked to make interdisciplinary connections between economics and other content areas. Ayers's study provides insight into what high-quality economics instruction and PCK looks like in three classrooms; however, her work focused exclusively on how teachers approached instruction in economics and not how students experienced the content or what they ultimately did or did not learn. The high-stakes AP context amplifies the complex interactions between instruction and assessment, and while research is silent on what happens in AP Economics classrooms specifically, there is a body of work that explores ambitious instruction in other social studies disciplines across AP.

While researchers question the efficacy and equity of AP (e.g., Gwartney, 2012; MacDonald & Siegfried, 2012), the number of students served by AP remains high (College Board, 2021). In the field of social studies, some research has explored the unique context of AP. For example, Chu (2000) and Lurie (2000) found that AP U.S. History teachers' concerns

about their students' performance on the AP examination led to increased test preparation (e.g., focusing on formulaic responses and acquisition of factual knowledge). A few key studies have explored ambitious instruction in this setting and have provided evidence that high-quality instruction can occur even with the pressures of the end-of-course AP examination. For example, in a case study that explored ambitious teaching in an AP European History course, Brooks (2013) reported that an experienced teacher was able to foster historical understanding while continuing to prepare her students for the end-of-course test. Brooks (2013) observed that this teacher was able to achieve her instructional aim of promoting historical understanding in this context through her "expertise in the field of history, well-developed beliefs about the purpose history can serve students, and familiarity with the AP European History exam" (p. 73). Furthermore, Parker et al. (2011) and Parker et al. (2013) conducted design-based studies in AP U.S. Government and Politics courses in two high schools. The research team worked with teachers at one school to develop five projects to capture the content knowledge of the government course, but, more importantly, to emphasize active learning. During each of these studies, the researchers compared student outcomes on the AP examination between a course that used a project-based instructional approach and one that took a more traditional, lecture-based, instructional approach. They found that students in the course focused on project-based learning performed at least as well as the students in the traditional AP setting, providing evidence that project-based learning and inquiry may provide an equally beneficial alternative to more traditional instructional approaches.

In sum, research on teaching and learning in AP courses in social studies content areas has found that the responsibility to cover large amounts of material can lead to a 'breadth versus depth' dilemma as teachers rely on traditional instructional approaches (e.g., lecture) to cover

content at a rapid pace (Chu, 2000; Lurie, 2000). Teachers with clear instructional aims (Brooks, 2013; Paek et al., 2005) or utilizing project-based learning (Parker et al., 2011; Parker et al., 2013) can generate different learning experiences (and outcomes) for students. These conceptions of best practice and navigating the 'breadth versus depth' divide present unique challenges for the AP economics classroom. The pressure of the AP exam, like any high-stakes end-of-course test, has an impact on the teacher decision-making surrounding classroom-based assessment and, as such, much of what we know about classroom-based assessment in social studies is impacted by the high-stakes accountability era.

Classroom-Based Assessment in Social Studies

Research on assessment in social studies has explored the impact of high-stakes end-of-course multiple-choice tests and has prescribed a vision of what better assessment could be (Grant & Salinas, 2008: Shemilt, 2018). Research that explores external accountability systems and their impact on the social studies classroom (Grant, 2006; Kornhaber, 2004; Shepard, 2001), however, is distinct from, but clearly related to, research that explores social studies classroom-based assessments. As Torrez and Claunch-Lebsack (2013) outline "the extant literature is replete with studies on assessment, testing, and evaluation, yet there is a paucity of empirical research focusing specifically on assessment in the social studies classroom" (p. 462).

Studies on classrooms under high-stakes external accountability systems have found that in states and grade-levels where social studies is tested, assessment drives instruction (Au, 2007; 2009; VanSledright, 2013). Much like the AP setting described above, the presence or absence of a high-stakes test can lead teachers to focus on coverage (i.e., rapidly moving through fact after fact) and control (i.e., narrowly focusing on information in the standards) of tested content (Au, 2007; Barton & Levstik, 2004; Grant & Salinas, 2008). In short, assessment policy acts as a lever

for instruction and a teacher's classroom-based assessment reflects the larger accountability system they inhabit. This tension can lead to a misalignment between the goals of the field of social studies education, the nature of high-stakes and classroom-based assessment, and a teacher's own instructional aims.

Teachers' own decision-making around their use of classroom-based assessments is influenced greatly by the expectations of their context, as Meuwissen (2013) found in his exploration of two teachers' assessment practices across two courses: a flexible elective and a more traditionally structured and tested course. Both teachers in this study had sound aims for their assessment: tests should serve as a feedback loop in the process of student learning. However, when faced when the high-stakes context of a state-tested Government course and an AP Government course, both teachers had to make concessions to their philosophy and navigate a tension between their vision of assessment and their context with students. Navigation of this tension for these two teachers was characterized by transparency with students, reluctant compliance with state and local policy, and pragmatic divergence from the pressures of the high-stakes assessment context. Meuwissen's (2013) study is key as an exploration of how teachers can use adaptive assessment practices to gatekeep (Thornton, 1989, 2006) their students' experience of high-stakes accountability measures in their day-to-day classroom experiences.

While there are a number of studies on how social studies instruction and classroom-based assessment are impacted by high-stakes testing, the body of work on the impact of this context on student thinking and student learning is much smaller. Reich (2009, 2013) conducted think-aloud interviews about questions on the 10th grade New York State Global History and Geography Regents exam with 13 students. Using selected test-items Reich explored student reasoning, thinking, and learning along with examining the validity of the multiple-choice test

questions. The findings of this study indicated a clear misalignment between the broader goals of social studies education (i.e., promoting disciplinary thinking) and the nature of questions on a multiple-choice test, which measured not only content knowledge, but also a student's literacy and *test-wiseness*. Reich's study remains relevant as teachers, especially in the AP setting, feel pressure to prepare their students for the end-of-course examination.

Reich's (2009) work contributed to an ongoing conversation concerning the validity and value of multiple-choice tests as a measure of student achievement. There is some evidence that well-constructed multiple-choice tests can be used as a valid measure of students' academic achievement of a given set of content standards (Haladyna 1999, 2004). However, studies of test-taker reasoning, much like Reich's (2009), show that there is a much more ambiguous relationship between what test-creators believe a test measures and what students actually do when processing information on a test (Nuthall & Alton-Lee, 1995). Because traditional multiple-choice tests measure much more than just content knowledge, such as a student's ability to read and interpret a question (Farr et. al, 1990) or skills of *test-wiseness* that are independent of the intended knowledge and skills being measured (Millman, Bishop, & Ebel, 1965), they should not be used as a 1:1 proxy of student achievement in a given content area. As multiple-choice test items are still used widely both in large-scale and classroom-based assessment contexts, the research that explores student opportunities to learn content and their thinking on summative assessments is of continued importance. Our study adapts a research protocol developed by Graham Nuthall and Adrienne Alton-Lee (1995) in order to understand the interactional relationship between teaching moves, student experiences, student performance on an end-of-unit assessment, and ultimately, student learning.

Nuthall (1999) found that student learning "results from the connections students make between newly evolving knowledge constructs and their background knowledge" (p. 335). Learning was not always based on what was explicitly taught by the teacher, but instead based on the "participation in those classroom activities in which students are required to recall and use their previous knowledge and experiences" (Nuthall, 2000, p. 248). Nuthall remained skeptical of social constructivist teaching promoting "tighter structuring and scaffolding of students' activities" (Brophy, 2006, p. 536). Despite this skepticism, Nuthall (1996) recognized that "every aspect of classroom life is complex, multilayered, and context dependent" (p. 209). Ultimately, Nuthall (1996) argued that student learning of substantive knowledge is a "dynamic interactive system" (p. 210) where "students' access to and participation in the learning activities of the classroom are structured by their negotiation of social status" (p. 211).

Focused on students in elementary and middle school, Nuthall and Alton-Lee (1995) highlight how learning is constructed between teachers and students as well as amongst students (Brophy, 2006). Unlike many other studies on student learning, Nuthall and Alton-Lee examined student learning within a classroom context and explored granular elements of classroom instruction. They investigated the interactional relationship between teaching and learning by working with classroom teachers to create assessments that aligned with the teachers' goals for the unit. Students were given a pre-test, observed during the unit, and given a post-test, and then they participated in a think-aloud interview. Focal students participated in a second think aloud of the post-test a year after the unit was completed to see what substantive knowledge students had retained. Nuthall then created "concept files" for each student and the concept they were supposed to learn. Based on the individual concept file, Nuthall could predict with relative accuracy (80-85%) whether students would answer test questions correctly (Brophy, 2006).

In order to understand Ms. Walter's instruction and student learning, we adapted a research protocol developed by Graham Nuthall and his colleagues (see Nuthall, 2000). Nuthall explored the processes (and outcomes) of how learning occurs and is shaped within and through social contexts; he focused on elementary and middle school students' learning in science and social studies. The methodological protocol employed by Nuthall and his colleagues involved 1) development and implementation of a pretest aligned with a teacher's articulated learning outcomes for a unit of study; 2) daily observations of instruction; 3) identification of 4-6 focal students to be continuously observed and recorded; and 4) post-tests and interviews of students. We adapted his protocol in order to capture, qualitatively, the interactional complexity of how students describe and experience classroom instruction (Nuthall, 1999).

Methods

We used case study methodology to examine how students' classroom experiences in economics related to learning as measured by an AP-style, teacher-constructed summative assessment. We were interested in what the students experienced during the unit on market structures, what the teacher's instructional aims for the unit were, and how students performed on the assessment; thus, the bound for the case study was the unit of study on market structures (Yin, 2018).

Context and Participants

The study took place in an AP Economics course at Valley High School (pseudonym), a large public school in the Commonwealth of Virginia. The school enrolls a diverse student population of approximately 1,900, ninth- through twelfth-grade students. The AP Economics course is an increasingly popular course for students due to a law in Virginia that requires students to take either an economics or a personal finance course prior to graduation (Code of

Virginia §22.1–200.03). Valley offers two courses for students to fulfill this requirement, personal finance (i.e., a semester-long course offered for seniors) or AP Economics. AP Economics is open to any student enrolled at Valley and is popular across grade levels, attracting sophomores, juniors, and seniors. Students are not required to take the AP test at the end of the course, but those who do so receive additional points on their grade point average (GPA) and have the potential to transfer credit to the college or university of their choice.

Ms. Walter, the AP Economics teacher at Valley, had been teaching AP Economics for three years. She was the only person in the 22-person social studies department who "had any type of major/minor [...] in econ" (Interview, 3/31/2016). When the previous AP Economics teacher retired, Ms. Walter was the natural choice to take over AP Economics as she had purposefully taken several economics classes as "job security" for her future career as a social studies teacher. When asked to describe her approach to teaching this course, Ms. Walter stated that she tried to "limit [her] amount of lecture time" and instead used more "simulations and case studies." She started each unit of instruction with content that students knew and worked toward "what they're less familiar with…so they can build on what they know" (Interview, 2/23/16). This classroom was of interest to the research team as Ms. Walter was recommended as a teacher who engaged in best-practice instruction (i.e., student-centered teaching, discussion, sparse lecture) in the social studies classroom context.

For this study, we also purposefully followed a small group of students, referred to as 'focal students,' to closely trace the classroom experiences and interactions with the teacher and classmates they had (Nuthall & Alton-Lee, 1995). From a class of 18 students, 4 students were selected as focal students, identified by the following pseudonyms: Allison, Patrick, Michelle, and Sean (See Table 1). These students were selected because they represented the class both in

terms of gender (i.e., eight females and ten males) and grade level (i.e., eleven seniors, three juniors, and four sophomores). It should be noted that while there were three juniors enrolled in the course, none of them consented to be in the study. Each of the four focal students were consented through the institutional review board (IRB).

Table 1

Student Participants

Name	**Age/Grade**	**Reason for taking AP Economics**	**Other social studies courses currently enrolled in**
Allison	18/12	"This is about to sound really bad, but because you needed personal finance or, like, an econ class to graduate, and so a couple of my friends told me about it last year, and I was like, "oh..." They said you do a lot of projects and stuff like that, so that's why I signed up"	AP Government
Patrick	18/12	"Personal finance is required for... to graduate for our class. We're one of the first classes that have to take it, and last year was the first year. And your choices are standard personal finance and AP Economics. So, a lot of us that are looking at big schools…But I know the AP looks so much better than a standard class so..."	AP Government
Michelle	15/10	"I wanted to take one AP class this year, but I didn't really want to take AP World History because, like, I don't know, I don't really like history that much, so I was like 'I'll take this.' And my brother took it last year because he's a junior"	None
Sean	17/12	"The fact that it would look good on a college resume. A lot better than, so I think, than personal finance. And I decided it'd be a little more interesting than regular personal finance, which I feel I already sort of know really well. Just through Boy Scouts, we had to take	AP Government

		a lot of merit batches, and one of them focused on personal finance specifically."	

Data Collection

Data collection included the following: two semi-structured interviews with the teacher, 11 ninety-minute video-taped classroom observations (two cameras throughout the classroom as well as detailed field notes), a semi-structured interview with each individual focal student (pre- and post- assessment data), and collection of all classroom artifacts (i.e., student notes, PowerPoint slides, lesson materials, etc.). The first interview with Ms. Walter occurred approximately one week before the unit began and focused on Ms. Walter's instructional aims for the unit. Ms. Walter also shared her post-assessment with the researchers and discussed her assessment approach.

Ms. Walter planned to assess the students through a teacher-constructed, AP-style test that included multiple-choice and free response questions as well as questions that required the analysis of graphs. The majority of her assessment questions came from released AP tests. It should be noted that the post-assessment had more items than the pre-assessment (aligning with Ms. Walter's instructional aims); for the purposes of analysis, we only compared the items that appeared on both (i.e., seven multiple choice questions). In addition to the test, Ms. Walter had the students write an analytical essay based on the documentary that they watched in class (*Schooled: The Price of College Sports)*. Through the essay, the students had to analyze what market structure they believed the National Collegiate Athletic Association (NCAA) most closely resembled. Students also engaged in four quizzes early in the unit: the first two quizzes mimicked the multiple-choice style of the AP-style while the second two were short answer

questions regarding oligopolies and game theory. The quizzes were graded and returned to students, but not reviewed over the course of the unit. Ms. Walter's unit plan engaged students in a number of instructional approaches to support students in learning the characteristics of market structures (See Table 2).

Table 2

Unit Overview

Day	**Content**	**Instructional Strategies**
Day 1	Unit Overview Perfect Competition	• Pre-Assessment • Unit anticipation chart • PowerPoint Lecture • Unit Packet: Graphing
Day 2	Monopolies	• Review: Perfect Competition • Unit Packet • Quiz • PowerPoint Lecture
Day 3	Oligopoly	• Review: Monopolies • Unit Packet • Quiz • PowerPoint Lecture • Video: *King of the Hill* (Backchannel chat)
Day 4	HHI: Oligopoly v. Monopoly v. Perfect Competition	• Quiz • Monopoly Board Game • Calculate Herfindahl-Hirschman Index (HHI) based on game results

Day 5	Game Theory and Nash Equilibrium	• Review: Market Structures • Video: Crash Course • Game Theory Matrix Practice • Video: excerpt from *A Beautiful Mind* • Skit about game theory • Video: British Game Show
Day 6	Game Theory Monopolistic Competition Market Analysis	• Present skits about game theory • Quiz • Unit Packet • PowerPoint Lecture • Market Analysis Project • *Plays local college's basketball game in background*
Day 7	Review	• Market Analysis Presentation • Gallery Walk: Market Structures
Day 8	Test	• Test
Day 9	Market Structures	• Video: *Schooled: The Price of College Sports*
Day 10	Market Structures/*Schooled*	• Socratic Seminar
Day 11	Market Structures	• Work Day: *Schooled* Paper

The second interview with Ms. Walter took place about three weeks after the unit's completion. During this interview, Ms. Walter engaged in a think aloud of the post-assessment where she was asked to talk through each test item. She anticipated students' responses and specified where and when she thought the students interacted with the information necessary to answer the test items correctly (i.e., textbook, lecture, group project, etc.).

Semi-structured interviews with the focal students occurred approximately two weeks after the unit's conclusion and focused on the students' backgrounds as well as a think aloud of why and how they answered questions from the post-assessment. Specifically, students were asked to cite their learning experiences inside and outside of class to describe why and how they answered for each test item. Follow-up questions were used to probe any student misconceptions or to clarify answers.

Data Analysis

Data analysis occurred through multiple, iterative rounds. The first round of data analysis consisted of a holistic read of all data sources to capture emergent themes. We then analyzed the post-assessment based on the AP Microeconomics topic outline (See Table 3). Each question was labeled with a concept descriptor that represented the knowledge students would need to know in order to correctly answer the test item.

Table 3

Assessment questions

Question	**Content Descriptor**	**AP Standard**
Question 5	Perfectly Competitive Markets/MR=MC/graphing	II.D.1.c II.D.2.a
Question 8	Perfectly Competitive Markets/MR=MC/graphing	II.D.1.c II.D.2.a II.D.2.d
Question 17	Pure monopolists demand curve/graphing	II.D.3
Question 33	Price Discrimination	II.D.3.d
Question 34	Characteristics of an oligopoly	II.D.4.a

Question 35	Characteristics of a monopolistically competitive firm	II.D.5

During the next round of analysis, we analyzed the student performance on the pre- and post-assessments as well as the interview think aloud of the post-assessment based on codes developed by Nuthall and Alton-Lee (1993). Each student assessment (pre, post, interview) was coded as to whether the students knew the information prior to the unit, learned it during the unit, never learned the information, or learned and forgot the information. We then created "item files" that consisted of the assessment and the instructional interactions students had with the content during the unit (See Table 4). This involved line-by-line coding of all data by content topic in Dedoose, a password-protected qualitative software program.

Table 4

Item files

Concept Descriptor	Number of test items	Days Taught	Instructional Approach
Perfectly Competitive Markets/MR=MC/graphing	2	Day 1 Day 2	PPT/Lecture Graph Packet Quiz
Pure monopolists demand curve/graphing	1	Day 2	PPT/Lecture Graph Packet Quiz
Price Discrimination	1	Day 3	PPT/Lecture Market Analysis Project
Characteristics of an oligopoly	2	Day 3 Day 4 Day 6	PPT/Lecture *King of the Hill* *A Beautiful Mind* clip Group Project—Skit Graph Packet Market Analysis Project Quiz

Characteristics of a monopolistically competitive firm	1	Day 6	PPT/Lecture Graph Packet
Price/Non-Price Competition	1	Day 3 Day 6	PPT/Lecture *King of the Hill* Market Analysis Project

Our next round of coding focused on the student interviews and Ms. Walter's post-unit interview for how they described their experiences in the classroom. Based on the instructional approaches in the item files, we coded the interviews for what instructional approaches students identified as contributing to what they learned as compared to what Ms. Walter stated in her interview. In our final round of coding, we used the AP economics standards (i.e., as defined by College Board) to understand when and how content was covered throughout the unit. Not only did this process ensure that all assessment items were covered at some point during the unit, but it also illuminated the facets of the framework appropriated by the teacher to explain certain economics concepts and necessitated by the post-assessment to show mastery of the content.

Findings

Analyses of student answers on the multiple-choice assessment paints a complex and varied portrait of what students learned or did not learn during the unit (See Table 5). The class average for the assessment was a 'C,' which Ms. Walter stated was "good for an Econ class," but this assessment "was a higher average with a C range…but only slightly higher so probably within the realm of normal statistics" (3/31/16, Interview). Ms. Walter was pleased with how the unit went and felt that "in general the class seemed to get it…they got most of their objectives" (3/31/16, Interview). Students were able to answer questions correctly about the market structures they learned early in the unit (i.e., perfectly competitive markets and monopolies), but

struggled to answer questions about content they learned later in the unit (i.e., oligopolies and monopolistically competitive firms).

Table 5

Student Performance on Assessment Questions (Correct Answers/Total Questions)

Student	Pre-Assessment	Post-Assessment	Interview
Allison	5/7	4/7	3/7
Patrick	3/7	4/7	3/7
Michelle	5/7	6/7	4/7
Sean	3/7	4/7	3/7

For the four focal students, their scores on three assessments (i.e., pre-assessment, post-assessment, and think-aloud interview) varied. Three of the four focal students scored higher on the post-assessment by one question, but by the time of the think-aloud interview, they scored the same or lower than they had on the pre-assessment. One student, Allison, had her score decline with each successive assessment.

The student assessment data presents a mixed picture as to whether, or what, students learned during the unit. Analysis of these questions demonstrates that the nature of the question (i.e., fact-based, application of knowledge) appeared to matter. Students performed better on factual questions but struggled on more complex questions that asked them to use procedural knowledge (i.e., knowledge required to 'do economics' such as math computations, data visualization, and graphing; Ayers, 2018) or on two-step multiple choice questions. The student interviews highlighted the individuality of student learning. Each of the students described their experiences of the classroom instruction in different ways. In order to further explore what economic content knowledge, understandings, and skills students learned and how they learned it

during the unit, we highlight four assessment questions: two questions that all students answered correctly and two questions that captured a variety of student responses. After exploring student thinking on each of these questions, we then look back to the unit to interrogate whether students had an opportunity to learn the tested economic content knowledge and understandings.

Questions 5 & 33: Recalling Information. All four students answered two assessment questions correctly, questions 5 and 33 (See Figure 1). Both of these questions could be described as fact-recall; students had to retrieve a specific piece of economic content knowledge in order to answer the question correctly. In question 5, students could answer correctly if they remembered that price equals marginal revenue (P=MR) for a perfectly competitive firm. In question 33, students had to remember the definition of price discrimination to select the correct answer. Both questions were straightforward examples of economic content knowledge, as students did not have to use procedural knowledge, combine different economic concepts, or employ economic reasoning to correctly answer the question.

Figure 1

Assessment Questions

Question 5.
For a perfectly competitive firm, if the market price is $8, then

A. marginal revenue is greater than $8.
B. marginal revenue is less than $8.
C. marginal revenue is equal to $8.
D. average revenue is greater than $8.
E. average revenue is less than $8.

Question 33.
Which of the following is true of monopolists that practice price discrimination?

A. They charge all customers the same price.
B. They earn a smaller profit than those that do not practice price discrimination.
C. They charge customers different prices according to different elasticities of demand.
D. They produce lower quantities than pure monopolists.
E. They produce the same quantity of output as pure monopolists.

Note: Correct answers are in bold.

For example, on question five, Allison stated that she knew the answer to the question because "marginal revenue equals margin, like price" (3/31/16, Interview). When asked if she remembered the answer coming up during class, Allison responded "we actually wrote it on like everything that we did…in my notes, I definitely wrote it and boxed it in because [Ms. Walter] said it was important" (3/31/16, Interview). Similarly, Michelle noted, "Because we learned that, like, MR=MC and I feel like for perfectly competitive firm would be equal and not, like, greater or less than." Sean recalled seeing this information in "a graph" that Ms. Walter drew, and Patrick recalled "taking notes in class and things like that and the graphs." When reflecting on the unit, Ms. Walter stated that she expected the students to learn this on the first day of the unit when "we talked about perfect market structure…and we created the graph" (3/31/16, Interview). She felt that the students would have interacted with P=MR "multiple times throughout [the unit], whenever we reviewed the market structures" (3/31/16, Interview).

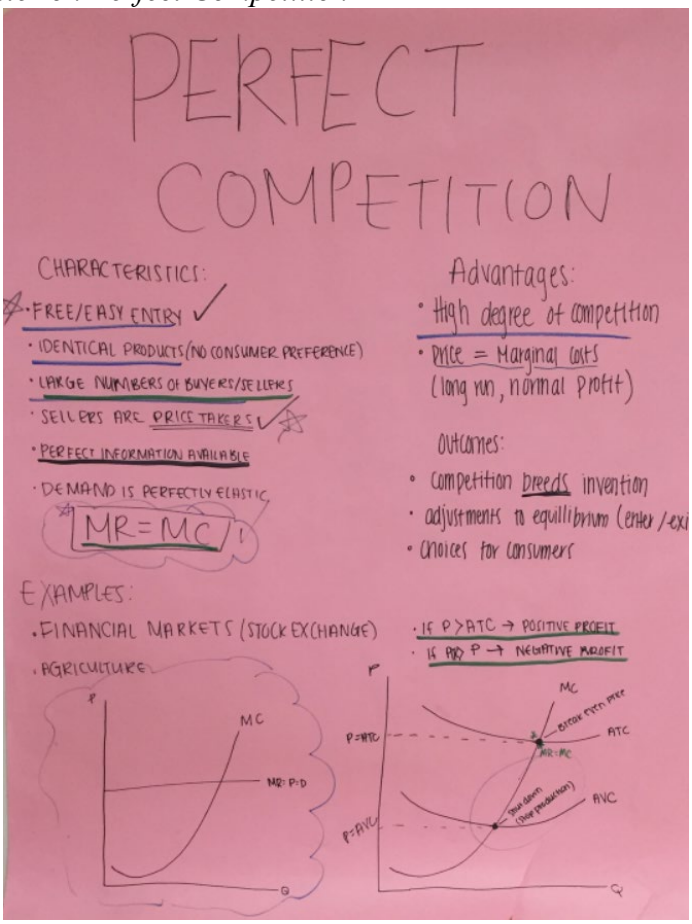
On question 33, each of the students provided a definition of price discrimination before describing when they remembered the term coming up during their class. Sean said, "I remember price discrimination is like charging different prices" (3/31/16, Interview), while Michelle defined price discrimination as occurring when companies "change the price based on different things and they're kind of unfair" (3/31/16, Interview). Allison and Sean remembered Ms. Walter providing examples of companies that used price discrimination. Sean knew that "airlines…give different ticket prices" (3/31/16, Interview), but did not remember exactly how an airline company used price discrimination. Patrick stated that "price discrimination isn't something difficult to figure out," but that they had discussed this topic in his government class, and in AP Economics, it was discussed in *King of the Hill* "for like twenty minutes" (3/31/16,

Interview). Michelle remembered the topic coming up, but was unable to pinpoint when or where she had heard about it, stating "we were talking about monopolies or oligopolies or something, one of those, I remember price discrimination was one of the slides" (3/31/16, Interview). In her reflection on the unit, Ms. Walter believed the students should have learned about price discrimination "when we talked about the different strategies that monopolists could use in order to increase their profits, I think that was day two" (3/31/16, Interview). Ms. Walter also mentioned that students could have engaged in price discrimination when they did their market analysis project; however, none of the focal students used price discrimination in their analysis.

Questions 5 & 33: Opportunities to Learn. Classroom observations, in this case, largely aligned with Ms. Walter's and the students' description of where they learned material in class. For question 5, students were introduced to P=MR during the first days of the unit as Ms. Walter covered perfect competition; the rule also appeared continually throughout the unit. For example, students reviewed this economic concept when they created posters summarizing the main characteristics of each market structure on the seventh day of the unit. The group that reviewed a perfectly competitive market wrote this equation prominently on their graph (See Figure 2).

Figure 2

Student Review Poster on Perfect Competition



For question 33, the primary time when students interacted with price discrimination was on the second day of the unit when Ms. Walter introduced the concept of monopolies and spent 10 minutes of a 40-minute lecture on price discrimination, providing examples and non-examples of how price discrimination works. During their poster review of the unit, the group describing the features of a monopoly wrote that they "can price discriminate," but did not provide an example or a definition as to what price discrimination is. It is also important to highlight that the example Sean was able to give of price discrimination (airlines giving different ticket prices) was a not an example explicitly covered by Ms. Walter during her classroom instruction, but rather from the course text.

Across both assessment questions, the four focal students had a foundational understanding of the economic concept or term and were able to correctly recall the content knowledge. In most cases, the four students pointed to similar moments in the class when they remembered the content knowledge being discussed. The students most frequently referenced Ms. Walter's lecture and teacher-centered instruction as the source of information they drew upon to answer these assessment questions. There also seemed to be alignment between when and where Ms. Walter thought the students would learn the information and when and where the students said that they learned the information. Ms. Walter had clear memories of lecturing on the information and providing the definition to students.

Questions 17 & 37: Mixed Results & Multiple Steps. For two of the questions on the assessment, questions 17 and 37, students had difficulty answering correctly and interviews revealed greater variability in student descriptions of where and how they learned information. Questions 17 and 37 asked students to engage in conceptual thinking and had the students connect various economic concepts and employ procedural knowledge (See Figure 3). To answer question 17 correctly, students had to know that in a pure monopoly, firms face a dilemma; to maximize revenue, they must either lower prices to sell more units or increase prices and face less marginal revenue. While this question may be conceptually complex, a visualization of the graph of a pure monopoly would greatly simplify the process of answering this question; however, one was not provided. That is, if a student can recall the graph of a monopolist's demand curve, they could clearly visualize that the demand curve physically "lies above" the marginal revenue curve. In question 37, students had to understand the definition and characteristics of an oligopoly and apply it to a list of statements. Question 37 was a two-step

multiple-choice question in which students had to determine which of three statements were correct.

Figure 3

Assessment questions

Question 17. Which of the following is true of a pure monopolist's demand curve? A. It is perfectly inelastic. B. It is perfectly elastic. C. It coincides with its marginal revenue curve. D. It lies below its marginal revenue curve. **E. It lies above its marginal revenue curve.**
Question 37. Characteristics of an oligopoly include which of the following? I. Collusion can increase oligopolists' profits. II. Oligopolistic firms are interdependent. III. Independent price decision making leads to lower returns. A. I only B. II only C. III only D. I and II only **E. I, II, and III**

Sean was the only student who has able to correctly answer question 17. Sean stated that he knew the answer because "the demand curve clearly is above marginal revenue" implying his ability to visualize the graph in order to answer the question (3/31/16, Interview). Sean remembered seeing the graph that Ms. Walter presented during class, but also from the textbook. On the pure monopoly quiz, Sean scored a 1/5 and outlined the textbook chapter as a way to raise his quiz score. When studying for the assessment, Sean said that he always made sure to review all of the graphs "because I think they appear a lot on the test, they're important" (3/31/16, Interview). The other three students appeared to have some misconceptions about the definitions of terms and seemed unsure about how to answer the question. Patrick oscillated

between choices A (inelastic) and B (elastic) stating "it's just the process of me remembering if inelastic is the one where you need it or is elastic the one where you need it." Ultimately, Patrick chose A "because that would make me think, you know, there's people willing to bargain for it…it's not like they could go without it. They need whatever the product is" (3/31/16, Interview). In this response, Patrick demonstrates a misunderstanding of the key vocabulary terms in the question and attempts to erroneously apply his concept of perfectly inelastic demand to a monopolist's demand curve.

Allison and Michelle faced similar difficulties as Patrick and could not define the vocabulary terms in the question. The three students also struggled to remember when or if they discussed the concept in class. Allison knew that they discussed monopolies in class and that a monopoly occurred when "there's no other options, so it's just the one [company] that controls everything else," but could not remember when they discussed the demand curve. Patrick said he relied on his knowledge of elasticity that was from a previous unit where they learned "how much a company wants to make…or how much people want a product with supply and demand affecting that" to answer this question and had no memory of the content of the question coming up in class for this particular unit. When Ms. Walter reflected on this question, she immediately responded "graphs, it's all about the graphs" that the class had looked at while she lectured on monopolies.

All four students incorrectly answered question 37. Michelle answered B—that oligopolistic firms are interdependent. She had a misunderstanding of oligopolies and what "interdependent" means. She selected this answer "because there are few firms and they're focused on themselves" (3/31/16, Interview). Michelle did not remember oligopolies coming up in class or any instances of discussing the concept and said, "I just kind of guessed." The other

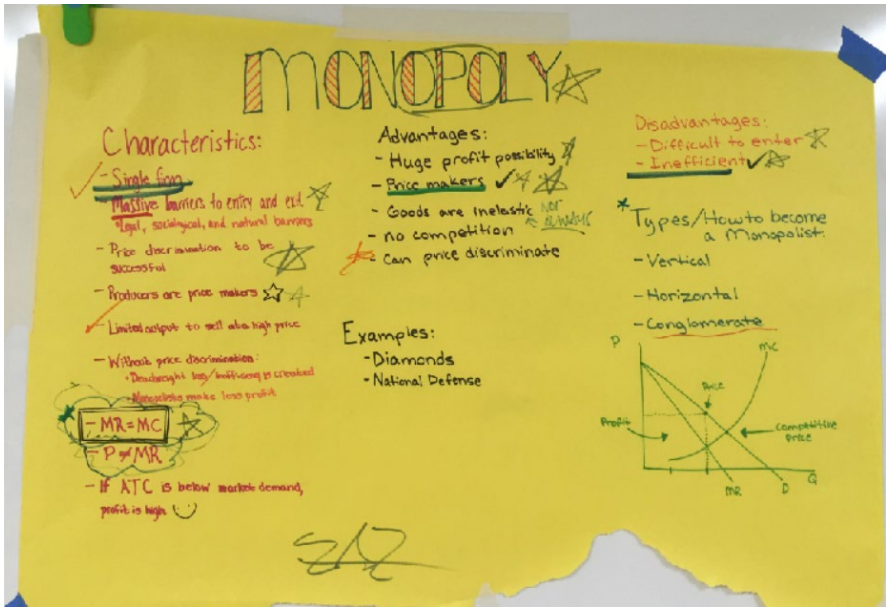
three students all selected D, which did not include III (independent price decision making leads to lower returns). Patrick talked through each Roman numeral, selecting I (collusion can increase oligopolists' profits) because of the episode of *King of the Hill* they watched in class. He rationalized why II (oligopolistic firms are interdependent) had to be correct because he knew that "if one firm lowers their prices because they're so few, the rest…have to lower their prices as well" (3/31/16, Interview). III "threw" Patrick "off a bit because if you lower prices and the others don't, it could get you more returns in the short term just because more are going to leave the other firms and come to you" (3/31/16, Interview). In his answer, Patrick understood the "lower returns" to apply to one particular company rather than to all firms within an oligopoly. Sean and Allison had similar answers to Patrick and were confused by what III meant. Even though option III is an extension of interdependence, which all students related to oligopoly by selecting II, they were unable to understand that oligopolists, through game theory, must work together to set prices. If a firm sets its price independent of the other firms in an oligopoly, a net loss will occur, thus lower returns in the long run.

Ms. Walter stated that the question asked students to "distinguish very specific things about oligopoly" (3/31/16, Interview). She remembered covering the content frequently throughout the unit and presenting the information through a variety of methods, including her lecture on the basic characteristics and the episode of *King of the Hill.* She also mentioned the documentary, *Schooled: The Price of College Sports* as how students might have learned this information, but the students took the assessment prior to watching the film. For questions 17 and 37, students had difficulty connecting the content asked by the test question with specific moments of classroom instruction, data analysis further exposed why this connection may have been tenuous.

Questions 17 & 37: Opportunities to Learn. Regarding question 17, concerning a pure monopolist's demand curve, on the second day of the unit, Ms. Walter had lectured about monopolies and showed the students a series of graphs of when a monopolist makes a profit, when a monopolist breaks even, and when a monopolist loses money. This five-minute excerpt of the lecture, when the graphs were first introduced, was the only time Ms. Walter directly related the two curves. She told the class "Marginal revenue is below what could be considered the price. The marginal revenue line is usually set by the natural market price, but it doesn't equal the demand. It makes more sense graphically" (3/11/16, Observation). During this lecture, she never explicitly stated that the demand curve is above the marginal revenue curve. Students were at least exposed to this graph a second time on their review day before the unit test because the group that created the poster on the characteristics of monopoly did include this visualization (See Figure 4)

Figure 4

Student Review Poster on Monopoly



When analyzing this assessment question and student interview data, it was clear that students understood the definition of a monopoly and how it interacts with a market but missed the question because it specifically related to graphing and visualization of the concept. That is, students lacked some procedural knowledge to make use of what they knew about monopoly. As mentioned earlier, Sean, the only student to answer this question correctly, went beyond the expectations of regular class participation in order to outline a textbook chapter to boost a quiz grade and was exposed to this content in a different format.

A large amount of class time was dedicated to the content covered by test item 37, concepts related to the characteristics of oligopolies. Analysis of classroom observations indicated that the class spent more time on oligopolies than the other three market structures with Ms. Walter devoting almost four full class periods to the subject. During these four days, the students interacted with the material in a variety of ways. Ms. Walter lectured on the foundational characteristics of oligopoly and game theory. The students also played Monopoly, watched video clips (*King of the Hill, A Beautiful Mind*), created a skit about game theory, and examined different scenarios and matrices about game theory. Most of this instruction was based in real world application of the knowledge, such as the skit about game theory and the problem-solving scenario with the prisoners' dilemma. Students were asked to think about these concepts in a novel way and, during class, were able to demonstrate their understanding of the concepts. However, when thinking aloud on the test, none of the students made the connection between game theory or Nash equilibrium and the interdependent nature of oligopolies. Interestingly, this was the quiz on which three students (Sean, Michelle, Allison) scored a 4/4, but they did not associate their learning on that day or the quiz with this assessment

question. It is important to note that the format of this quiz did not match the format of the end-of-unit assessment.

The disconnect between student responses to question 37 and their memory of classroom experiences may also point to the structure of the AP-style test item as a source of misalignment and tension. Classroom observations and post-unit interviews provide evidence that students experienced activities related to the characteristics of oligopolies and, subsequently, had an acute understanding of their characteristics. Multiple students referenced the *King of the Hill* episode and the concurrent backchannel chat in their post-unit interviews as they accurately defined the concept of collusion and the characteristics of an oligopoly. When presented with a multi-step question however, students faltered and were not given an avenue to provide evidence they had learned this content. One student, Sean, engaged in construct irrelevant reasoning when engaging with question 37, eliminating III from contention because it "doesn't look right" (3/31/16, Interview).

Discussion

Ms. Walter represented a teacher who engaged in best practice instruction and modeled the components of pedagogical content knowledge for economics (Ayers, 2018). She engaged students with instruction that pushed them beyond the foundational content knowledge of AP Microeconomics and invited them to experience content through a multi- and inter- disciplinary approach; she worked to make content relevant to students' lives, asked students to participate in discussion about economics topics, and designed classroom experiences (e.g., projects) that were active and student-centered. However, student scores on an AP-style end-of-unit assessment in conjunction with students' responses to this assessment in post-test interviews indicate a misalignment between instruction, assessment, and student learning.

Students did not perform well on the end-of-unit assessment on market structures. However, our findings indicate the nature of this poor performance points to a more complex story than instructional failure. There is no doubt, from our post-test interviews, that students learned content associated with AP Microeconomics. The test was able to capture some of this knowledge, but it was not able to capture all of it. The test also revealed weaknesses in students' knowledge of AP Microeconomics, which were exacerbated by the style and structure of the AP-released test questions. These questions sometimes pushed students to engage in "construct irrelevant" reasoning (Smith et al., 2019), and asked students to practice disciplinary skills (i.e., procedural economic knowledge; Ayers, 2018) which were not explicitly emphasized during instruction.

Ms. Walter's goals, even as she defined them in the pre-unit interview, were never explicitly about student learning. When Ms. Walter talked about her hopes for students during the unit, she spoke in terms of what experiences she wanted students to have (e.g., limited time with lecture, simulations) rather than what she wanted students to know, understand, and be able to do (i.e., learning objectives). The end-of-unit assessment that Ms. Walter constructed using AP released test items communicated a different goal, success on the end-of-course examination. This goal, while clearly important to the students and Ms. Walter based on their interviews, was rarely discussed during class. Students were measured on outcomes that were not fully represented through instruction in this AP Microeconomics classroom (i.e., performance on a multiple-choice test) and were, therefore, unable to practice or receive feedback on their work towards these outcomes. Students did engage in two practice quizzes that mimicked their end-of-unit test, but they did not received feedback outside of their score on these five question quizzes. These findings indicate the important distinction between students' classroom experiences and

learning outcomes and between a teachers' instructional goals for their students and their learning goals.

Further, Ms. Walter's decision-making was impacted by a number of factors including her understanding of best practice social studies instructional methods, her own philosophy of teaching, the AP curriculum defined by College Board, and the pressures of the end-of-unit test. The pressures Ms. Walter faced and how she interpreted and adapted the course in response to them corroborate Meuwissen's (2013) findings that place teachers at the center of assessment and curriculum sensemaking. Teachers distill innumerable pressures to create a particular classroom experience for their students. These findings further emphasize the importance of teacher's role as gatekeeper for student experiences (Thornton, 1989, 2006) but suggest an expansion of this role to include how students face the pressures of both classroom-based and large-scale assessment.

Our findings build on the body of literature on student thinking about their learning on multiple-choice tests. Students were successfully able to answer questions that required fact-recall of foundational economic concepts. Further, students and their teacher all pointed to similar moments in class when this content was learned. Corroborating the work of Reich (2009, 2013) and Smith and colleagues (2019), we also found that students engaged in 'construct irrelevant' reasoning unrelated to economic concepts and skills to answer test questions. Student success or failure on the multiple-choice test items revealed more complexity than a disconnect between what they had learned about economics and how they were able to perform on the test. Student think-alouds revealed a lack of experiences in class that aligned with the skills required to successfully answer questions on the end-of-unit assessment. In particular, students were not exposed to economic ways of thinking that were necessary to success. These findings extend the

argument of Ayers (2018) that successful PCK in economics focuses not just on knowledge of economics content and concepts or interdisciplinary knowledge, but also procedural knowledge that draws on math, graphs, and data visualization.

Implications and Conclusions

The student experience of this unit of study and their sensemaking of the end-of-unit assessment have important implications for teaching, teacher education, and future research on the connections between instruction and student learning in the secondary social studies classroom. This case study indicates there may be a need for support for teachers in the field with regards to alignment between instructional planning, instruction, and assessment. Professional development for in-service teachers should focus on the connections between best practice instruction and assessment. Further, the reality of the classroom and the broader policy context cannot be ignored as teachers face the pressures of measuring student knowledge and new assessment initiatives that ask them to engage in project-based and performance-based learning. In-service teachers need support in evaluating test items, designing their own test items, aligning test items to their instruction and their goals, and in teaching students to think about testing as a part of their learning process. More assessment-based research like the work of Reich (2009; 2013) and Smith and colleagues (2019) is necessary to unpack the complex questions of alignment raised in this manuscript.

In teacher education, this case study pushes us to think deeply about how we expose pre-service teachers to the concept of assessment and how this relates to their understanding of instructional planning and enactment. Assessment in teacher education programs is often treated as content neutral; that is, courses are often offered for assessment across the secondary content areas (social studies, science, mathematics, and English language arts). This model of instruction

ignores that high quality assessment, much like high quality instruction, is disciplinary in nature. When assessment is woven into our understanding of instruction, pre-service teachers can be taught disciplinary instructional moves alongside disciplinary methods of assessment to document student progress toward success in discipline-specific content knowledge, understandings, and skills. As pre-service teachers leverage social studies best practices such as discussion, inquiry, and source analysis, they should also learn disciplinary ways to measure student growth and success in these realms. In addition, teacher educators and researchers should explore how these disciplinary ways of knowing and assessing may vary across disciplines within social studies (e.g., economics, history, geography, civics, etc.). While much high-quality work has been done to understand what it means to "know" and "do" history or civics, much less is known about disciplinary knowledge in economics.

Ms. Walter's classroom raises important questions about how teachers make sense of student learning while navigating an externally imposed assessment context. Research should continue to explore more than the actions of just teachers or just students in the secondary social studies classroom, but the instructional dynamic that connects them (Ball & Forzani, 2007). Teachers and schools now have more flexibility with regards to the types of classroom assessments they employ due to shifting federal and state policy that endorses the use of portfolios and performance assessments (Darling-Hammond et al., 2016). Future research should explore how teachers and students navigate this new policy context and how this impacts a classroom level understanding of learning.

References

Abrams, L.M, Pedulla, J.J., & Madaus, G.F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice*, *42*(1), 18-29. https://doi.org/10.1207/s15430421tip4201_4

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258-267. https://doi.org/10.3102/0013189X07306523

Au, W. (2009). Social studies, social justice: W(h)ither the social studies in high-stakes testing?. *Teacher Education Quarterly*, *36*(1), 43-58. https://www.jstor.org/stable/23479200

Ayers, C. A. (2018). A first step toward a practice-based theory of pedagogical content knowledge in secondary economics. *Journal of Social Studies Research*, *42*(1), 61–79. https://doi.org/10.1016/j.jssr.2017.01.003

Ball, D. L., & Forzani, F. M. (2007). What makes education research" Educational?" *Educational Researcher*, *36*(9), 529.

Barton, K. C., & Levstik, L. S. (2004). *Teaching history for the common good*. Routledge.

Barton, K. C., & Avery, P. G. (2016). Research on social studies education: Diverse students, settings, and methods. In D. Gitomer & C. Bell (Eds,), *Handbook of research on teaching* (pp. 985-1038). American Educational Research Association.

Brooks, S. (2013). Teaching for historical understanding in the Advanced Placement program: A case study. *The History Teacher*, *47*(1), 61-76. https://www.jstor.org/stable/43264185

Brophy, J. (1990). Teaching social studies for understanding and higher-order application. *The Elementary School Journal, 90*(4), 351-417. https://www.jstor.org/stable/pdf/1001938.pdf

Brophy, J. (2006). Graham Nuthall and social constructivist teaching: Research-based cautions and qualifications. *Teaching and Teacher Education 22*(5), 529 – 537. https://doi.org/10.1016/j.tate.2006.01.008

Brophy, J., McMahon, S., & Prawatt, R. (1991). Elementary social studies series: Critique of a representative example of six experts. *Social Education, 55*(3), 155-160.

Brozo, W.G., & Tomlinson, C.M. (1986). Literature: The key to lively content courses. *The Reading Teacher, 40*(3), 288-293. https://www.jstor.org/stable/pdf/20199384.pdf

Butters, R.B., & Asarta, C.J. (2011) A survey of economic understanding in U.S. high schools. *The Journal of Economic Education 42*(2), 200-205. https://doi.org/10.1080/00220485.2011.555723

Chu, J. (2000). Preparing for the AP Exam: The dangers of teaching to the test. *The History Teacher, 33*(4), 511-520. https://doi.org/10.2307/494947

College Board. (2021, November 4). *AP program participation & performance data 2021*. Retrieved January 30, 2022, from https://research.collegeboard.org/programs/ap/data/participation/ap-2021

Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E. L. (2016). Pathways to new accountability through the Every Student Succeeds Act. *Palo Alto, CA: Learning Policy Institute*.

Ercikan, K., & Seixas, P. (2015). *New directions in assessing historical thinking*. Routledge.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*(3), 209-226. https://www.jstor.org/stable/pdf/1434927.pdf

Fitchett, P.G. & Heafner, T.L. (2010). A national perspective on the effectiveness of high-stakes testing and standardization on elementary social studies marginalization. *Theory and Research in Social Education, 38*(1), 114-130. https://doi.org/10.1080/00933104.2010.10473418

Fitchett, P. G., Heafner, T. L., & Lambert, R. (2014). Assessment, Autonomy, and Elementary Social Studies Time. *Teachers College Record*, *116*(10), 10.

Goodlad, J. (1984). *A place called school: Prospects for the future*. McGraw-Hill.

Grant, S.G. (Ed.). (2006). *Measuring history: Cases of state-level testing across the United States*. Information Age Publishing.

Grant, S.G., & Salinas, C. (2008). Assessment and accountability in the social studies. In L.S. Levstik & C.A. Tyson (Eds.), *Handbook of research in social studies education* (pp. 219–238). Routledge.

Guzzetti, B.J., Kowalinski, B.J., & McGowan, T. (1992). Using a literature based approach to teaching social studies. *Journal of Reading, 36*(2), 114-122. https://www.jstor.org/stable/40016443

Gwartney, J. (20012). What should we be teaching in basic economics courses? *The Journal of Economic Education, 43*(3), 300-307. https://doi.org/10.1080/00220485.2012.686398

Haladyna, T.M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Erlbaum.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Erlbaum.

Kornhaber, M.L. (2004). Appropriate and inappropriate forms of testing, assessment, and accountability. *Education Policy, 18*(1), 45-70. https://doi.org/10.1177/0895904803260024

Koretz, D. (2008). *Measuring up.* Harvard University Press.

Lurie, M. N. (2000). AP U.S. history: Beneficial or problematic? *History Teacher*, *33*(4), 521–525. https://doi.org/10.2307/494948

MacDonald, R. A., & Siegfried, J. J. (2012). Refreshing the voluntary national content standards in economics. *Journal of Economic Education*, *43*(3), 308–314. https://doi.org/10.1080/00220485.2012.686779

Martin, D.M., Maldonado, S.I., Schneider, J., & M., Smith (2011). *A report on the state of history education: State policies and national programs.* https://doi.org/10.13140/RG.2.2.12801.40802

Maxwell, N. L., Mergendoller, J. R., & Bellisimo, Y. (2005). Problem-based learning and high school macroeconomics: A comparative study of instructional methods. *The Journal of Economic Education, 36*(4), 315-331. https://doi.org/10.3200/JECE.36.4.315-331

Mergendoller, J. R., Maxwell, N. L., & Bellisimo, Y. (2001). Comparing problem-based learning and traditional instruction in high school economics. *The Journal of Educational Research*, *93*(6), 374-382. https://doi.org/10.1080/00220670009598732

Mergendoller, J. R., Maxwell, N. L., & Bellisimo, Y. (2006). The effectiveness of problem-based instruction: A comparative study of instructional methods and student characteristics. *Interdisciplinary Journal of Problem-Based Learning*, *1*(2), 49-69. https://doi.org/10.3200/JECE.36.4.315-331

Meuwissen, K.W. (2013). Readin', writin', ready for testin'? Adaptive assessment in elective and standardized-tested social studies course contexts. *Theory and Research in Social Education*, *41*(3), 285–315. https://doi.org/10.1080/00933104.2013.812049

Miller, S. L. and VanFossen, P. J. (2008). Recent research on the teaching and learning of pre-collegiate economics education. In, L. Levstik and C. Tyson (Eds.), *Handbook of Research in Social Studies Education*, (pp. 284-306). New York, NY: Routledge.

Millman, J., Bishop, C.H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*(3), 707-726.

National Council for the Social Studies (NCSS). (2013). The college, career, and civic life (C3) framework for social studies state standards: Guidance for enhancing the rigor of K-12 civics, economics, geography, and history. Silver Spring, MD

National History Day. (2011). *National history day works: National program evaluation.* Retrieved from https://www.nhd.org/why-nhd-works

Nuthall, G. (1996). Commentary: Of learning and language and understanding the complexity of the classroom. *Educational Psychologist*, *31*(3-4), 207-214. https://doi.org/10.1080/00461520.1996.9653267

Nuthall, G. (1999). The way students learn: Acquiring knowledge from an integrated science and social studies unit. *The Elementary School Journal, 99*(4), 303-341. https://www.jstor.org/stable/1002174

Nuthall, G. (2000). The anatomy of memory in the classroom: Understanding how students acquire memory processes from classroom activities in science and social studies units. *American Educational Research Journal*, *37*(1), 247-304. https://doi.org/10.3102/00028312037001247

Nuthall, G., & Alton-Lee, A. (1993). Predicting learning from student experience of teaching: A theory of student knowledge construction in classrooms. *American Educational Research Journal*, *30*(4), 799-840. https://doi.org/10.2307/1163205

Nuthall, G., & Alton-Lee, A. (1995). Assessing classroom learning: How students use their knowledge and experience to answer classroom achievement test questions in science and social studies. *American Educational Research Journal*, *32*(1), 185-223. https://doi.org/10.3102/00028312032001185

Paek, P. L., Ponte, E., Sigel, I., Braun, H., & Powers, D. E. (2005). *A portrait of Advanced Placement teachers' practices.* New York, NY: College Entrance Exam Board.

Parker, W. C., Lo, J., Yeo, A. J., Valencia, S. W., Nguyen, D., Abbott, R. D., Nolen, S.B., Bransford, J.D., & Vye, N. J. (2013). Beyond breadth-speed-test: Toward deeper knowing and engagement in an Advanced Placement course. *American Educational Research Journal*, *50*(6), 1424-1459. https://doi.org/10.3102/0002831213504237

Parker, W., Mosborg, S., Bransford, J., Vye, N., Wilkerson, J., & Abbott, R. (2011). Rethinking advanced high school coursework: Tackling the depth/breadth tension in the AP US government and politics course. *Journal of Curriculum Studies*, *43*(4), 533-559. https://doi.org/10.1080/00220272.2011.584561

Pedulla, J., Abrams, L., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state mandated testing programs on teaching and learning: Findings from a national survey of teachers.* National Board on Educational Testing and Public Policy.

Popham, W.J. (2003). *Test better, teach better: The instructional role of assessment.* ASCD.

Reich, G. A. (2009). Testing historical knowledge: Standards, multiple-choice questions and student reasoning. *Theory & Research in Social Education*, *37*(3), 325-360. https://doi.org/10.1080/00933104.2009.10473401

Reich, G. A. (2013). Imperfect models, imperfect conclusions: An exploratory study of multiple-choice tests and historical knowledge. *The Journal of Social Studies Research, 37*(1), 3-16. https://doi.org/10.1016/j.jssr.2012.12.004

Santiago, M. (2017). Erasing Differences for the Sake of Inclusion: How Mexican/Mexican American Students Construct Historical Narratives. *Theory and Research in Social Education*, *45*(1), 43–74. https://doi.org/10.1080/00933104.2016.1211971

Saye, J. (2017). Disciplined inquiry in social studies classrooms. In M. M. Manfra & C. M. Bolick (Eds.), *The Wiley handbook of social studies research* (pp. 336–359). Chichester, UK: John Wiley & Sons

Segall, A. (2003). Teachers' perceptions on state mandated standardized testing: The Michigan educational assessment program (MEAP) as a case study of consequences. *Theory and Research in Social Education, 31*(3), 287-325. https://doi.org/10.1080/00933104.2003.10473227

Sewall, G. (1988). American history textbooks: Where do we go from here? *Phi Delta Kappan, 69*, 552-558.

Shemilt, D. (2018). Assessment of learning in history education: Past, present, and possible futures. In S.A. Metzger & L.M. Harris (Eds.), *The Wiley international handbook of history teaching and learning* (1st ed., pp. 449-471). John Wiley & Sons, Inc.

Shepard, L.A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 1066-1101). American Educational Research Association.

Smith, M. (2017). Cognitive validity: Can multiple-choice items tap historical thinking processes? *American Educational Research Journal. 54*(6), 1256-1287. https://doi.org/10.3102/0002831217717949

Smith, M., Breakstone, J. & Wineburg, S. (2019). History assessments of thinking: A validity study. *Cognition and Instruction*. *37*(1), 118-144. https://doi.org/10.1080/07370008.2018.1499646

Soper, J. C. & Walstad, W. B (1988). What is high school economics? Posttest knowledge, attitudes, and course content. *Journal of Economic Education, 19*(1), 37-51. https://www.jstor.org/stable/1182018

Thornton, S.J. (1989, March 27-31). *Aspiration and practice. Teacher as curricular-instructional gatekeeper in social studies* [Paper Presentation]. AERA 1989 Conference, San Francisco, CA. https://files.eric.ed.gov/fulltext/ED315347.pdf

Thornton, S.J. (2006). What matters most for gatekeeping? A response to VanSledright. *Theory and Research in Social Education, 34*(4), 416-418. https://doi.org/10.1080/00933104.2006.10473316

Torrez, C.A., & Claunch-Lebsack, E.A. (2013). Research on assessment in the social studies classroom. In J.H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 461-472). SAGE.

VanFossen, P.J. (2005). "Reading and math take so much time…: An overview of social studies instruction in elementary classrooms in Indiana. *Theory and Research in Social Education, 33*(3), 376-403. https://doi.org/10.1080/00933104.2005.10473287

VanSickle, R. L. (1992). Learning to reason with economics. *The Journal of Economic Education*, *23*(1), 56-64. https://doi.org/10.2307/1183479

VanSledright, B.A., (2013). Can assessment improve learning? Thoughts on the C3 Framework. *Social Education, 77*(6), 334-338.

Volger, K. (2006). The impact of high school graduation examination on Mississippi social studies teachers' instructional practices. In S.G. Grant (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 273-302). Information Age Publishing.

Walstad, W. B., & Rebeck, K. (2001). Assessing the economic understanding of U.S. high school students. *American Economic Review, 91*(2), 452-457. https://doi.org/10.1257/aer.91.2.452

Walstad, W. B., Rebeck, K., & Butters, R. B. (2013). The test of economic literacy: Development and results. *The Journal of Economic Education, 44*(3), 298-309. https://doi.org/10.1080/00220485.2013.795462

Wentworth, D. R. (1987). Economic reasoning: turning myth into reality. *Theory Into Practice*, *26*, 170–175. https://doi.org/10.1080/00405848709543270

Worthington, T. A. (2018). Letting students control their own learning: Using games, role-plays, and simulations in middle school U.S. history classrooms. *Social Studies*, *109*(2), 136–150. https://doi.org/10.1080/00377996.2018.1460791

Yin, R. K. (2018). *Case study research and applications, 6th Edition.* Thousand Oaks, CA: Sage.

Top Down, Bottom Up…and Then What Happened? Assessment Policy Change in Middle School History Classrooms

Michael Gurlea[1], Tyler Woodward[2], Stephanie van Hover[1]

[1]University of Virginia

[2]St. John Fisher College

Researchers of the secondary social studies classroom have called for a move away from high-stakes multiple-choice tests and their emphasis on fact-recall in favor of summative assessments that focus on inquiry and disciplinary skills (Ercikan & Seixas, 2015; VanSledright, 2014). Curriculum and assessment initiatives such as Stanford's *Beyond the Bubble* project (Breakstone, et al., 2013) and the *College, Career, and Civic Life (C3) Framework* (National Council for the Social Studies [NCSS], 2013) echo these calls by creating frameworks and materials designed to push educators to use primary sources, inquiry, and disciplinary writing to help understand their students' learning. Large-scale assessment policy has begun, slowly, to reflect this vision of high-quality classroom-based assessment. Across the curriculum, in the wake of *No Child Left Behind* (NCLB), assessment contexts across the country are changing. The *Every Student Succeeds Act* (ESSA) requires that states "implement assessments that measure 'higher-order thinking skills and understanding'" and "allows for the use of 'portfolios, projects or extended performance tasks'" (Darling-Hammond, et al., 2016). In social studies specifically, some states, divisions, and schools have begun to reduce the number of high-stakes tests they administer while concurrently implementing alternative models of large-scale assessment (Grant, 2017).

However, little is known about how these new assessment policies are interpreted and implemented in practice. Virginia, as one of the first states to reduce the number of mandatory end-of-course multiple-choice tests and require alternative assessments, represents a key case that can provide insight into this shifting context. Research on policy realization in Virginia has the potential to offer guidance to current and future efforts to shift social studies assessment away from an overemphasis on multiple-choice and toward approaches that measure disciplinary thinking skills. In addition, studying this policy shift can provide insight into a process that

reflects what Linda Darling-Hammond (1994) described as "top-down support for bottom-up reform" where teachers and division-leaders have more power and autonomy in creating assessments with the support of the state.

This manuscript investigates the implementation of Virginia's efforts to replace two state mandated end-of-course multiple-choice assessments in secondary social studies with locally developed alternative assessments. By focusing on the process of policy realization this study focuses not on the fidelity of implementation (i.e., the extent to which the policy is enacted as intended), but rather the adaptations to the implementation (i.e., the way actors in a system interpret a policy to suit their own contexts and needs; Century & Cassata, 2016). This is important because assessment policy at the state, division, and school level acts as a lever for teacher's instruction (Au 2007; 2009). Therefore, the ways that state actors and division-leaders interpret and adapt a policy to implement it within their schools is connected to teachers' decision-making and students' classroom experiences.

Virginia has long viewed itself as on the cutting edge of assessment policy as one of the first states to enact a standards-based accountability system in 1995, well before the passage of NCLB (van Hover et al., 2010). The *Virginia Standards of Learning for History and the Social Sciences* (SOLs) outlined what students needed to learn and measured student progress toward these goals with state required end-of course testing in social studies in grades 3, 5 and 7-11. Test pass rates in social studies were a component of school accountability measures, thus creating a system for rating school and division performance and outlining consequences for schools that did not perform adequately (Duke & Reck, 2003). In 2014, the Virginia General Assembly took a step away from this high-stakes accountability system and toward assessment policy innovation by replacing five end-of-course standardized tests with "Local Alternative

Assessments," to be developed at the division-level. Two of these tests were removed in social studies courses usually offered in middle school: 'United States History to 1865' (USI) and 'United States History from 1865' (USII). The Virginia Department of Education (VDOE) gave freedom to divisions to design these "Local Alternative Assessments" as they saw fit. The initial guidelines document, however, made it clear that the new legislation "provides for the use of authentic performance assessments and portfolios" (VDOE, 2014, p. 3).

Therefore, in this study, we explore Virginia's enactment of a new assessment policy in secondary social studies. We first reviewed policy documents and news articles in order to understand and trace the history and development of this policy shift. Then, to capture how the policy was implemented across divisions in Virginia, we administered surveys to division-level social studies coordinators and conducted follow-up interviews with selected coordinators. Through this documentary analysis, survey data, and interview data we present a key case of policy realization and answer the following research questions:

1. To what extent is the adopted state policy of local alternative assessments being implemented across Virginia school divisions?
2. How do different division level social studies coordinators interpret and adapt this assessment policy?

This manuscript is organized as follows. We begin with a review of the literature to situate Virginia's assessment policy within the larger narrative of the impact of assessment policy on the classroom. We then, in order to understand the intentions of Virginia's policy, present the argument from the literature for what form effective assessment in secondary social studies should take. Next, we present the conceptual framework, followed by the methods section, which includes contextual information about Virginia and the history of implementation from our

analysis of policy documents and news articles. Following methods, the findings are presented in three parts, organized by theme: 1) Uneven Implementation, 2) Interpretations of the Policy Change, and 3) Assessment as a Lever for Instruction. We conclude with our limitations, discussion, and implications.

Literature Review

Policy and The Social Studies Classroom

Virginia's current efforts to promote performance-based "local alternative assessments" in social studies classrooms is best understood in the broader context of the NCLB-era. NCLB legislated test-based school accountability throughout the United States by asking states, divisions, and schools to administer standards-based assessments. A number of studies emerged in the wake of NCLB that indicated while there were some positive effects, test-based accountability had mixed consequences for student achievement and school performance. Using a quasi-experimental design to compare states that had implemented some version of test-based accountability prior to NCLB to states that did not, Dee and Jacob (2010) found that these testing and accountability reforms had moderate positive effects (for a relatively low cost) on student performance in fourth and eighth grade math, as assessed by the National Assessment of Educational Progress (NAEP). In addition, these gains were particularly large for students of historically disadvantaged groups and for students who were eligible for subsidized lunch.

Despite evidence of some gains, however, many researchers and educational leaders question these results in light of the potential unintended consequences of test-based accountability. For example, teachers and schools may respond strategically to standardized testing by changing the pool of tested students through election to special education or language learner programs (Cullen & Reback, 2006; Figlio & Getzler, 2002; Jacob, 2005), concentrating

largely on tested content (i.e., teaching to the test; Koretz & Barron, 1998), focusing on test-specific skills (i.e., teaching the test; Klein, et al. 2000), or fixating on specific students on the "bubble" (Booher-Jennings, 2005). These perceptions led to broad-based support across the country for a reevaluation of assessment policy and sparked movements of parents and stakeholders to voice their discontent through test opt-outs and protests (e.g., Layton, 2013).

Most research on assessment in social studies specifically, given the context of NCLB, also explores the impact of high-stakes end-of-course multiple-choice tests (Grant & Salinas, 2008; Shemilt, 2018). Research indicates that the presence of a high-stakes test at the end of a social studies course does not have a clear positive relationship with improved student learning or improved teaching. (Journell, 2010; Grant & Salinas, 2008; van Hover, 2006; van Hover & Heinecke, 2005). We do have insights, however, into how classroom-based assessments in social studies were impacted by the accountability era. In their review of the literature on classroom-based assessment in social studies Torrez & Claunch-Lebsack (2013) describe this impact:

> "American social studies classrooms, for the early part of the 21st century, have been in a near state of atrophy as the disconnect between classroom realities and best social studies practices widens. Diminished social studies instruction has resulted in the silencing of constructive assessments of student learning which may have constrained new classroom research assessment projects" (Torrez & Claunch-Lebsack, 2013, p. 465)

The pressure of external accountability led social studies teachers and programs to turn to the use of externally developed social studies curricula (Segall, 2003) which have been shown previously to stymie student depth of understanding (Brophy, 1990; Brophy et al., 1991; Brozo & Tomlinson, 1986; Guzzetti et al., 1992; Sewall, 1988). These classroom-based assessments often were not designed to help the teacher understand the learning and progress of their students, but rather to provide practice for high-stakes tests (Abrams et al., 2003; Pedulla, et al., 2003; Volger, 2006). In addition, when social studies is not explicitly tested it is often

marginalized in the curriculum, especially in elementary schools (VanFossen, 2005; Fitchett & Heafner, 2010; Fitchett et al., 2014). The mixed results of the accountability era have led to more recent calls for policy to endorse performance assessments (Gareis, 2019) and assessments created locally, where teachers have a greater voice (Reich, 2018).

In recent years there has been important descriptive work outlining the various approaches states are taking to reduce the number of traditional multiple-choice tests required. Stosich and colleagues (2018) explored the ways that 12 states, all of which were members of the Innovation Lab Network (ILN), implemented performance assessments in some capacity through semi-structured interviews with state education agency personnel. The authors found four distinct approaches that states took to the implementation and promotion of performance assessments in schools within each state: 1) Supporting teachers in implementing classroom-based performance assessments, 2) Using performance assessments as a component of graduation requirements, 3) Using performance assessments for the purpose of statewide school accountability, specifically in replacing traditional multiple-choice tests, and 4) Using performance assessments to seek a federal waiver to alter testing requirements for federal accountability. Virginia's policy efforts fall in Stosich and colleague's (2018) third category: developing and implementing local alternative assessments within a state-level accountability system. Grant (2017), as well, describes the innovations developing at the state-level, specifically regarding social studies classrooms. In his explanation of the current landscape of these new assessment practices Grant (2017) calls "for empirical study with pragmatic as well as pedagogical benefits…to continue existing lines of study into how teacher and students (and the public) navigate the extant assessment landscape" (p.472) To that point, there is very little research that explores policy that seeks to bridge the gap between best-practice and reality in the

social studies classroom.

Effective Assessment in Social Studies

The policy of "local alternative assessments" in Virginia intends to avoid the pitfalls outlined above and allow teachers to employ a vision of effective assessment. What can this best practice assessment look like? Performance assessments have long been promoted by policymakers as tools for educational reform (Linn, 1993) as they can 1) help teachers capture what is not measurable by other assessment formats (Resnick & Resnick, 1992); 2) be a learning tool in and of themselves rather than just an indicator of achievement (Bennet, 2010; Bennett & Gitomer, 2009); and 3) improve instruction by exposing teachers to new visions of what is important for students to learn (Lane, 2010; 2013). In social studies, a vision of best-practice assessment aligns with this vision of performance assessment through allowing students to exhibit mastery of disciplinary knowledge and skills through inquiry and projects (Ercikan & Seixas, 2015).

A handful of studies have been conducted in secondary social studies classrooms that explore student thinking and envision a better assessment practice. Reich (2009; 2013) conducted think-aloud interviews with 13 10th graders using selected test items from the New York State Global History and Geography Regents exam in order to explore student thinking and learning. Findings indicated that social studies multiple-choice test questions often do not measure disciplinary thinking and content knowledge, but rather measure a combination of history content knowledge, literacy, and test-taking skills. Smith and colleagues (2019), similarly, used think-aloud interviews with 26 high school students who had completed the Advanced Placement (AP) US History course and scored a 3 or higher on the end-of-course exam (showing some level of proficiency in course content). In these interviews they had

students consider both traditional forced-choice test items and their answers to open-ended performance-based History Assessments of Thinking (HATs). Corroborating and expanding on Reich's (2009; 2013) findings Smith and colleagues (2019) found that while both question types pushed students to engage with historical thinking, multiple-choice items often led to students engaging with "construct irrelevant" reasoning like eliminating distractors and considering the best-fit answer to a given multiple choice question. These studies of student thinking on traditional multiple-choice tests provide us with important insights into how assessments can and should be aligned with the goals of social studies education. Policy that supports performance assessment has the potential to promote the kinds of assessments that focus on disciplinary thinking.

Importantly, there have been key examples in social studies assessment research where a performance-assessment has been used and shown to be effective at helping students learn and retain this knowledge. Parker and colleagues (2011; 2013) conducted design experiments in AP US Government & Politics courses across two secondary schools comparing a course taught in a more traditional, didactic way with a course that was taught using a project-based curriculum and sustained inquiry. A rigorous project-based curriculum, like the one enacted by Parker and colleagues, must include the following characteristics: The project must 1) carry the full subject matter load of the course, 2) be authentic, related to the world beyond the classroom, 3) focus on a meaningful learning goal and, 4) have an appropriated external summative assessment (Parker & Lo, 2016). The AP test remained as a final measure of student performance at the end of each course. Students were still able to perform well on this traditional end-of-course assessment while engaging in more ambitious instruction throughout the year. There was, however, tension, surrounding students' adaptation to a new kind of course format and new definitions of success

and knowledge. Similarly, National History Day (2011) conducted a longitudinal study across multiple states from 2008 to 2010 in which student achievement was measured after engaging in student-driven historical research. While both the Parker (2011; 2013) and National History Day (2011) studies were not explicitly centered on assessment, both studies are effective in presenting the complex and interconnected relationship between curriculum, instruction, and assessment. The National History Day (2011) results, in particular, suggest that when presented with more ambitious instruction students were likely to out-perform their peers on traditional assessments.

In addition, there is evidence that when teachers are given the power to navigate these complex assessment contexts in ways that align with their beliefs and expertise, they are more likely to acknowledge the constructed nature of disciplinary knowledge in social studies. Studies that support this idea call for teachers to have more autonomy in their classroom-based assessment decision-making. Meuwissen (2013) presents a case study of one novice and one experienced teacher and their assessment practices across two course contexts: a traditional curriculum culminating in a standardized test and a flexible elective curriculum in which teachers had more autonomy. While both teachers had sound and consistent beliefs that the purpose of assessment was a mechanism to give students feedback about their learning, this philosophy manifested differently in their elective and tested course. These teachers' discussions of assessment with their students in their electives focused on feedback loops and the constructed nature of social studies knowledge. On the other hand, discussions of assessment in the traditional government course context was more didactic, though both teachers did push back against traditional means of assessment in these courses with adaptive assessments. The amount that each teacher advocated for alternative assessments and publicized their deviation from traditional assessment was mediated by their years' experience and political capital in the school.

Meuwissen's (2013) findings indicate that despite the ways that classroom-based assessment has been co-opted and influenced by its relationship with achievement, evaluation, promotion, and retention when teachers are given more autonomy to design classroom-based assessments they have the potential to create assessments that are more aligned with the goals of the field of social studies education. That is, the use of assessments that focus on student learning and disciplinary knowledge and skills. These findings are echoed in Reich's (2018) argument that current assessment contexts feature an overemphasis on the reliability of tests rather than validity.

Implementation Research and Policy Realization

Our study, while situated in the literature on assessment in social studies is also situated in the literature on implementation and policy realization research (Ball, 1997; Century & Cassata, 2016). Implementation research is anchored in exploring what an innovation or new policy can and should be, what happens throughout and following innovation or new policy enactment, and what we can learn through these explorations about enhancing education (Century & Cassata, 2016). For this study, we specifically utilized a pro-adaptive perspective to implementation research, which places the focus on school division leaders' ability to adapt, rather than to strictly comply, to new policy initiatives (Century & Cassata, 2016). This perspective emphasizes that perfect implementation is never feasible (Durlak, 2010; Moore et al., 2013) and that adaptation is the typical propensity of those who implement new policy across varied school contexts (Berman, 1981; Dearing, 2009; Hall & Hord, 2015). This perspective also considers the role of policy developers and the ways they may alter policy elements across time to ensure implementation success—an approach better known as "mutual adaptation" (Dearing, 2009; Dusenbury et al, 2003). This approach considers that policy should be adaptable for a

variety of school division contexts and all actors involved should be willing to adapt the policy to improve the chances of implementation.

We also draw on Ball's (1997) conception of policy realization, which seeks to make sense of the complex and non-linear nature of policy implementation. Ball (2017) argues that policies are not neutral, nor are they static, and that the path of a policy from its inception to dissemination to enactment is an interpretative, localized and context-based process that is "ongoing, interactional, and unstable" (p. 10). Policy realization recognizes the "socially and politically constructed nature" of schooling (Ball & Goodson, 1984 p. 3). Ball (1997) makes clear that within institutions, "policies do not normally tell you what to do; they create circumstances in which the range of options available in deciding what to do is narrowed or changed or particular goals or outcomes are set" (p. 270.) Considering both a pro-adaptive perspective to implementation research and a conception of policy realization research allowed us to investigate the interrelationships that exist when a policy is implemented and how those who work within educational institutions respond and take action. These bodies of literature informed our construction of a conceptual framework to make sense of the role of the division-level social studies coordinator in the process of policy realization.

Conceptual Framework

We ground our study in a framework adapted from Spillane (1998) in his study of division-level implementation of a state reading policy. Spillane's understanding of the role of division leaders in a shifting policy context helped us conceptualize the role of division-level social studies coordinators in Virginia. Further, the concepts of fragmented centralization (Meyer & Scott, 1983) and segmentalism (Kanter, 1983) helped situate the division-level social studies coordinator's role within the division and the state. These division leaders, in many ways, act as

local policymakers as they engage in a process of "constructing" policy rather than simply accepting or rejecting it (Spillane, 1989, p. 36). This process of construction involves an interpretation and adaptation of the policy to fit a specific division's context and can, in some cases, also lead a division leader to work to adapt their context to better suit the policy.

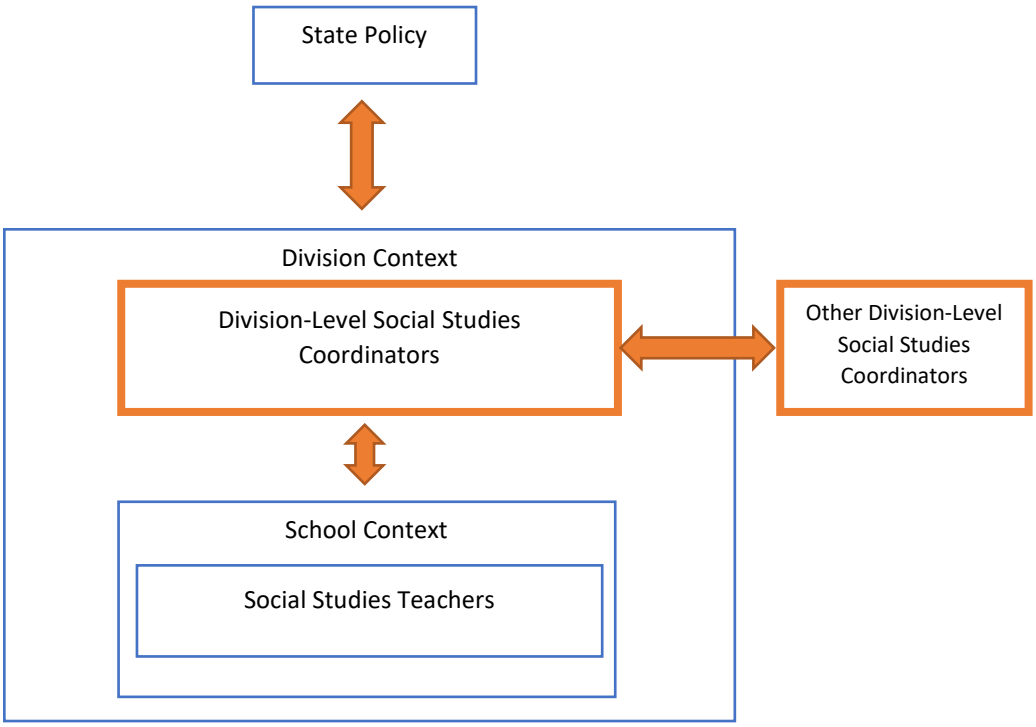
Fragmented centralization (Meyer & Scott, 1983) suggests that a very particular power is centralized in the office of a division-level social studies coordinator that is distinct from the power of any other local actor. In the case of Virginia, division-level social studies coordinators do not have any evaluative power over teachers, but interpret state policy to set an agenda each year concerning social studies curriculum, instruction, and assessment. This power is importantly distinct from the evaluative power of the school principal or the more localized influence of a social studies department chair. Further, a division-level social studies coordinator might levy their influence to help facilitate and pay for professional development for social studies teachers or invite them to state-level trainings and conferences. The nature of segmentalism (Kanter, 1983) indicates that due to the fragmented nature of power in the structures of school different subsections of an organization will respond to policy change differently, influenced by their own agendas and contexts. For instance, social studies division-level coordinators from two different divisions in the state will likely construct a policy imperative differently. Similarly, even within a division, a social studies coordinator and a science coordinator may construct a policy differently and will be unlikely to communicate about their interpretation and adaptation.

A given division-level coordinator must navigate the pressures of context from the state, their division, and the school communities within their division (See Figure 1). These contextual factors include, but are not limited to, 1) division and school leadership (e.g., the superintendent's office, school principals), 2) community characteristics, 3) funding, 4)

relationships with the state department of education, and 5) relationships with local colleges and universities. All of these factors, in combination with a division leaders' own philosophy of education, can influence the way that they choose to construct a state policy.

Figure 1

Division-Level Social Studies Coordinators' Role



This conceptual frame not only impacted the questions we asked, but also the interpretative lenses and coding schemes we used in analyzing our findings. Aligned with pro-adaptive approach to implementation research (Century & Cassata, 2016) our team designed research questions that focused closely not on the fidelity of policy implementation, but rather the ways that actors across the state were forced to interpret and adapt to a new policy based on their unique contexts. Spillane's (1998) framework helped us developed codes in our analysis of survey and interview data that centered division leaders as policy constructors.

Methods

We used a qualitative research design method (Miles et al., 2014) to explore the following research questions:

1. To what extent is the adopted state policy of local alternative assessments being implemented across Virginia school divisions?
2. How do different division level social studies coordinators interpret and adapt this assessment policy?

These questions help us explore the consequences, intended and unintended, of an assessment policy that aims to fill the vacuum left by a reduction in high-stakes multiple-choice tests. In addition, understanding how division leaders interpret this policy can help make sense of other reforms that center localities in the creative process are implemented and adapted.

Data Sources, Participants and Context

The data sources for this study include policy documents, news articles, surveys, and semi-structured interviews. Data collection included three phases: 1) document analysis to construct a framework for the inception, implementation, and dissemination of the policy; 2) distribution of a survey to all state-level and division-level social studies administrators and college educators; 3) semi-structured follow-up interviews with select survey respondents. We collected policy documents and news articles to first trace the inception of this shift in assessment policy. Then, we made the decision to create and administer the survey to social studies coordinators, as these were the individuals that were most likely leading the effort to implement the new policies within their school divisions. Lastly, we identified specific coordinators to interview to obtain more in-depth views from coordinators about what was happening in their respective school divisions and to better triangulate our data sources.

In total, 30 coordinators completed our survey and four were selected for post-survey semi-structured interviews. The coordinators were selected for interviews based on various criteria. First, the coordinators had to be a current or former school division leader for social studies curriculum and instruction. Often these coordinators were just responsible for social studies, but in smaller divisions they sometimes oversaw multiple content areas. Second, we selected coordinators that represented a range of differing school divisions regarding size, demographics, and geographic location (Table 1). Lastly, we were interested in interviewing coordinators that had varying levels of experience.

We identified the coordinators by the pseudonyms Dolores, Haley, John, and Debbie. Dolores, a White female, was the former social studies coordinator for a division of approximately 5000 students for 10 years. Haley, a White female, was the current social studies coordinator for her division of 30,000 students. John, a White male, was the current social studies coordinator for a large (~50,000 students), school division for the last eight years, and Debbie, a White female, was the current social studies coordinator for a large (~60,000 students), school division for the past 15 years. Each coordinator interviewed had a great deal of experience working with the various assessment policy initiatives put forth by the General Assembly and VDOE. While these participants were by no means representative of every social studies division-level coordinator from across the state they came from key districts that had worked with experienced partners (often from local universities) and made an effort to implement the policy change.

Table 1

Social Studies Coordinator Participant Demographic Information

Participant Name	Gender	Race	Division Size	Years as Coordinator
Dolores	Female	White	~5000 students	10 years
Haley	Female	White	~30,000 students	1 year
John	Male	White	~50,000 students	8 years
Debbie	Female	White	~60,000 students	15 years

***Policy Documents and News Source*s**

The document analysis relied on systematic collection of local, state, and federal policy and news documents and presentation materials from conferences and workshops. Policy documents and conference materials were accessed through the VDOE website, including PowerPoints that were presented by various state-level leaders at VDOE conferences, official memos written by Virginia's Superintendent of Public Instruction, and other VDOE-approved documents that outline important topics related to school divisions' adoption of performance and balanced assessment. News articles were acquired through searches in national, regional, and local news organizations' digital archives. The consulted papers included *The Washington Post*, *The News Virginian*, *The Richmond-Times Dispatch*, *The Daily Progress*, *Orange County Review*, *Roanoke Times*, and the *Greene County Record*. Search terms used in all searches included, 'Education,' 'Assessment,' 'Local Alternative Assessment,' 'Testing,' 'SOL Test'/'Standards of Learning', 'Performance'/'Performance-Based,' and some iterations of these terms. All document titles and headlines resulting from these searches were considered for inclusion in document analysis.

Surveys

Surveys were administered to social studies coordinators using a protected online survey administration system (i.e., Qualtrics). The survey consisted of 26 questions (see Appendix A) that varied in format (e.g., short answer response, Likert items, etc.). The survey asked respondents open-ended questions regarding their division's implementation of "local alternative assessments" in addition to Likert items regarding fidelity, teacher perception, assessment quality and perceptions of policy execution. Our research team developed survey items aligned with our conceptual framework considering how division leaders interpreted and adapted to the new state policy. Our survey was piloted with four social studies education researchers in the state of Virginia who were familiar with the state policy. While the quantitative trends among the Likert items are included (see Appendix C) due to the low number of survey respondents these were not considered in our final data analysis and interpretation.

Before the administration of the survey, we generated a list of 83 social studies coordinators from the VDOE and division websites. We systematically e-mailed each coordinator every two to three weeks for a total of four reminders across three months. If the coordinator completed our survey, they were no longer sent reminder e-mails. In total, 30 out of the 83 coordinators completed the survey, which resulted in a 36% completion rate. No clear patterns emerged in analyzing the division coordinators who chose to participate and those who did not. While the sample of 30 division coordinators was by no means representative, it did feature leaders from divisions of various sizes, socio-economic status, and region across the state of Virginia. Each survey respondent was given a unique numerical identifier in order to simplify data analysis. While the demographic data (e.g., years of experience, division of employment) for each of these respondents was considered in our data analysis, it cannot be shared in the full manuscript. Often, at the division-level, there is only one person who takes on the role of social

studies coordinator. Therefore, we are unable to reveal even the region of origin for our survey respondents to protect their anonymity.

Post-Survey Semi-Structured Interviews

We conducted post-survey semi-structured interviews with four participants. The participants had completed the survey and were contacted based on their survey responses to ask if they could partake in the interviews. We attempted to interview individuals who completed the survey and represented a characteristically distinct school division or who were involved in the assessment policy shift from the start. All four participants agreed to an interview and completed a digital consent form. We also audio-recorded and transcribed each interview utilizing the Zoom transcription tool. The interview protocol focused on clarification and extension of survey responses, the division leaders' experience of the policy change, as well as in-depth discussion of division-level history assessments (please see the interview protocol in Appendix B). Each interview lasted for an hour each for a total of 4 hours of interview data.

Data Analysis

Document analysis was comprised of the systematic collection of local, state, and federal policy and news documents and presentation materials from conferences and workshops. Starting with documents published in 2014, the first and second authors determined emergent themes through holistic coding (Miles et al., 2014) and plotted the dates on a timeline to trace the trajectory of the policy over time. Our plotting of each document on a timeline made it possible to triangulate this information with the survey responses and post-survey interviews with social studies coordinators. All researchers met to review the emergent themes from the document analysis that were written in narrative form by the first author.

Analysis of the short answer questions of the surveys consisted of three rounds of iterative coding. First, the short answer survey data was read separately and holistically for initial codes by the first and second authors. The first and second authors then met to go over these codes and determined which ones would be used in the second round of coding. The second round of coding consisted of the first and second authors separately reapplying the more refined codes derived from the first round. The authors then met to go over this second round. Finally, the first and second authors met to write short analytic memos for each survey question which included a summary of each response, potential disconfirming evidence, and evident conceptual connections.

The post-survey semi-structured interview data were coded in two major stages (Saldaña, 2013). The first stage included applying holistic codes to data chunks. All researchers worked together to identify these descriptive codes during the first stage for the interview data (see Table 2 below). As we worked through one interview of the four, we compiled a list of holistic codes and sub-codes. We then separately read through the other interviews with the list of holistic codes and sub-codes in mind and then met to code the other three interviews for a second major stage of coding. We also created a codebook that outlined each holistic code, applicable sub-codes, and ways we defined each code.

The document, survey, and interview data were triangulated to confirm potential findings. We determined that each data source largely corroborated each other, which is described further in our findings section. However, in the instances that we identified disconfirming data, we worked together to interrogate its relationship with our potential findings. We actively looked for rival explanations to reduce our biases and to better support our findings.

Table 2

First and Second Stage Coding Cycles for Interview Data

Interview Coding	
First Stage	Second Stage
Assessment+	Definitions of Purpose of Type of Philosophy of Rubrics and Scoring
Instruction+	Impact of Assessment on Instruction
Players+	State Legislature VDOE State Organizations Division Historic Sites University Partnerships
Professional Learning/ Professional Development+	
Teachers+	Trust Accountability
Students+	Learning
Policy Shift+	Reasons for Process of Perceptions of Stakes
	SOLs+

Note. + indicates a primary code, all other codes are sub-codes.

History and Context

Virginia led the charge on standards-based reform, creating a system in the mid-1990s that closely aligned tests in math, English, social studies, and science with standards and tied student performance on these tests with consequences for schools and divisions. When NCLB was passed in January of 2002, Virginia was already at the forefront of the standards-based

educational reform movement and only had to adapt this system slightly to meet the new federal conceptions of school success (van Hover et al., 2010). Due to the complex consequences of NCLB in recent years policymakers have considered reducing the number of standardized tests students take or changing these assessments all together.

These national debates were reflected locally in discussion of Virginia policy. In the movement away from standardized test-based accountability Virginia policymakers once again wanted to take on a role of national leadership: rather than adopting the Common Core, legislators sought to create the SOL 2.0. The 2013 gubernatorial race was characterized by both candidates, Terry McAuliffe (D) and Ken Cuccinelli (R), proposing a reduction to the number of SOL tests so students and teachers could focus less on rote memorization (Meola, 2013). Interests across the state converged as school boards joined together to pass resolutions requesting a decrease in the number of SOL tests (at the time 34) that students were required to take (Strauss, 2013). Several Virginia educational organizations including The Virginia Association of School Superintendents (VASS) and the Virginia Education Association (VEA), the state teacher's union, expressed support for a reform of the assessment and accountability process (Meola, 2013; Reid, 2013).

Chris Braunlich (2013), a former member of the Fairfax County School Board and contributor to the conservative newspaper the Washington Examiner, in an op-ed for Charlottesville's *Daily Progress* expressed some reservations with SOL reform, "How do we meet the demands of high performers," he asked, "while simultaneously meeting the needs of students who too often lack the basic skills necessary for employment?" Many school boards answered this question in their resolutions, arguing that the traditional SOL tests were not a quality way to measure student achievement, and asked, instead, for assessments that more

accurately reflected student knowledge and understanding against rigorous standards. Teachers, these resolutions said, should be at the center of design; creating assessments that provoke inquiry and engender a lifelong love of learning (Chandler, 2014). Delegate Thomas Greason, one of the bill's sponsors outlined its purpose:

> "The vision is moving to a world where not every assessment has to be a high-stakes, end-of-year, high-pressure bubble test…we augment it with projects and portfolios and other kinds of highly organized and thoughtful assessments that give students and teachers a chance to get deeper in the content, to integrate content across different disciplines, and to create the kind of education experience we all want to have happen in our schools." (As quoted by Meola, 2014)

Virginia's proposed changes received some positive attention nationally in the broader context of the standards-based assessment debate. By not fully eliminating standardized tests, but instead rolling back the number that students were responsible for taking, and by replacing these tests with assessments that focused on "problem-solving and critical thinking" these proposals, in the eyes of some, represented a sensible change (Don't Ditch, 2014). Some critics within the state, however, questioned how effective these changes might be without proper funding. Divisions that were already advantaged in terms of resources may have a much easier time developing and implementing authentic performance assessments with fidelity than other smaller divisions. ("Perspective," 2014).

After the General Assembly's April 2014 adoption of this new assessment policy, the VDOE began to review, interpret, and reinterpret the act. In August and September, the first policies from the VDOE were communicated to school divisions and the state at-large through a superintendent's memo and a "Guidelines" document (Staples, 2014, August 15; VDOE, 2014). These initial directives already began to mark a departure from the language of the legislation from the General Assembly. The legislature implied that "local alternative assessments" were synonymous with "authentic performance assessments," but this initial guidelines document

indicated that this was not the case, giving schools freedom to create a variety of assessments to fulfill these new requirements. In addition, it left room for schools to develop their alternative assessments over the course of a few years. The VDOE followed through on the legislature's promise of support for professional development—offering, in October of 2014, incentive grants across the state to assist with the creation of "local alternative assessments" and the training of teachers. Through 2015, each of Virginia's eight superintendent's regions received at least $14,000 in support. (Staples, 2014, October 24; VDOE, 2015) It is worth noting that each of these regions varies greatly in terms of student demographics, school and division size, and accreditation status (See Appendix D).

A year after the initial passage of the "local alternative assessment" legislation the VDOE's surveys of school divisions were reported in April of 2015. These surveys found that across the state 90% of divisions were administering a common, local assessment, however, most (~65%) were still giving standards-based, multiple-choice exams in US History part I and part II. Schools also reported a desire to plan and implement "local alternative assessments" that more closely resembled "authentic performance assessments" and moved away from a traditional multiple-choice test. Across the eight superintendent's regions, most school divisions sought support in the development of their alternative assessments from either a private consultant or with the help of Virginia's colleges and universities. (VDOE, 2015).

In the following years the VDOE also introduced the concept of a "balanced assessment" which bridged the gap between traditional multiple-choice exams and "authentic performance assessments" and gave schools a middle-ground to work towards (VDOE, 2017). Much of the work of the VDOE remained the same into 2017: they continued to produce supporting materials for the creation and administration of authentic performance assessments and conduct "desk

reviews" of school divisions. A quality criterion tool was created and distributed along with a glossary of assessment terms to support divisions in the creation and administration of their new assessments. The superintendent's memo outlining these supports was also the first public document to explicitly state the goals of the changing assessment policy. These goals included college and career readiness and student access to a variety of instruction and assessment strategies that support relevant skills, which put the student at the center of the learning process. (Staples, 2017). These goals marked a shift in the narrative about local alternative assessments coming from the VDOE; no longer were authentic performance assessments a distant, regulative goal, but now a concrete one with a coherent rationale.

Through 2018 and 2019 the VDOE continued to host conferences and professional development opportunities to support the administration of authentic performance assessments and conduct "desk reviews" to understand what was happening across the state. In early 2019 the timeline of implementation was updated and expanded once again. This new timeline included the expectation that divisions would use VDOE common rubrics and begin cross-scoring student assessments within and across schools in the division. (VDOE, 2019). Currently, there is very little data about what is happening in schools with regard to local alternative assessments. The reports and updates provided by the VDOE did not continue beyond 2016. In 2019, the state once again reduced the number of end-of-course multiple-choice tests that divisions had to administer even as the process of replacing these tests with local alternative assessments remains murky.

Findings

Survey and interview data supported the complex, iterative story of policy change presented by policy documents and news sources. Analysis of the data led to the emergence of

three major findings: (1) Implementation of the policy across the state was uneven based on access to resources and professional development. (2) Interpretations of the policy change appeared to be mediated by relational interactions and networks of trust between the VDOE, division leaders, and teachers. (3) The intention of the policy shift, to impact teachers' instruction at the classroom level, is a slow and ongoing process.

Uneven Implementation

Policy documents, survey results, and interviews all corroborate the idea that, especially in the first years of implementation of local alternative assessments, there was variability in interpretation and execution of the state's policy. The survey of division leaders revealed the various kinds of products that were being considered to meet the requirement of administering a local alternative assessment. Table 3, shown below, displays this variability. Division leaders were able to select multiple types of assessment if they felt they applied to their implementation process. In addition, the frequency at which these local alternative assessments were administered, regardless of their format, changed based on the division. 32% of respondents indicated that they occurred at least once a month, 39% once a quarter, 25% once a semester. Most division leaders indicated that these assessments should be administered more frequently than their current standard.

Table 3

Type of Local Alternative Assessment

Type of Assessment	**Total**
Interdisciplinary Problem-based Assessment	13
Document Based Question	21
Inquiry Project	10
Research Project	16

Multiple Choice Test	16
Portfolio	9
Public Product (e.g., Podcast, YouTube Video, Letter, etc.)	14
Public Speaking	8

Each division leader described their local alternative assessments in differently; some schools indicated that they had totally replaced multiple-choice tests with project-based experiences that emphasize student skills of evaluation, analysis, and synthesis and focus on primary sources. Other divisions indicated that multiple-choice assessments were still an important part of their approach through the use of a balanced assessment system using multiple modes of assessment. The language used to describe these assessments across the divisions represents a wide variety of terms, all present in the literature on assessment, including inquiry, performance tasks, performance-based assessment, authentic tasks, or project-based assessments. The freedom granted by the VDOE in the initial interpretation of the policy contributed to this wide variety of interpretations.

Qualitative responses in surveys and interviews indicate divisions' access to resources and the allocation of resources by the VDOE also contributed to uneven implementation across the state. Some school divisions were able to respond quickly to policy changes as many were already invested in training teachers and administrators in designing and implementing in project-based learning and alternative assessments to multiple-choice tests. One division leader described their schools as having "complete control" in creating "inquiry-based and DBQ division-wide assessments" (Survey, R13). This division's concern was not with the creation and implementation of performance-based assessments, but rather with the "consistency of rubric utilization, and the instructional shift to inquiry" (Survey, R13).

Other divisions were greatly involved in the process of advocating for the initial legislation, members of a state-level consortium working to create performance tasks, or in close geographic proximity to flagship universities leading the charge. While these divisions dealt with challenges early on, they were able to adapt and make incremental changes over the first few years of the policy. Haley shared her division's experience of these early challenges,

> "In that first year of implementation, in my opinion, we had almost zero guidance from the Virginia Department of Education in what an alternative assessment would look like. We had in our head what we hoped it would look like. So, in that first year we actually developed a portfolio project that all of the sixth and seventh grade students did throughout the year hitting all of the essential skills. We then heard from other division is other people, their alternative assessment was just a multiple-choice test...I believe it was the following year, maybe two, that the state gave a little bit more guidance on what that should look like, and we started moving away from the portfolio and towards a series of performance tasks and performance assessments" (Interview)

Haley was not alone in this assessment of what was happening across the state, and survey responses corroborated this tension between work being done within divisions, but anxiety about the landscape of policy implementation across divisions. One respondent echoed this,

> "The only thing I am concerned about with the policy is the consistency of instruction across the state. Currently there is no way to measure if students are getting to the same level of skill because each division is designing their own PBAs. Though I am confident in the work we have done; I am not sure how it compares to other school divisions." (Survey, R21)

On the other hand, divisions that were not involved in the processes of change, did not have access to funds, or the expertise of experienced partners often were slow or unable to change their assessment practices. Haley's interview and the survey response above reference how these divisions were perceived from other actors across the state. Some survey respondents felt this pressure and expressed their feelings of a lack of support and guidance, "The work is tremendous and the VDOE has not provided the necessary assistance" (Survey, R25). Another division leader succinctly said, "Teachers need more than snapshot professional development"

(Survey, R18). Division leaders in these cases express the pressures they and teachers feel to create something new with a lack of support from the state.

It was the perception of some division leaders across the state, as indicated by Haley's interview and a number of survey responses that divisions that were not implementing changes to their large-scale and classroom-based assessment were unwilling to adapt to the new state policy. This perspective represents a different argument than the one presented here that uneven implementation was a product of funding, capacity, and access to resources. While it is possible that across the state there were districts that were unwilling rather than unable to adapt, we did not find evidence for this rival hypothesis in our survey and interview data. Surveys of districts-leaders that represented districts that were slow to implement the change did not reflect a criticism of the policy writ-large, but rather criticisms of implementation based in uneven distribution of resources, training, and funds.

Interpreting the Policy Change

Analysis of surveys and interview data led to the development of codes that captured the way the policy change appeared to be mediated by relational interactions and networks of trust between the VDOE, division leaders, and teachers. Any given division's implementation was influenced by these factors and was further influenced by division leaders' philosophies of assessment, instruction, and relationships with teachers and state actors. As this policy represents "top down support for bottom up reform" (Darling-Hammond, 1994) it became clearly important how various actors in this system perceived their own ability and the ability of others to enact and interpret a policy that was not purely regulative. It is important to note that these findings can only describe a limited vision of relational trust (Bryk & Schneider, 2002). While networks of trust were at play in the implementation of this policy, our analysis of data could only surface

how these networks were perceived by division-level social studies coordinators. Trust, in this case, was uni-directional, and the data presented below only describes how division-level coordinators thought about the teachers in their district. Data surfaced that to what extent division-leaders respected teachers, perceived their competence, and considered their capacity for change had an impact on where they placed the locus of control for the construction of local alternative assessments.

Many division leaders expressed trust for teachers, not only in the process of implementing this specific policy in their classrooms, but also as experts in what works for students, indicating a high level of respect for the teachers in their division and a belief in their competence. One division leader cited the expertise of teachers as the main reason the policy shift had gone smoothly, "Implementation of local alternative assessments in USI and USII have gone very well due to the high capacity of the teachers in those courses. Members of our team have served on state-level committees and bring tremendous knowledge and resources to the table. Challenges have arisen in the areas of duration of the tasks and time that it takes to score the assessments" (Survey, R17). In our interview Dolores echoed this trust of teachers, hoping that this policy would give them back power that she believed the multiple-choice tests of the past had taken away,

> "[There is a] huge disconnect when teachers lose the ability to create the assessments and relate it back to the instruction that they provided children and it becomes a guessing game just as much for teachers as it does kids. We've lost all credibility here. So, I believe that giving kids multiple chances, letting kids have some choices, and moving away from things that can be Googled and memorizing dates and states and capitals and factoids and really giving kids a chance for authentic research and to follow their own natural curiosity and interest." (Dolores, Interview)

The process of teacher-created assessments was of particular importance to Dolores as she lauded even tasks developed early in the process of policy implementation as "project-based

learning, small projects, and writing" and "not DBQs or canned lessons" (Dolores, Interview). For Dolores, the policy worked well when teachers were creating assessments so they could tie it directly to instruction they had given.

This trust of teachers and the ability of localities to create local assessments was sometimes at odds with the direction provided by the VDOE. While Haley asked for more support in implementation, she found the conferences and professional development that the state did provide inadequate for their needs,

> "We have attended pretty much all of the conferences offered by the VDOE we because we feel it's very important to keep a pulse on the expectations of the state…We always try to bring somebody that's not from central office, just so that they're involved in the process too. They tend to say, 'well, we were doing this two years ago. This is as far as they've gotten? Gee, you know, we're way ahead on this.' …We to date do not feel that the VDOE has provided us with anything that we feel we need to take back to our division and utilize because I think we feel confident that what we've developed locally is more reliable and more in line with our vision for performance assessments" (Haley, Interview).

Haley's conception of trust and competency was in her own colleagues and the teachers she supports in her division rather than in the leadership of the state. Haley also emphasized the importance of teacher independence in deciding what represented student success in scoring and what went into their gradebooks to represent growth to students and parents,

> "They have a mandatory performance task for each unit throughout the year, but that performance task does not necessarily need to be graded. The performance task a lot of times is used as the tools to teach the analysis of the document or the development of an argument, you know, how do I write a claim, what would be appropriate evidence? So we wanted to honor the teachers' independence in selecting whether or not that was going to be scored depending on the proficiency of their students. Once a quarter the performance assessment is required to be scored. So those four assessments are required to go into the gradebooks all of the other tasks are not. Some people put them in some people don't, and then they most certainly still have their own individually developed or school developed of quizzes, multiple-choice tests" (Haley, Interview).

It is clear that teacher autonomy was important to Haley and that she worked to bridge the gap between the large-scale assessment policies of the past and the ways that a new state policy could

impact the classroom level.

Debbie, a division leader at a large suburban division, also emphasized the importance of giving teachers autonomy in the design and implementation of local alternative assessments. In describing her division's most recent vision of the policy she described,

> "I am not going to do it from the division level. But what I am going to do is give you the autonomy at your school, in your PLC, to develop based on the context of your needs. I want you to develop a performance assessment for your 6th graders at your school. And you are going to come together to do that. Now, you have children with special needs, you have gifted kids and so…yes you are going to take that performance assessment and you are going to scaffold it to fit the needs of your gifted population and also your students with disabilities…the foundation is going to be the same performance assessment and then you are going to come together and you are going to have these conversations with your PLC about student progress and their skills and how they are progressing and how can we tweak this. So I had just rolled that out in January to say – this is what you need to do. And I gave them time to do it. Not like, "This is what you have to do I need it tomorrow" (Debbie, Interview).

John echoed this, not only trusting teachers to create assessments, but to use the results of the assessments to influence future instruction, "These assessments include both content and skills we need to be need to be revising instruction…based on these results. There isn't a lot of formal follow up on that, so I can't speak to whether that's happening. Sort of one of those 'trust the professionalism to the professionals'" (John, Interview). John, however, did have reservations about how the state would proceed as they plan to replace more multiple-choice tests with local alternative assessments, "We wondered if it's going to count for accreditation, or count for graduation credit, can they trust local divisions to create their own?...We have to count social studies credit for high school graduation, and right now there is a blend of SOL testing and locally-created things. I know state lawmakers don't like localities to create their own things that count" (John, Interview). John's response supports the idea that middle school history courses were a safe space to enact this new policy with low stakes, but that the future of assessment in social studies in the state is still uncertain.

Other division leaders were concerned about how teacher perceptions of the policy would influence implementation. Would teachers see this as more unnecessary work? Would they buy-in to this new system? Some division leaders characterize this as the slow and lengthy process of changing the culture from the NCLB-era, "It is good for students if they are made well. Teachers see it as more work and will require time for acclimation" (Survey, R13). One division leader described this as a mental shift, "The biggest difficulty was getting social studies teachers to buy into using PBAs. The mind shift away from SOLs was not easy and we are still working with some teachers on best practices of doing PBAs well and improving their assessments with rigor" (Survey, R23). Another respondent said,

> "I also believe that greater direction from the state could help in terms of verifying that school divisions are actually taking this opportunity to improve social studies instruction and assessment. I have no worries that teachers are still assessing knowledge in terms of multiple-choice test - there is a time and place for this type of objective, knowledge-based assessment, but I feel that teachers are still doing too much of it at the individual classroom level" (Survey, R27).

One division leader expressed concern, not that teachers were taking time to acclimate to a new policy environment, but rather that they had to challenge what they conceived of as student success, "It is great that the VDOE has given divisions the options to do local alternative assessments. It reduces the stress on the teachers and students. The biggest challenge is to have the teachers trust the use of the local assessments as a viable option of student mastery" (Survey, R23). These conceptions of teacher 'buy-in' and having to make a mental shift indicate a different vision of trust between division-level leaders and the teachers they support.

In our interview Debbie described how the initial policy in 2014 came at a time when teachers were feeling particular pressures in her division,

> "At that same time that this was happening there was a movement in [our division] among the teachers because the teachers, at that time, were so overwhelmed already with mandates about things they had to do from the central office of the division. So, there was

> a movement across the county which was called, 'Just Let Me Teach.' They had signs on their classroom doors, they had t-shirts made, I mean, this was a very vocal movement. And, so, in the thick of that here was one more thing that went on their plates" (Debbie, Interview).

In sum, some of the variety of interpretations of the policy was mediated by division-leaders' perceptions of teachers, the VDOE, the policy, state lawmakers, and their own philosophies. Some division leaders indicated that their vision of successful 'local alternative assessments' hinged on the expertise and capacity of teachers to create, implement, score, and interpret the results of these tests. Other division leaders questioned how well the policy could be implemented when teachers were acclimated to the world of high-stakes testing. Others still questioned how teachers would have the capacity to develop and implement these tests among the number of other initiatives and expectations they faced.

Assessment as a Lever for Instruction

As the expressed purpose of the policy was to reduce the amount of time teachers spent on test preparation activities in their classrooms, our data analysis also explored how division leaders interpreted this policy and its impact on instruction. The philosophy of the policy, as made clear by this purpose, is that assessment can act as a lever for instruction. With new expectations of outcomes for students, teachers will change the way they teach in order to respond. Our analysis found that, while divisions leaders generally believed in this change and observed it in their districts, the process was often slow, and required a shift in the culture of actors across all levels of implementation.

The new policy change did not completely relieve the perceived pressures of NCLB-era multiple-choice testing; one survey respondent expressed the tension between these new tests, and necessary content coverage, "We definitely do not need to 'throw the baby out with the bathwater.' How many of these do we do? How do they fit within the curriculum within the

crush of trying to get it all done?" (Survey, R12). Another division leader questioned the state's chosen policy route of changing assessments, "It has been refreshing to rethink assessment, but we also need to reform curricula in order to truly produce instructional change" (Survey. R27).

This mixed support for the policy due to its impact on instruction was presented differently by another division leader,

> "I believe it [local alternative assessment] is good for students when properly utilized and constructed. I have mixed responses as it relates to teachers. Teachers, historically, have been judged by their students' performance on the SOL test. Many focused on test preparation and memorization to convey the content well enough to pass the test. Test scores misrepresented the quality of instruction taking place in the classroom. Super-star teachers were not always those with the highest pass rate. I think the policy should revert to the original intent of school-based assessments designed to merge content areas and local history." (Survey. R13)

This respondent seems to support school-based assessments insofar as they better represent good teaching and are good for students, but questioned the broader process of implementing local alternative assessments and their dual nature as both state-mandated and locally constructed. This division leader, in other survey responses, cited the "ever changing guidance" from the state as a major challenge (Survey, R13).

One survey respondent commented on the slow nature of this change, "Students struggle with analysis and writing. We are slowly shifting our instructional practices because of poor student performance on assessments" (Survey, R28). Other survey and interview respondents indicate that this change can take years as teachers adapt to student performance on these assessments, and then adapt their assessments, and then their instruction, "They've gotten *so* much better. It was really great to have [a university consult] look at our tasks and say, 'You guys are really on the right track. I would be willing to put these up against any of the ones we have put up as examples'…We knew it was getting better…We started moving more and more kids from the 'mostly got it' to the 'definitely got it' pile" (Dolores, Interview).

John characterized the instructional shift as a challenging but important part of the change, “This is definitely good for students and teachers. Making the instructional shift is the hardest part, but once that’s done, it will be even better for students” (John, Survey). He echoed this response in his follow-up interview, “We've taken the approach all along that these are a work in progress. You know, no assessment is ever perfect and we're trying to develop something that's good and we want teachers to buy in. And what we preach to them, and I say preach intentionally, is that you know assessments should guide instruction” (John, Interview). Division leaders were varied, in their belief in the ability of the new state policy to impact instruction, but largely believed that the assessment shift was good for students.

Limitations

This study traces the macro-level interpretation of a state level policy from the VDOE to division level social studies leaders across the state. While the study does provide key insights into policy implementation, it is also important to acknowledge the limitations of this study. First, it is important to note that any claim made about teachers and their interpretation of the policy change was made from the perspective of a division-level leader. Teachers were not consulted for the purposes of this study, but this is a rich vein for future assessment policy implementation research. Secondly, the results of our survey and semi-structured interview data are not generalizable beyond the contexts in which they reside. Our survey response rate makes it impossible to make claims about division leaders across the state, rather providing some key examples of reactions and interpretations to the policy of local alternative assessments. It is a distinct possibility that our survey and interview data did not capture key perspectives across the state, both because of a low response rate, and because of a willingness for division-leaders to share certain perspectives with our research team.

Discussion and Implications

The implementation of local alternative assessments in Virginia provides an important case and context to understand policy implementation more broadly and, specifically, in the shifting context of assessment. Exploring this case can provide insights into how other states, divisions, and schools might implement a policy that seeks to give teachers autonomy while still endorsing a particular kind of assessment. Our analysis of data surfaced important questions about equity, the nature of policy change, the nature of assessment in social studies, and directions for future research.

Not all divisions had equal access to resources that supported assessment change. State policy, in an attempt to ensure local autonomy, did not appear to take into account the vast differences between divisions across the state. Well-resourced divisions were able to respond quickly by providing their teachers with training and support to develop and score new assessments, while other divisions were unable to enact the change even partially. Policies that seek to give teachers and localities power over instructional and assessment decision-making should ensure that divisions are able to apply these changes equitably through funding, training, and other means of support.

Even in the pursuit of equity across all divisions, we acknowledge the imperfect nature of policy implementation (Durlak, 2010; Moore et al., 2013). In the implementation of any policy, it is important to understand the way that all stakeholders (e.g., lawmakers, state leaders, division leaders, and classroom teachers) interpret and adapt based on their personal philosophies and contexts. This complex process indicates the need for policymakers, even when endorsing bottom-up reform, to have a clear and shared vocabulary about the reasons for a change and the component parts of a policy. As researchers, as well, we still need to establish a shared

vocabulary about high quality assessment in the social studies classroom. In particular, this study highlights the importance of actors such as division-level social studies coordinators who, in this specific case, wielded a seemingly substantial amount of interpretative power in editing the state policy their divisions.

Further, in the pursuit of this shared vocabulary it is important to acknowledge the place that both multiple-choice and performance assessments have in social studies classrooms. Taken by their nature alone neither of these assessment types is inherently good or bad for teachers and students, they serve a particular purpose. Our exploration of the enactment of this policy across the state helped illuminate how divisions use multiple-choice style tests to discuss data, check-in with students, and build to greater understanding. We also encountered performance assessments that did not engage students in higher order thinking and simply 'checked a box' of perceived engagement. Policies should endorse assessments that promote and focus on student learning rather than indicate that a task has been completed in the classroom (Bennett, 2010; Bennett & Gitomer, 2009). In addition, more research needs to be done at all levels of the policy implementation process. While this study can provide insights to implementation in the broadest sense, more research should explore teachers' interpretations of policy and student experiences and outcomes.

This study builds on the important descriptive work that has begun to analyze how schools and classrooms are responding to federal and state policy changes after the era of high-stakes accountability (Grant, 2017; Stosich, et al., 2018). Future research should continue to work towards an understanding of how local actors adapt to policy decisions. The field of social studies education needs more empirical studies that trace policy implementation to the classroom level in order to understand how these shifting policy contexts impact teacher decision-making.

Future studies should work to connect not just assessment policy with assessment practice, but also the impact of these policies on student classroom experiences and student learning.

References

Abrams, L.M, Pedulla, J.J., & Madaus, G.F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice*, *42*(1), 18-29. https://doi.org/10.1207/s15430421tip4201_4

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

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258-267. https://doi.org/10.3102/0013189X07306523

Au, W. (2009). Social studies, social justice: W(h)ither the social studies in high-stakes testing?. *Teacher Education Quarterly*, *36*(1), 43-58. https://www.jstor.org/stable/23479200

Ball, S. J. (1997). Policy sociology and critical social research: A personal review of recent education policy and policy research. *British Educational Research Journal*, 23(3), 257–74. https://doi.org/10.1080/0141192970230302

Ball, S. J. (2017). *The education debate*. Policy Press.

Ball S. J., & Goodson, I. (1984). Introduction: Defining the curriculum; histories and ethnographies. In I. Goodson & S. J. Ball (Eds.), *Defining the curriculum: Histories and ethnographies* (pp. 1-14). London: Falmer.

Bennett, R.E. (2010). Cognitively based assessment of, for and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives*, *8* (2-3), 70-91. https://doi.org/10.1080/15366367.2010.508686

Bennett, R.E., & Gitomer, D.H. (2009). Transforming k-12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), Educational assessment in the 21st century (pp. 44-61). Springer.

Berman, P. (1981). Educational change: An implementation paradigm. In R. Lehming & M. Kane (Eds.), *Improving schools* (pp. 253–286). London: Sage.

Booher-Jennings, J. (2005). Below the bubble: "Educational Triage" and the Texas Accountability System. *American Educational Research Journal, 42*(2), 231-268. https://doi.org/10.3102/00028312042002231

Braunlich, C. (2013, October 27). SOL reform's double-edged sword. *The Daily Progress.* https://www.dailyprogress.com/opinion/columns/sol-reform-s-double--edged-sword/article_3addd620-3f0e-11e3-8d7a-0019bb30f31a.html

Breakstone, J., Smith, M., & Wineburg, S. (2013). Beyond the bubble in history/social studies assessments. *Phi Delta Kappan*, *94*(5), 53-57. https://doi.org/10.1177/003172171309400512

Brophy, J. (1990). Teaching social studies for understanding and higher-order application. *The Elementary School Journal, 90*(4), 351-417. https://doi.org/10.1086/461623

Brophy, J,, McMahon, S., & Prawatt, R. (1991) Elementary social studies series: Critique of a representative example of six experts. *Social Education, 55*(3), 155-160.

Brozo, W.G., & Tomlinson, C.M. (1986). Literature: The key to lively content courses. *The Reading Teacher, 40*(3), 288-293. https://www.jstor.org/stable/20199384

Buckles, S., Schug, M. C., & Watts, M. (2001). A national survey of state assessment practices in the social studies. *The Social Studies*, *92*, 141-146. https://doi.org/10.1080/00377990109603992

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. Russell Sage Foundation.

Century, J., & Cassata, A. (2016). Implementation research: Finding common ground on what, how, why, where, and who. *Review of Research in Education*, 40, 169-215. https://doi.org/10.3102/0091732X16665332

Chandler, M.A. (2014, January 13). Virginia lawmakers call for fewer SOL tests. *The Washington Post*. https://www.washingtonpost.com/local/education/virginia-lawmakers-call-for-fewer-sol-tests/2014/01/13/a7461654-789a-11e3-8963-b4b654bcc9b2_story.html

Cullen, J. B., & Reback, R. (2006). *Tinkering toward accolades: School gaming under a performance accountability system*. Emerald Group Publishing Limited.

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, *64*(1), 5-31. https://doi.org/10.17763/haer.64.1.j57n353226536276

Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E. L. (2016). Pathways to new accountability through the Every Student Succeeds Act. *Palo Alto, CA: Learning Policy Institute*.

Dearing, J.W. (2009). Applying diffusion of innovation theory to intervention development. *Research on Social Work Practice*, 19, 503–518. https://doi.org/10.1177/1049731509335569

Dee, T. & Jacob, B. (2010). Evaluating NCLB: Accountability has produced substantial gains in math skills but not in reading. *Education Next, 10*(3), p. 54-61.

Don’t ditch standardized tests, improve them (2014, February 14). *The Washington Post*. https://www.washingtonpost.com/opinions/dont-ditch-standardized-tests-improve-them/2014/02/14/eeb9e722-950a-11e3-84e1-27626c5ef5fb_story.html

Duke, D.L., & Reck, B.L. (2003). The evolution of educational accountability in the old Dominion. In D.L. Duke, M. Grogan, P.D. Tucker, & W.F. Heinecke (Eds.) *Educational leadership in an age of accountability: The Virginia experience* (pp. 36-38). SUNY Press.

Durlak, J. A. (2010). The importance of doing well in whatever you do: A commentary on the special section, "Implementation Research in Early Childhood Education." *Early Childhood Research Quarterly*, 25, 348–357. https://doi.org/10.1016/j.ecresq.2010.03.003

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research*, 18, 237–246. https://doi.org/10.1093/her/18.2.237

Ercikan, K., & Seixas, P. (2015). *New directions in assessing historical thinking*. Routledge.

Figlio, D. N., & Getzler, L. S. (2002). *Accountability, ability and disability: Gaming the system?*. Emerald Group Publishing Limited.

Fitchett, P.G. & Heafner, T.L. (2010). A national perspective on the effectiveness of high-stakes testing and standardization on elementary social studies marginalization. *Theory and Research in Social Education, 38*(1), 114-130. https://doi.org/10.1080/00933104.2010.10473418

Fitchett, P. G., Heafner, T. L., & Lambert, R. (2014). Assessment, Autonomy, and Elementary Social Studies Time. *Teachers College Record*, *116*(10), 10.

Gareis, C. (2019) "Teaching and assessing for deeper learning" in D.G. Wren's *Assessing Deeper Learning*. (pp. 1-19). Rownan & Littlefield.

Grant, S.G. (2017). The problem of knowing what students know: Classroom-based and large-scale assessment in social studies. In M.M. Manfra & C.M. Bolick (Eds.), *The Wiley handbook of social studies research* (1st ed., pp. 461-476). John Wiley & Sons, Inc.

Grant, S.G., & Horn, C. (2006). The state of state-level history tests. In S.G. Grant (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 9-27). Information Age Publishing.

Grant, S.G., & Salinas, C. (2008). Assessment and accountability in the social studies. In L.S. Levstik & C.A. Tyson (Eds.), *Handbook of research in social studies education* (pp. 219–238). Routledge.

Grant, S.G., Swan, K., & Lee, J. (2017). *Inquiry-based practice in social studies education*: *Understanding the inquiry design model*. Routledge.

Guzzetti, B.J., Kowalinski, B.J., & McGowan, T. (1992). Using a literature based approach to teaching social studies. *Journal of Reading, 36*(2), 114-122. https://www.jstor.org/stable/40016443

Hall, G. E., & Hord, S. (2015). Implementing change: Patterns, principles and potholes (4th ed.). New York: Pearson.

Hedges, L., & Nowell, A. (1998). Black-white test score convergence since 1965. In c. Jencks & M. Phillips (Eds.), *The black-white test score gap* (pp. 149 – 181). Washington, DC: Brookings Institute.

Horn, C. (2003). High-stakes testing and students: Stopping or perpetuating a cycle of failure. *Theory into practice*, 42(1), 30–41. https://doi.org/10.1207/s15430421tip4201_5

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5-6), 761-796. https://doi.org/10.1016/j.jpubeco.2004.08.004

Journell, W. (2010). The influence of high-stakes testing on high school teachers' willingness to incorporate current political events into the curriculum. *The High School Journal, 93*(3), 11-125. https://www.jstor.org/stable/40864929

Kanter, R. (1983). *The change masters.* Simon & Schuster.=

Klein, S. P., Hamilton, L., McCaffrey, D. F., & Stecher, B. (2000). What do test scores in Texas tell us?. *Education Policy Analysis Archives*, *8*, 49. https://doi.org/10.14507/epaa.v8n49.2000

Koretz, D. M., & Barron, S. I. (1998). *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS)*. RAND Corporation.

Kornhaber, M., Orfield, G., & Kurlaender, M. (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education.* Century Foundation Press.

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

Lane, S. (2013). *Performance assessment*. In J.H. McMillan (Ed.), SAGE handbook of research on classroom assessment (pp. 313-329). SAGE.

Layton, L. (2013, April 14). Bush, Obama focus on standardized testing leads to "opt-out" parents" movement. *The Washington Post*. https://www.washingtonpost.com/local/education/bush-obama-focus-on-standardized-testing-leads-to-opt-out-parent-movement/2013/04/14/90b15a44-9d5c-11e2-a941-a19bce7af755_story.html

Linn, R.L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*(1), 1-16. https://doi.org/10.3102/01623737015001001

Meola, O. (2013, October 22). GOP delegates seek reform of SOLs. *Richmond Times-Dispatch* https://www.richmond.com/news/gop-delegates-seek-reform-of-sols/article_fb82e905-a4b4-5205-b03f-76c8ea284e06.html

Meola, O. (2014, February 10). House advances bill on school openings, SOLs, school rating. *Richmond Times-Dispatch. https://www.richmond.com/news/virginia/government-politics/general-assembly/house-advances-bills-on-school-openings-sols-school-ratings/article_3ea27d4e-9281-11e3-8aa1-001a4bcf6878.html*

Meyer, J., & Scott W. (1983) *Organizational environments: Ritual and rationality*. Sage.

Meuwissen, K.W. (2013). Readin', writin', ready for testin'? Adaptive assessment in elective and standardized-tested social studies course contexts. *Theory and Research in Social Education*, *41*(3), 285–315. https://doi.org/10.1080/00933104.2013.812049

Miles, M.B., Huberman, A.M., & Saldaña, J.M. (2014). *Qualitative data analysis: A methods sourcebook, 3rd Edition*. Thousand Oaks, CA: Sage.

Moore, J., Bumbarger, B., & Cooper, B. (2013). Examining adaptations of evidence-based programs in natural contexts. *Journal of Primary Prevention*, 34, 147–161. https://doi.org/10.1007/s10935-013-0303-6

National Council for the Social Studies (NCSS). (2013). The college, career, and civic life (C3) framework for social studies state standards: Guidance for enhancing the rigor of K-12 civics, economics, geography, and history. Silver Spring, MD

National History Day. (2011). *National history day works: National program evaluation.* Retrieved from https://www.nhd.org/why-nhd-works

Parker, W.C. & Lo, J.C. (2016) Reinventing the high school government course: Rigor, simulations, and learning from text. *Democracy and Education, 24*(1), 1-10.

Parker, W. C., Lo, J.C., Yeo, A. J., Valencia, S. W., Nguyen, D., Abbott, R. D., Nolen, S.B., Bransford, J.D., & Vye, N. J. (2013). Beyond breadth-speed-test: Toward deeper knowing and engagement in an Advanced Placement course. *American Educational Research Journal*, *50*(6), 1424-1459. https://doi.org/10.3102/0002831213504237

Parker, W.C., Mosborg, S., Bransford, J., Vye, N., Wilkerson, J., & Abbott, R. (2011). Rethinking advanced high school coursework: Tackling the depth/breadth tension in the AP US government and politics course. *Journal of Curriculum Studies*, *43*(4), 533-559. https://doi.org/10.1080/00220272.2011.584561

Perspective: A half-hearted swing at SOL's. (2014, May 3). *Roanoke Times*. https://www.dailyprogress.com/archives/perspective-a-half-hearted-swing-at-sol-s/article_e277631a-d202-11e3-829d-0017a43b2370.html

Pedulla, J., Abrams, L., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state mandated testing programs on teaching and learning: Findings from a national survey of teachers.* National Board on Educational Testing and Public Policy.

Reich, G. A. (2009). Testing historical knowledge: Standards, multiple-choice questions and student reasoning. *Theory & Research in Social Education*, *37*(3), 325-360. https://doi.org/10.1080/00933104.2009.10473401

Reich, G. A. (2013). Imperfect models, imperfect conclusions: An exploratory study of multiple-choice tests and historical knowledge. *The Journal of Social Studies Research, 37*(1), 3-16. https://doi.org/10.1016/j.jssr.2012.12.004

Reich, G. (2018). "The center fails: Devolving assessment authority to educators" In P.G. Fitchett & K.W. Meuwissen (Eds). *Social Studies in the New Education Policy Era: Conversations on Purposes, Perspectives, and Practices.* (pp. 152-157). New York, N.Y.: Routledge.

Reid, Z. (2013, November 20). Group backs cutting some of 34 SOL tests. *Richmond Times-Dispatch.* https://www.richmond.com/news/local/education/group-backs-cutting-some-of-sol-tests/article_6a94c82c-bc76-5438-adae-ab5e97ec9b83.html

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.G. Gifford & M.C. O'Conner (Eds.), *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 37-55). Kluwer.

Richardson, A. (2014, May 5). Effect of upcoming SOL changes largely unclear. *The Daily Progress*. https://www.dailyprogress.com/news/effect-of-upcoming-sol-changes-largely-unclear/article_751bc2e8-d4be-11e3-9d4e-001a4bcf6878.html

Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). London: Sage.

Segall, A. (2003). Teachers' perceptions on state mandated standardized testing: The Michigan educational assessment program (MEAP) as a case study of consequences. *Theory and Research in Social Education, 31*(3), 287-325. https://doi.org/10.1080/00933104.2003.10473227

Sewall, G. (1988). American history textbooks: Where do we go from here? *Phi Delta Kappan, 69*, 552-558.

Shemilt, D. (2018). Assessment of learning in history education: Past, present, and possible futures. In S.A. Metzger & L.M. Harris (Eds.), *The Wiley international handbook of history teaching and learning* (1st ed., pp. 449-471). John Wiley & Sons, Inc.

Smith, M., Breakstone, J. & Wineburg, S. (2019). History assessments of thinking: A validity study. *Cognition and Instruction*. *37*(1), 118-144. https://doi.org/10.1080/07370008.2018.1499646

Spillane, J.P., (1998). State policy and the non-monolithic nature of the local school district: Organizational and professional considerations. *American Educational Research Journal*. *35*(1), 33-63. https://doi.org/10.3102/00028312035001033

Staples, S.R. (2014, August 15) *Superintendent's memo no. 216-14: Clarifications regarding 2014 legislation impacting the standards of learning testing program.* Virginia Department of Education. https://www.doe.virginia.gov/administrators/superintendents_memos/2014/216-14.shtml

Staples, S.R. (2014, October 24). *Superintendent's memo no. 292-14: Alternative assessment assistance incentive grant.* Virginia Department of Education. https://www.doe.virginia.gov/administrators/superintendents_memos/2014/292-14.shtml

Staples, S.R. (2015, May 1). *Superintendent's memo no. 108-15: Local alternative assessment plan reviews.* Virginia Department of Education. https://www.doe.virginia.gov/administrators/superintendents_memos/2015/108-15.shtml

Staples, S.R. (2017, April 28). *Superintendent's memo no 135-17: Update on the implementation of local alternative assessments.* Virginia Department of Education. https://www.doe.virginia.gov/administrators/superintendents_memos/2017/135-17.shtml

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). https://doi.org/10.14507/epaa.26.2906

Strauss, V. (2013, October 27). Virginia school boards pass anti-SOL resolutions. *The Washington Post*. https://www.washingtonpost.com/news/answer-sheet/wp/2013/10/27/virginia-schools-boards-pass-anti-sol-resolutions/

Torrez, C., & Claunch-Lebsack, E. (2013). Research on assessment in the social studies classroom. In J. McMillan (Ed.). *Handbook of research on classroom assessment*. SAGE.

van Hover, S. D. (2006). Teaching history in the old dominion: The impact of Virginia's accountability reform on seven secondary beginning history teachers. In S. G. Grant (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 195–219). Information Age Publishing.

van Hover, S. D., & Heinecke, W. (2005). The impact of accountability reform on the "'wise practice" of secondary history teachers: The Virginia experience. In E. A. Yeager & O. L. Davis Jr. (Eds.), *Wise social studies teaching in an age of high-stakes testing* (pp. 89–105). Information Age Publishing.

van Hover, S., Hicks D., & Sayeski, K. (2012). A case study of co-teaching in an inclusive secondary high-stakes world history I classroom. *Theory & Research in Social Education, 40*(3), 260-291. https://doi.org/10.1080/00933104.2012.705162

van Hover, S., Hicks, D., Stoddard, J., & Lisanti., M. (2010). From a roar to a murmur: Virginia's history and social science standards, 1995 – 2009. *Theory & 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*, 34–51.

VanFossen, P.J. (2005). "Reading and math take so much time…: An overview of social studies instruction in elementary classrooms in Indiana. *Theory and Research in Social Education, 33*(3), 376-403. https://doi.org/10.1080/00933104.2005.10473287

VanSledright, B. (2014). *Assessing historical thinking and understanding: Innovative designs or new standards.* Taylor & Francis.

Virginia Department of Education. (2014) *Guidelines for local alternative assessments for 2014-2015.* https://www.doe.virginia.gov/testing/local_assessments/guidelines_for_local_alternative_assessments-2014-2015.pdf

Virginia Department of Education (2015, April 23). *Update on local alternative assessment plans.* https://www.doe.virginia.gov/boe/meetings/2015/04_apr/agenda_items/item_f_presentation.pdf

Virginia Department of Education. (2016). *Local alternative assessment guidelines for 2016-2016 through 2018-2019.* https://www.doe.virginia.gov/testing/local_assessments/guidelines_for_local_alternative_assessments-2016-2019.pdf

Virginia Department of Education. (2016b). *Assessing for deeper learning: A transformative pathway to prepare Virginia students for the future.*

https://www.doe.virginia.gov/testing/local_assessments/professional-development/index.shtml

VDOE: Committee on School and Division Accountability (2016, March 16). *Local alternative assessments*. https://www.doe.virginia.gov/boe/committees_standing/accountability/2016/03-mar/local-alternative-assessments-presentation.pdf

Virginia Department of Education (2017). *History and social science SOL institutes.* https://www.doe.virginia.gov/instruction/history/professional_development/institutes/2017/

Virginia Department of Education (2019). *Guidelines for local alternative assessments for 2018-2019 through 2019-2020.* https://www.doe.virginia.gov/administrators/superintendents_memos/2019/

Volger, K. (2006). The impact of high school graduation examination on Mississippi social studies teachers' instructional practices. In S.G. Grant (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 273-302). Information Age Publishing.

Wei, R. C., Pecheone, R. L., & Wilczak, K. L. (2014). Performance assessment 2.0: Lessons from large-scale policy and practice. *Stanford Center for Assessment, Learning, and Equity*.

Appendix A
Social Studies Assessment Survey

Start of Block: Introduction and Demographic Questions

Q1 You are being invited to participate in a research study on **the implementation of local alternative assessments in social studies classrooms.** This study is being done by **Stephanie van Hover, Tyler Woodward, and Michael Gurlea** from the University of Virginia. The purpose of this research study is **to gain insight to the implementation and influence of local alternative assessments** and will take you approximately **20-30 minutes** to complete. Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to omit any question.

We believe there are no known risks associated with this research study; however, as with any online related activity the risk of a breach is always possible. Your answers in this study will remain confidential. We will minimize any risks by collecting data through a password protected survey tool (Qualtrics) and maintaining data on the password-protected file-hosting service UVA Box. All data will be de-identified and at the conclusion of the study the data will be destroyed. If you have any questions or concerns regarding this survey, please contact Stephanie van Hover.

Q2 Please put your full name on this line:

*(**Please note: Upon receipt of your survey, we will assign your survey a numerical code and remove your name from the survey.)*

Q3 Please write your e-mail address in the text box provided:

Q4 Please list your school division:

Q5 What is your current title or role? (e.g., division coordinator of social studies curriculum)

Q6 How many years have you served in your current position?

Q7 Roughly how many teachers do you currently supervise?

Q8 What content areas do you currently supervise? Select all that apply.

- ▢ English/Language arts (1)
- ▢ World Languages (2)
- ▢ Mathematics (3)

- ▢ Science (4)
- ▢ Social Studies (5)
- ▢ Health & Physical Education (6)
- ▢ Fine Arts (7)
- ▢ Career and Technical Education (8)
- ▢ Special Education (9)
- ▢ Other (Please specify) (10) _____

Q9 How many years did you teach in the classroom as a teacher?

Q10

What grade level(s) do/did you teach? *Select all that apply.*

- ▢ Kindergarten (1)
- ▢ Grade 1 (9)
- ▢ Grade 2 (10)
- ▢ Grade 3 (11)
- ▢ Grade 4 (12)
- ▢ Grade 5 (13)
- ▢ Grade 6 (14)
- ▢ Grade 7 (2)
- ▢ Grade 8 (3)
- ▢ Grade 9 (4)
- ▢ Grade 10 (5)
- ▢ Grade 11 (6)
- ▢ Grade 12 (7)
- ▢ Other *(please specify)*: (8) _____

Q11 What content areas do/did you teach? Select all that apply.

- ▢ English/Language arts (1)
- ▢ World Languages (2)

- ▢ Mathematics (3)
- ▢ Science (4)
- ▢ Social Studies (5)
- ▢ Health & Physical Education (6)
- ▢ Fine Arts (7)
- ▢ Career and Technical Education (8)
- ▢ Special Education (9)
- ▢ Other (Please specify) (10) _____

Q12 To which gender do you most identify?

O Male (1)

O Female (2)

O Prefer not to say (3)

O Prefer to self-describe: (4) _____

Q13 How would you describe your ethnicity?

O Hispanic or Latino (1)

O Indian or Alaska Native (2)

O Black or African American (3)

O White or Caucasian (4)

O Asian (5)

O Native Hawaiian or Other Pacific Islander (6)

O Other (Please Specify) (7) _____

End of Block: Introduction and Demographic Questions

Start of Block: Short Response Questions

Q14 How do you define authentic assessment? Do you believe that the assessments that you are currently giving to students matches your definition of authentic assessment?

Q15 What type of local alternative assessments are you implementing in your classroom? Check all that apply.

- ▢ Interdisciplinary Problem-based Assessment (1)
- ▢ Document-Based Question (2)
- ▢ Inquiry Project (3)
- ▢ Research Project (4)
- ▢ Multiple Choice Test (5)
- ▢ Portfolio (6)
- ▢ Public Product (e.g., Podcast, YouTube Video, Letter, etc.) (7)
- ▢ Public Speaking (8)

Q16 As you know, in 2014, the state transitioned away from end-of-the-year multiple-choice standardized tests in U.S. History I and II. How has your division replaced them? Please describe your local alternative assessments.

Q17 Who designed the local alternative assessment? (e.g., a department chair, a collaborative team of teachers, division supervisors, etc.) To the best of your ability, describe the design process.

Q18

To what extent are you able to make modifications to your local alternative assessment?

Q19 In 2-3 sentences describe your divisions' implementation of local alternative assessments in U.S. History I and II. How is implementation going? What challenges has your division experienced?

Q20 What is your opinion of the adoption of local alternative assessments in lieu of the multiple-choice SOL? Is this good for students? Is this good for teachers? Do you think anything about the policy needs to change?

End of Block: Short Response Questions

Start of Block: Likert Items

Q21 Questions About Implementation

	Strongly Disagree (1)	Disagree (2)	Agree (3)	Strongly agree (4)

I would describe our end-of-course assessments in U.S. History I and II as performance assessments. (1)	O	O	O	O
Our divisions' U.S. History I and II end-of-course assessments are aligned with the Virginia Standards of Learning. (2)	O	O	O	O
The state has effectively implemented local alternative assessments in U.S. History I and II classrooms. (3)	O	O	O	O
The states' approach to the implementation of local alternative assessments has allowed our division to have autonomy. (4)	O	O	O	O
The data collected from local alternative assessments has been more valuable than the data collected from our previous division benchmarks. (5)	O	O	O	O

Q22 Questions About Teachers

	Strongly Disagree (1)	Disagree (2)	Agree (3)	Strongly agree (4)
Teachers benefit from the policy change towards local alternative assessments (1)	O	O	O	O
Teachers spend less time on test preparation thanks to the state's implementation of local alternative assessments (2)	O	O	O	O
Teachers' instruction has improved because the state has adopted local alternative assessments (3)	O	O	O	O
Teachers' collaboration has improved because the state has adopted local alternative assessments (4)	O	O	O	O
The state's adoption of local alternative assessments has created an unreasonable amount of work for teachers (5)	O	O	O	O
Teachers feel empowered by the implementation of local alternative assessments (6)	O	O	O	O

Q23 Questions About Students

	Strongly Disagree (1)	Disagree (2)	Agree (3)	Strongly agree (4)
Students benefit from the policy change toward local alternative assessments (1)	O	O	O	O
Performance-based assessments are a fair way to judge student mastery of content in a course (2)	O	O	O	O
Students learn social studies best when they are engaged in performance-based tasks and assessments (3)	O	O	O	O
All students can learn challenging social studies content through performance-based tasks and assessments (4)	O	O	O	O
Our economically disadvantaged students are better served because the state has implemented local alternative assessments (5)	O	O	O	O

The VA SOL multiple choice test for U.S. History I and II was a more effective measure of student knowledge than our current local alternative assessment (6)	O	O	O	O

Q24 Questions About Frequency

	More than once a month (1)	Once a month (2)	Once a quarter (3)	Once a semester (4)	Once a year (5)	Never (6)
How often are performance-based assessments administered, on average, in the secondary social studies classrooms you supervise? (1)	O	O	O	O	O	O
How often do you believe is the ideal frequency off administration of performance-based assessments? (2)	O	O	O	O	O	O

How often do teachers participate in professional development to prepare them to develop and administer performance-based assessments? (3)	O	O	O	O	O	O
How often, on average, do teachers meet to discuss planning and grading performance based assessments? (4)	O	O	O	O	O	O

End of Block: Likert Items

Start of Block: Exit Questions

Q25 Who else in your division would be a good resource to discuss local alternative assessments in U.S. History I and II with? (e.g., Teachers, Department Chairs, Administrators). Include e-mail address if possible.

**Remember that your answers in this study will remain confidential. We will minimize any risks by collecting data through a password protected survey tool (Qualtrics) and maintaining data on the password-protected file-hosting service UVA Box. All data will be de-identified and at the conclusion of the study the data will be destroyed.

Q26 Would you be willing to discuss this topic further in a follow-up interview?

O Yes (1)

O No (2)

End of Block: Exit Questions

Appendix B
Social Studies Coordinator/Curriculum Leadership Interview Protocol

1. **Could you please describe your role in the county?** *(Specific title may be used if known from survey data)*
    a. *Sub-question: What are your main objectives as social studies coordinator? How long have you served in this role?*
2. **Could you please describe what you're looking for regarding assessment in social studies classrooms? What does successful assessment of student learning look like to you?**
3. **Could you describe if you think that your division has developed a local alternative assessment for U.S. History I and II?**
    a. *Sub-questions: If so, how was it developed? How were you involved in the process? When did this process begin? Was there any training involved for the implementation of this assessment?*
4. **Could you describe if you have attended any VDOE workshops/conferences regarding the creation or implementation of such assessments?**
    a. *Sub-questions: Did these workshops/conferences change your beliefs about assessment? Did these workshops/conferences help with your instruction?*
5. *If the division has not developed an assessment yet*—**Would you be willing to have your social studies teachers work together to create such an assessment or should the division perhaps take on this role? What are your thoughts on the creation of such an assessment?**
6. **How do you think your social studies teachers use assessment data? Do they work together or separately to create assessments?**
    a. *Sub-questions: Do they use it to have a grade? To meet state requirements? To inform instruction?*
7. **Please describe if social studies teachers' assessments are tied to their evaluation as a teacher.**
    a. *Sub-questions: SMART Goals? Are there any professional incentives or consequences for teachers attached to how well students do on this assessment? Are there any incentives or consequences for you as a principal, if students do or do not well on teachers' assessments?*
8. **Do you think your beliefs about assessment match the beliefs of your social studies teachers? Do they match the aims of the VDOE?**
    a. *Sub-question: Would you change, in any way, how students are assessed at your school?*

Appendix C
Quantitative Trends in Survey Data

Questions About Implementation

- 93% said they agree or strongly agree with describing their assessments as performance assessments
- 93% agree or strongly agree that their assessments are aligned with the SOLs
- When asked if the state has effectively implemented local alternative assessments results were mixed, though most agreed or strongly agreed (~62%)
- 89% agree or strongly agree that the state's approach to the implementation of assessments has allowed their division to have autonomy
- When asked if the data collected from LAA was more valuable than data collected from benchmarks responses were mixed, with most selecting 'disagree' (~48%)

Questions about Teachers

- 93% agree or strongly agree that teachers benefit from the policy change
    - 64% agree, 28% strongly agree
- 70 % agree or strongly agree that teachers spend less time on test prep thanks to the state's implementation of LAA.
    - 48 % agree, 20% strongly agree
- 73% agree or strongly agree that teachers' instruction has improved because the state adopted assessments
    - 53% agree, 20% strongly agree
- 83% agree or strongly agree that teacher collaboration has improved
- 75% disagree or strongly disagree that the policy change has created more work for teachers
- 72% agree or strongly agree that teachers feel empowered by the new policy
    - 62% agree, 10% strongly agree

***When asked about changes in instruction or empowering teachers, division leaders were less likely to mark "strongly agree" even if they agreed

Questions About Students

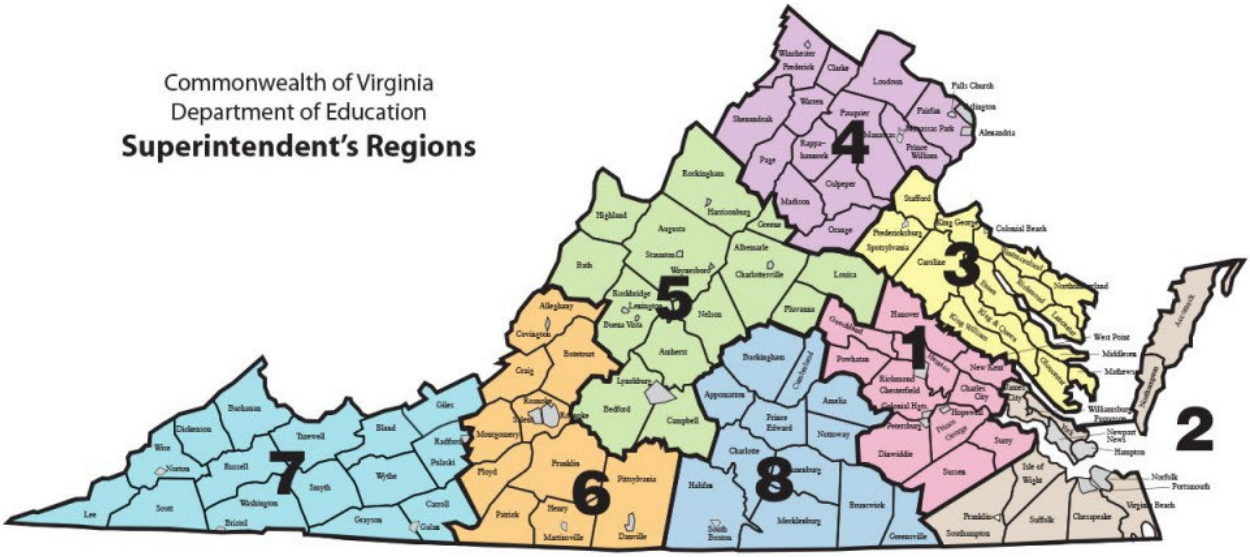
- 96% agree or strongly agree that students benefit from the policy change
- 86% agree or strongly agree that performance assessments are a fair way to judge student mastery of content in a history course
- 97% agree or strongly agree that learn social studies best when they are engaged in performance-based tasks and assessments
- 86% agree or strongly agree that economically disadvantaged students are better served through this policy change
- 86% disagree or strongly disagree that the VA SOL multiple-choice tests for U.S. I and II were more effective measures of student knowledge compare to the new alternative assessments

Questions about Frequency of Assessment

- In terms of frequency, divisions tend to administer performance-based assessments, between once a month and once or twice a semester. Twice a semester (or once a quarter) was the most common answer.
    - 32% once a month, 39% once a quarter, 25% once a semester
    - One respondent marked “never”
- In terms of frequency division leaders tend believe that performance assessments should be administered *more frequently* than they currently are.
    - 18% more than once a month, 32% once a month, 29% once a quarter
    - Two respondents marked “never”
- 52% believed teachers should participate in PD to prepare to develop and administer performance assessments once a year
- Division leaders were varied in response to how often teachers meet to discuss planning and grading performance assessments – all responses from more than once a month to never were represented.
    - Once a quarter was the most common response (28%)

Appendix D

Map of VDOE Superintendent's Regions



Toward A Shared Praxis:
Best Practice Assessment in Secondary Social Studies

Michael Gurlea
University of Virginia

The field of social studies education has spent decades debating its goals and purposes but recently has coalesced around *The College, Career, and Civic Life (C3) Framework* which emphasizes disciplinary skills of inquiry, including: gathering and evaluating sources, developing claims using evidence, communicating and critiquing conclusions, taking informed action, civic participation and deliberation, economic decision-making, and historical argumentation and perspective-taking (National Council for Social Studies [NCSS], 2013). This framework – aimed at influencing what is taught and how it is taught – presents a clear vision of effective instruction centered on disciplined inquiry. Questions remain, however, about how, within the social studies classroom, these ambitious curricular goals and effective instruction will be tangibly linked to assessment (NCSS, 1999) and whether the adoption of this framework influences "how we know what students know" (Grant, 2017, p. 461).

Historically, the most common assessment used to measure student progress in social studies has been the multiple-choice test (Goodlad, 1984; Martin et al., 2011). Multiple-choice tests, while designed to measure student fact-recall and content knowledge (Haladyna 1999; 2004), have been shown to not have much utility in measuring more ambitious disciplinary goals (Reich 2009; 2013; Smith et al., 2019). The *No Child Left Behind* (*NCLB) Act* privileged the use of these tests, but the arrival of *The Every Student Succeeds Act (ESSA)* and the *C3 Framework* presents an opportunity to ask: how is the field of social studies approaching assessment in a changing context? This question is important because assessment is the way we understand the invisible process of learning and measure how students are (or are not) progressing toward learning goals.

While recent literature reviews have investigated the disconnect between the instructional aims of social studies and the widespread use of multiple-choice tests both on the large- and

classroom-scale (Grant & Salinas, 2008; Torrez & Claunch-Lebsack, 2013), the current shifting assessment context offers a unique opportunity to return to the literature to synthesize and critically analyze the current state of research on large-scale and classroom-based assessment in secondary social studies. In conducting this systematic review of the literature, this manuscript seeks to 1) examine recent empirical evidence surrounding classroom-based and large-scale assessment in secondary social studies, 2) identify and discuss the themes within the research base in social studies assessment and consider gaps within the literature, and 3) discuss the implications for future research and how we might begin to connect ambitious curricular and instructional goals presented by the *C3 Framework* to disciplinary assessment. In order to understand the recent empirical literature, however, it is important to elucidate the findings surfaced in past reviews and situate these findings in the current policy context.

Past Reviews

In their chapter in *Handbook of Research in Social Studies Education* entitled *Assessment and Accountability in the Social Studies*, Grant and Salinas (2008) focused on empirical and theoretical works related to large-scale assessment and accountability in secondary social studies. Large-scale assessment was defined and outlined by the authors as, "standardized, social studies tests developed, administered, and scored by state departments of education and/or their designees" (Grant and Salinas, 2008, p. 220). They found both the empirical and theoretical base of research on social studies assessment to be thin and called for more research and theory surrounding assessment in social studies at the state and national level. They argued that, given the paucity of research, it was not possible to make empirical claims about the relationships between the large-scale assessment policy context and improved teaching or student outcomes. Several years later, in a chapter for the *SAGE Handbook of Research on Classroom Assessment*,

Torrez and Claunch-Lebsack (2013) took a different approach, focusing on classroom-based assessment in social studies. Their definition of classroom-based assessment acknowledges that it can be formative or summative in nature and they argue that "The primary characteristic of social studies classroom-based assessment is that it serves a constructive purpose; it benefits teaching and learning" (Torrez and Claunch-Lebsack, 2013, p. 462). While the authors found some research that explored the use of disciplined inquiry, project-based learning, and other ambitious means of classroom-based assessments these examples represented the exception rather than the rule. They contend that scholarship had largely focused on high-stakes testing and its implications, but "not on what happens in classrooms with assessment" (Torrez & Claunch-Lebsack, 2013, p. 468). Both of these reviews of the assessment literature in social studies made calls for more empirical inquiry in the realm of assessment.

Other recent scholarship has focused not on the discipline of social studies broadly, but specifically on assessment in history education. These works have explored the goals of the field, the design of assessments, and the impact of large-scale assessments (Ercikan & Seixas, 2015; Shemilt, 2018). In their edited volume, *New Directions in Assessing Historical Thinking*, Ercikan and Seixas (2015) include several theoretical and empirical chapters on the goals of history education, issues with assessing various disciplinary skills in history, large-scale assessments in history, and the validity of score interpretations of history tests. Similarly, in his handbook chapter, Shemilt (2018) theoretically compares the changing landscape and goals of history education against the costs and benefits of various means to assess these more ambitious outcomes. These works explore both the validity of multiple-choice and constructed-response items as measures of historical thinking and corroborate earlier studies indicating that multiple-choice tests are not particularly good measures of ambitious disciplinary goals.

Grant's (2017) most recent handbook chapter draws on literature outside of social studies education in order to evaluate the validity of classroom-based and large-scale assessments. He corroborates other reviews in contending that we know very little about what happens in the classroom-based setting in social studies with regard to assessment. In the realm of large-scale assessment, he synthesizes work on psychometrics and claims that these tests often meet standards of reliability, but struggle to meet a high standard for validity. Further, echoing the tension between the goals of the field and the current state of assessment, Grant (2017) questions whether these large-scale test measure anything of value. To that end, Grant (2017), provides one of the only descriptive explanations of the changing testing landscape across social studies in the era of ESSA by describing funded projects, single-state and multi-state initiatives designed to develop alternative assessments to multiple-choice tests on the large-scale.

Current Context

The reviews described above largely focus on the era of *NCLB* and the impact of this policy on state accountability systems and classrooms. This policy context pressured social studies teachers to focus on the coverage and control of tested content and to often teach in ways that supported success on an end-of-course assessment rather than growth in disciplinary knowledge and skills (Au 2007; 2009). *NCLB* also pushed researchers to focus on the pressures of the high-stakes test, and pay less attention to classroom-based assessment, as Torrez and Claunch-Lebsack (2013) outline in their recent review of the literature:

> American social studies classrooms, for the early part of the 21st century, have been in a near state of atrophy as the disconnect between classroom realities and best social studies practices widens. Diminished social studies instruction has resulted in the silencing of constructive assessments of student learning which may have constrained new classroom research assessment projects (Torrez & Claunch-Lebsack, 2013, p. 465)

With the arrival of the *Common Core Standards* (and associated *C3 Framework*) and the new policy context of *ESSA*, states are now able to choose measures such as portfolios and performance-based assessments to measure student growth rather than multiple-choice tests (Darling-Hammond, et al., 2016). In social studies, specifically, we have evidence that a small number of state-level entities are replacing or adding to large-scale multiple-choice assessments with performance-based assessments (Grant, 2017). Due to this changing landscape a return to the literature on assessment in secondary social studies is warranted. What do we know about tracking student progress towards ambitious disciplinary goals? What do we know about how districts, schools, and teachers are navigating changing policy contexts? Can new avenues for summative assessment accurately measure more than social studies content knowledge? Are teachers' classroom-based instructional and assessment practices influenced by this shifting policy context?

From this wide range of potential theoretical and empirical perspectives explored by the literature, I chose to focus only on empirical studies across the disciplines of secondary social studies. I chose not to explore theoretical perspectives in this review as many recent handbook chapters and books have included these in their approach. As these reviews have all, to some extent, highlighted the lack of empirical research surrounding assessment in social studies, I believed it was important to outline what conclusions can be drawn from recent empirical studies. Because of the small body of empirical studies available, it was possible to include studies that focused on both large-scale and classroom-based assessment.

The years since 2008 are significant as, in this time, the field of social studies education has coalesced around the *C3 Framework* for its ability to communicate ambitious disciplinary goals. In addition, in the years since Grant and Salinas' review, the federal and state policy landscape

has shifted. *ESSA* has allowed states more freedom in how they choose to construct assessments. We know very little about this shifting landscape except that some states are working to adopt ambitious large-scale assessments in social studies that attempt to measure higher-order thinking skills. These assessments vary in character: We have some evidence of states developing performance-based assessments in social studies (Grant, 2017) but also of some states taking a "balanced" assessment approach – continuing to use multiple-choice tests in addition to questions that require a written response (see Miller, 2018; Meuller & Colley, 2015).

Methods

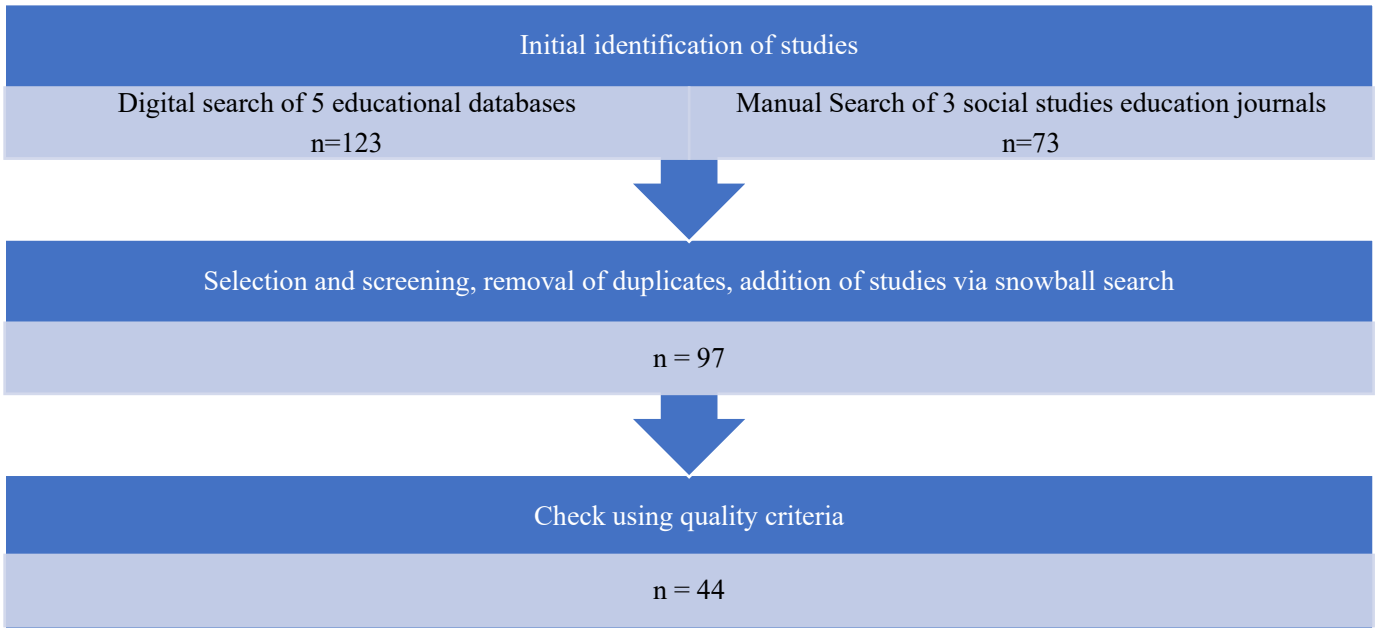
This extensive and systematic search of the assessment literature in social studies education was conducted using five online databases: 1) the Education Resources Information Center (ERIC), 2) Academic Search Complete, 3) Education Research Complete, 4) Psychology and Behavioral Sciences Collection, and 5) APA PsycINFO. The search process was conducted in two broad phases. In the first phase, articles were selected based on the follow criteria: 1) includes the keywords *secondary education, assessment, social studies/social science/history* in the abstract, 2) empirical or theoretical articles in English published in academic peer-reviewed journals and dissertations, and 3) published between January of 2008 and December of 2021. The second phase was expanded to include keywords that were emergent in the first phase of the search. These key words may include terms associated with assessment in social studies such as *performance-based, project-based*, *inquiry,* and *discussion.* After an initial screening of article titles and abstracts these searches resulted in the inclusion of 123 potential articles for further scrutiny and potential inclusion.

Using the same keywords and date ranges, I also manually searched *Theory and Research in Social Education* (TRSE), *The Journal of Social Studies Research* (JSSR), and *The History*

Teacher as these journals publish work that is particularly relevant to the secondary social studies classroom. This manual search resulted in the initial discovery of 73 articles relevant to the topic of assessment in social studies education. Using a snowball sampling method, the introductions and literature reviews of these manuscripts were also searched for empirical research that might also be relevant. The results of this manual search and snowball sample were then checked against and merged with the results of initial systematic search. After eliminating duplicate articles and conducting an initial screening process, the total of articles before the application of an inclusion/exclusion criteria was 97. These articles then underwent a further quality criterion check (described below) for inclusion/exclusion. The final number of articles included in this review was 44. For a visual representation of this process, see Figure 1.

Figure 1

Systematic Review Process



Analysis of the data yielded by these searches included descriptive information about each study including the nature of the article (i.e., empirical or theoretical), the focus of the study

(i.e., teachers, students, or curriculum), and the type of journal. The articles were analyzed following the framework suggested by Wilson and Anagnostopoulos (2021) in their methodological guidance piece concerning qualitative literature reviews. This framework determined the quality of the qualitative research surfaced (Garcia et al., 2019) and situated the research more broadly in the current contexts within the field of social studies education. The following sections outline the process of screening, selection, and quality check and the frameworks borrowed from the methodological guidance of Wilson and Anagnostopoulos (2021).

Selection and Screening Process

All resulting publications from the phase one and phase two searches were included in a systematic review and judged against an inclusion criterion. Studies were included if there was an explicit or implicit link to assessment in the secondary social studies classroom. For example, while some studies did not explicitly mention "assessment" they were included because of the inclusion of student outcome data and a discussion of its relationship with any number of variables. Some studies were excluded as they were not about assessment in secondary social studies classrooms but were still captured by our search. For example, the search criteria sometimes captured articles that referenced the "history of education" or a student's "educational history" in their abstract, but were not relevant to this systematic review. Despite the search terms, the initial search also surfaced many manuscripts related to assessment at the elementary level, which were excluded as they go beyond the scope of this review.

Quality Criterion

While the intention of the search was to be as inclusive as possible, it was necessary to apply a quality criterion to articles to determine whether they were fit for inclusion in this

systematic review. The 97 articles that surfaced through the initial selection and screening process were systematically assessed for quality. In order to determine quality, a list of questions (developed by Garcia et al., 2019) was applied to each article (see Table 1). The data from these questions was applied to an annotated bibliography spreadsheet in which each question was coded with a 0 (no) or a 1 (yes). Garcia and colleagues (2019) determined that a 6 or higher on the empirical scale along 10 criterion measures and a 4 or higher on the nonempirical scale along 6 criterion measures would warrant inclusion in the review. This same standard was applied to the articles for this review. As this systematic review only focused on empirical studies, the rule of 6 was applied to this review. Studies that scored lower than a 6 often had a misalignment between their research questions, their sample, and their findings. This process resulted in a final count of 44 articles for inclusion in the literature review. See Appendix A for a table of articles included and their final organization based on data analysis.

Table 1

Quality criterion

Category	Question
Empirical	1. Is the article empirical? 2. Is the research purpose or objective clear? 3. Is the literature review, conceptual, or theoretical framework appropriate and driving the research questions and/or methods? 4. Is the method used appropriate for addressing the purpose or objective? 5. Is there sufficient sample/data to address the purpose or objective? 6. Is the research context adequately described? 7. Is the analysis adequate or appropriate for addressing the purpose/ objective? 8. Are the results findings clearly presented and connected to the data? 9. Are the methodological limitations and or trustworthiness stated? 10. Are the conclusions drawn from or connected to the data and empirical evidence?
Nonempirical	1. Is the article nonempirical (descriptive or theoretical or program evaluation or trends analysis)? 2. Is the research purpose or objective clear?

	3. Is the problem statement, introduction, literature review, conceptual or theoretical framework appropriate and connected to the purpose? 4. Is the description or theoretical argument or evaluation methods sufficient for responding to the purpose? 5. Is the context adequately described? 6. Are the conclusions drawn from the description or theoretical analysis or evaluation?

Data Analysis

Data analysis resulted in descriptive information for each article including the journal of publication, country/state of study, and the focus of the study (i.e., students, teachers, standards, or multifocal). All empirical studies in the data set were sorted by their methodologies, that is, whether they used quantitative, qualitative, or mixed methods. Studies were further categorized and subdivided by their disciplinary focus within social studies (i.e, history, economics, civics, and geography) and their assessment focus (classroom-based assessment or large-scale assessment). A thematic review of the literature was undertaken in order to review the content of the articles included in the final systematic review. This process included multiple iterations of reading and analytic note taking in order to capture the key themes within the literature base.

Results

Data analysis resulted in the organization of the literature across three major themes 1) Studies of assessments that attempt to measure social studies content knowledge and skills using the multiple-choice format, both at the large- and classroom-scale, 2) Studies of ambitious assessments that attempt to measure higher-order disciplinary goals, and 3) Studies that center teachers and students navigating complex and changing assessment contexts.

This review of the literature echoes the results of previous reviews (Grant & Salinas, 2008; Torrez & Claunch-Lebsack, 2013). While there is a small body of compelling empirical literature that explores ways to measure student growth toward ambitious disciplinary goals,

teachers, schools, and students largely still exist in the policy and practice vacuum in transition between *NCLB* and *ESSA*. Large-scale multiple-choice measures still report achievement disparities along sociocultural variables (i.e., race, class, and gender). Teachers still feel the pressure of state-level end-of-course assessments: which impacts their classroom-based instructional and assessment practices.

An even smaller body of literature has begun the important work of exploring how teachers and students navigate new assessment contexts. Further, it is important to note that classroom-based formative assessment, which, as Shepard (2000) articulated, is most closely related to the processes of teaching and learning, remains largely unstudied in social studies. I argue that the current state of the assessment literature in social studies calls for further inquiry into the processes and practices of classroom-based assessment: both in how teachers navigate complex and changing assessment contexts and in how they and their students experience formative assessment for learning rather than summative assessment of learning.

Multiple-Choice Testing: The Pressures of Large-Scale Assessment

The empirical studies surfaced in this review build on the body of literature that provide evidence that large-scale multiple-choice tests are not particularly good measures of students' pursuit of more ambitious skills within the disciplines of social studies, and, further, that there are several issues with this test format in measuring disciplinary content knowledge. Teachers, as an artifact of the NCLB-era of high stakes accountability and as a result of accountability measures at the state level, still feel pressure to teach in ways that serve end-of-course tests in social studies.

Multiple-Choice Large-Scale Assessments

Multiple-Choice Measures. A small body of literature has explored student think-alouds on multiple-choice test items in order to capture student reasoning and assess item validity. Reich (2009; 2013) conducted interviews with 13 students using items from the 10th grade New York Global History and Geography Reagents Exam. Reich explored how students employed historical reasoning and thinking while examining the validity of the multiple-choice test items. Findings indicated a clear misalignment between the disciplinary goals of history education (i.e., thinking like a historian) and what was measured by the multiple-choice test (i.e., content knowledge, literacy, and test-taking skills).

Reich's (2009; 2013) findings are further corroborated and expanded by the work done by Mark Smith (2018) in his exploration of the effectiveness of the test items on the *Historical Thinking Test* (*HTT*). The *HTT* was a measure designed by Reisman (2012) in order to capture student success in engaging in historical thinking using multiple-choice measures. Smith (2018) explored the validity of these measures by conducting think aloud interviews with 12 high school juniors. Smith found that in answering the sample multiple choice items from the *HTT* participants did, to varying degrees, tap into historical thinking skills such as sourcing, corroboration, and contextualization. However, students were also able to reach correct answers on these test items by engaging in construct irrelevant reasoning such as guessing, analyzing irrelevant item features (i.e., the student was able to determine the correct answer because of contextual information from the question), pure fact recall (i.e., a student did not have to engage in a historical thinking process because they already knew the content), and decontextualized reasoning (i.e., the student was able to determine the correct answer using ahistorical logic).

Smith (2017; 2018) also explored the strength of multiple-choice test items on the NAEP US History test in eliciting historical thinking processes. The NAEP US History test is

administered to a nationally representative sample of US 8th grade students in order to determine achievement in US History. The NAEP Exam is often at the center of national discussion in the media upon its release, generating discourse and criticism about how much modern students actually know about US history (see Patterson & Shuttleworth, 2020). Smith's (2017) exploration of 27 high-school students' think aloud interviews considering NAEP exam items further echoes the findings of the studies mentioned earlier in this section: these multiple-choice items did not elicit student responses along the intended measure of "Historical Analysis and Interpretation" but rather engaged their use of test-taking strategies, fact recall, and literacy. These items were also often imprecise indicators of student content knowledge.

Large-Scale Analysis of NAEP Results. Beyond an exploration of the test-items themselves, several studies have explored the relationship between student outcome data on the 2010 12th grade NAEP US History and the 2006 12th grade NAEP Economics assessments and student characteristics and experiences of instruction. In US History content knowledge, there is still an 'achievement gap' on display in these NAEP results. That is, variance in achievement on this test is still somewhat explained by sociocultural variables (i.e., race, class, and gender; Fitchett et al., 2017; Heafner & Fitchett, 2015; Heafner & Fitchett 2017). These studies also report a positive relationship between student interest in the subject and text-dependent methods of instruction (i.e., reading, writing, and discussion) and success on the NAEP US History assessment. Heafner, Fitchett, and VanFossen (2019) found that these achievement gaps were also pervasive on the NAEP Economics assessment, with 17% of the variance of outcomes explained by sociocultural variables.

Largely the findings elucidated by these studies emphasize what we already knew about student performance on the NAEP US History and NAEP Economics tests but extend this

knowledge into the era after *NCLB*. A renewed focus in the field of social studies education on the need for culturally responsive teaching (Au, 2010; Gay, 2000; 2002; Salinas, 2006) since the previous analysis of NAEP US History results (see Zwick & Erkican, 1989) had not resulted in the narrowing of achievement gaps. These findings are further complicated by other recent research in social studies education, outlined above, that calls into question the ability of these high-stakes test items to measure content knowledge or skills well. As Grant (2017) explains, there has been much focus on creating test measures that are reliable (i.e., scores are consistent across groups and over time), but not valid (i.e., scores are measures of the content and skills goals we intend them to track).

State Standards and High-Stakes Tests. In their review of the literature, Grant and Salinas (2008) found that the expectations of testing and accountability in social studies varies widely from state to state. This ambiguity of state testing measures and the constant state of flux between which states are focusing on testing in social studies and which states are scaling back their state tests made it hard for Grant and Salinas (2008) to capture any national trends with regard to state-level testing. This state of the landscape remains largely unchanged, however two studies, drawn from the same larger project attempted to contrast the disciplinary goals of a number of these tests alongside student data considering their instructional experiences.

Dewitt and colleagues (2013) analyzed the state high school social studies standards and associated high-stakes tests across four states (i.e., Ohio, New York, Texas, and Virginia) to capture their cognitive demands and to capture the alignment between these standards and their accompanying tests using Bloom's Taxonomy (Anderson et al., 2001). In order to conduct this analysis, researchers used a dichotomous rubric divided between "higher order" and "lower order" thinking skills and applied this rubric to the state standards document and then to the

respective state's high stakes test. These rubrics were interrogated to explore the alignment of cognitive demands between the standards and the tests. Across each of the states, researchers found that standards often suggested higher-order thinking skills from students (e.g., critical thinking, problem solving), but there was a disconnect between the demands of the standards and the way that knowledge and skills were operationalized on the high-stakes summative assessments. These findings verify that the goals of the field of social studies education, while reflected in some ways in standards documents, are not well captured by high-stakes multiple-choice tests.

The study conducted by Dewitt and colleagues (2013) drew on a larger study conducted by John Saye and the Social Studies Inquiry Research Collaborative (SSIRC; 2013) which sought to connect student experiences of authentic pedagogy (i.e., teaching that aims to construct student knowledge through disciplined inquiry) to their performance on high-stakes tests. Across six states (i.e., Alabama, Georgia, New York, Ohio, Texas, and Virginia) researchers used a critical case sampling strategy to select history teachers (n=52) who had been identified by key peers as teachers who engaged in authentic pedagogy. Results suggested that, even within the sample of teachers selected for their use of high-leverage practices, students experienced relatively low levels of authentic pedagogy. However, the authors did note a consistent positive pattern between student experiences of authentic pedagogy and student scores on state tests. In other words, given the small chance that students did experience authentic pedagogy, they were more likely to perform well on the respective state-mandated assessment. Taken together these studies indicate a continued pattern in secondary social studies of a misalignment between the goals of the field and the tasks required on large-scale high-stakes tests. In order to understand

how this misalignment impacts instruction, it is important to turn to studies that seek to understand teachers' experiences of these high-stakes contexts.

Teachers in Large-Scale Assessment Contexts

Research since 2008 has contributed to the growing body of literature that connects a teacher's instructional decision-making with the pressures of the larger testing and accountability system they inhabit. How teachers make decisions around instruction and classroom-based assessment is undeniably influenced by the presence or absence of a high-stakes test at the state, district, school, or course level (Au 2007; 2009). Further, how teachers make decisions about what classroom-based assessments their students experience expands a vision of teacher as curricular-instructional gatekeeper that includes student assessment experiences (Thornton 1989; 2006).

Much of the work in recent years has corroborated the narrative that a high-pressure accountability context has an impact on teachers' instructional decision-making, specifically in the ways that stymie autonomy, limit instructional authority, and narrow the curriculum. Girard and colleagues (2021) found through an explanatory sequential mixed methods study that secondary history teachers reported feeling limited by required assessments. Through surveys (n=260) and follow up interviews (n=23), teachers stated that while they wanted to center content that was related to students lives and focus on historical significance rather than just historical facts, they felt constrained by what needed to be covered in service of the end-of-course high-stakes assessment. These participants' sense of control seemed to be more closely related to their broad assessment policy context (e.g., district assessment policy, state assessment policy, external testing requirements like the Advanced Placement test) as opposed to school-level factors.

Data from the *Survey of the Status of Social Studies (S4)* builds on this tension between assessment policy and control. The *S4* is a nationally-inclusive dataset that surveys social studies teachers on the organizational structure of the discipline, their instructional decision-making, their professional attitudes and their demographics (Fitchett & VanFossen, 2013). In a quantitative study, Hong and Hamot (2015; 2020) sought to understand associations between state testing policy, teachers' professional characteristics, school characteristics, and teachers' instructional autonomy. Findings indicated that teachers in low-income, high-minority schools with state-mandated tests reported much less instructional autonomy than their peers. This relationship, however, was mediated by experience and high-quality teacher preparation. That is, teachers with more experience or high-quality preparation may be able to better balance the demands of a high-stakes test with meaningful student learning.

Two recent studies have also explored how pre-service teachers make sense of and enact assessments in a broader high-stakes policy context. These studies are all small scale, qualitative case studies in association with social studies methods courses. Pre-service teachers experience of the high-stakes accountability context has the potential to narrow what they can learn in their student teaching experience as they must make sense of contradictions between their field experience and their coursework (Hawley & Whitman, 2020) and navigate high-quality disciplinary instruction with less guidance with regard to high-quality disciplinary assessment (Drake Brown, 2013). Pre-service teachers also observe the model of their mentor teacher navigating high-stakes accountability systems and may develop learned helplessness in focusing on success on a summative assessment rather than student learning (Hawley & Whitman, 2020).

Recent studies on large-scale assessment in social studies confirm the continued existence of an achievement gap and the potential bias and inaccuracy that exists in high-stakes multiple-

choice tests. Across the disciplines of social studies education, there is still a disconnect between the goals touted by researchers and the goals implied by high-stakes tests. As the shape of these large-scale assessments change under the freedom provided by *ESSA*, continued research is important to understand how these tests align with the ambitious disciplinary goals of history, civics, economics and geography.

Projects, Inquiries, and Performance-Based Assessments: Ambitious Classroom-Based Assessment

A small, but growing, body of research has emerged exploring ambitious assessments that aim to measure goals of social studies education that extend beyond history or civics content knowledge and include measures to assess skills outlined by the C3 Framework, such as gathering and evaluating sources, developing claims using evidence, communicating and critiquing conclusions, taking informed action, civic participation and deliberation, and historical argumentation and perspective-taking (NCSS, 2013). Most of these research projects are case studies, focusing on one particular school, classroom, body of students, or instructional tool. The nature of the results of these studies makes it difficult to generalize findings about the effectiveness of any of these given assessment tools across multiple contexts, but they do provide evidence that, at the classroom level, there are means for teachers to measure student growth towards more ambitious goals.

Projects

While "projects" might refer to any number of instructional endeavors a teacher and their students undertake over the course of a class period (or multiple class periods), *Project-Based Learning (PBL)* is a discrete and defined method of instruction. Rigorous projects under *PBL* have four characteristics: 1) They must carry the full subject matter load of the course, 2) They

must be authentic, connecting to the world beyond the classroom, 3) They must focus on a meaningful learning goal, and 4) They must include and appropriate, external, summative assessment (Parker & Lo, 2016). While a *PBL* curriculum is not necessarily an assessment in-and-of-itself the nature of inquiry, the connections to the world beyond the classroom, and the need for a summative assessment allows teachers and policymakers to use projects to track and measure student growth toward a number of disciplinary goals. Recent studies have explored the benefits and challenges of enacting a project-based curriculum and connected assessment both inside and outside of the traditional classroom.

Action Civics, a summer civics institute offered to 5th to 9th graders, while not classroom-based, provides evidence of activities and assessments that improve student efficacy across six competencies of civics achievement defined as: 1) 21st Century Positive Youth Leader, 2) Active and informed citizen, 3) Academically successful student, 4) Youth Civic Participation, 5) Youth Civic Creation, and 6) Civic and Cultural Transformation. (Blevins et al., 2016; LeCompte et al., 2020). In order to meet and measure these competencies the *Action Civics* project engages students in a process of research, action, and reflection about locally and personally relevant problems in government. Blevins and colleagues (2016) explored two iterations of the week-long summer institute through a mixed methods study. Through engaging in a curriculum that focused on the powers and process of local government and participating in authentic civic experiences (e.g., meeting local leaders, participating in mock trials, reading and interacting with primary sources, engaging in research, and group discussion) students improved their civic competency and engagement as assessed by semi-structured interviews with 36 students and pre- and post-institute survey data constructed to assess civic knowledge and engagement. LeCompte and colleagues' (2020) related study analyzed four years of student survey data (n=295) to determine

that the *Action Civics iEngage* program was effective in improving students' practical civics skills and knowledge (e.g., to care about a local problem, to organize a meeting, to write formal letters to community leaders, and express their opinion in the written form). The methods of the iEngage program could be explored in civics and government course contexts to improve student outcomes on measures beyond civics content knowledge and skills.

Levy (2011) explored similar civic advocacy projects and their relationship with students' political efficacy in a mixed methods case study of one high school course. Through classroom observations, interviews, and surveys, Levy found a positive relationship between students' participation in civic advocacy projects and students' development of political efficacy (i.e., belief that an individuals' actions can influence political processes) and understanding of the challenges of political participation. The case study teacher, Mr. Kendall, provides an exemplar case of an instructor understanding the complexity of skill building and scaffolding necessary for students to engage in an ambitious project. His scaffolds included whole class discussions of ethics, instruction on effective communication, source evaluation, and self-evaluation. Levy's (2011) case is instrumental both in its findings surrounding student progress toward political efficacy and in its focus on the formative feedback loops necessary to support student success in an ambitious assessment.

In a key study exploring the implementation of a project-based curriculum, Walter Parker and a team of researchers conducted design-based studies in AP US Government and Politics courses in two high schools (Parker et al., 2011; Parker et al., 2013). Researchers worked closely with teachers in one setting to develop projects that captured both government content knowledge and emphasized active learning. Student outcomes on the AP test were compared between the students who experienced the project-based curriculum and the students who

experienced traditional models of instruction and researchers found that the students who experienced the project-based curriculum performed at least as well as their peers on the AP test. This finding provided evidence that the traditional breadth versus depth dilemma in the AP setting might be mitigated by using high-quality projects rather than solely direct instruction.

The *National History Day (NHD)* project is an example of a disciplined inquiry project that has students (Grades 6 – 12) engage in original historical research. The *NHD* program is intended to be able to be integrated and implemented into any social studies classroom across the United States. The *National History Day* organization (2011) published the results of their own longitudinal study of the effects on history achievement of engaging in these historical research projects. Using data from multiple states from 2008 and 2010, the authors explored the relationship between participation in the *NHD* project and measures of student success (e.g., statewide history tests, AP tests, end-of-course exams, grades, and feedback from teachers). Results indicated that the NHD participant students outperformed their peers across these measures of student success. As Torrez and Claunch-Lebsack (2013) state in their review of the literature on classroom-based assessment, however, more research is necessary to understand the formative and summative assessment feedback loops and processes that underly this student success in the *NHD* project. Fehn and Schul (2011) corroborate these findings through their analysis of *NHD* contest winners in the documentary filmmaking category. They found that students practiced and displayed skills aligned with the disciplinary goals of history education such as source analysis, interpretation, evaluation, and synthesis in the creation of their projects. Documentary filmmaking, in general, may represent a project-based method of student creation that can promote authentic intellectual work (Swan & Hofer, 2013).

In sum, the recent works on *PBL* provide a hopeful picture of the potential benefits of students engaging in a project-based curriculum and the data provided to teachers by the associated assessments. However, the call for research made by Torrez & Claunch-Lebsack (2013) remains important: More studies should analyze not only how participation in projects can enhance student progress toward disciplinary skills, but also how the formative, instructional assessments, that take place within the project-based setting, build student knowledge, and inform instruction. Further, we know very little about what happens when a district or school chooses to implement a project on a larger-scale.

Classroom-Based Discussion

Very little research has explored the potential of classroom-based discussion as a means of assessment, but two studies in civics suggest that it may have value as both a tool for teaching and for formative and summative assessment, especially when teachers are given the resources to facilitate and evaluate high-quality discussion. Lin and colleagues (2016) conducted a randomized intervention study of middle school students (n=5870) and collected their reports of political self-efficacy after exposure to the *Word Generation* program. *Word Generation* engages students in learning academic vocabulary through discussion of a controversial issue each week. Students who engaged consistently with this academic vocabulary did self-report increased confidence in participating in discussions surrounding controversial issues. *Word Generation* is cross-disciplinary, and it is important to note that Lin, Lawrence, and Snow (2015) also found an increase in quality of classroom discussion using this intervention in a science classroom. Kohlmeier and Saye (2019) also studied student reasoning in a classroom-based discussion and found that students engaged in higher-order thinking and reasoning when teachers purposefully instructed students on the norms of quality deliberation, explicitly practiced discussion, and

allowed students to engage with each other rather than just the instructor. There is no recent research that considers any form of discussion as an assessment of student learning in social studies content, despite the central role that discussion plays in considerations of effective disciplinary practice in social studies (see Barton & Avery, 2016).

Historical Perspective Taking

Outside of the United States there is an increased focus on other disciplinary goals such as historical perspective taking (*HPT*; i.e., the ability of students to understand how people in the past saw their world) and historical empathy (i.e., the ability of students to understand the thoughts, actions and feelings, of people in the past). The studies in this section, while not explicitly about classroom-based assessment, attempt to measure students along these two competencies and are included to outline a means of assessment toward a different kind of historical thinking.

Huijgen and colleagues (2017) surveyed 15 and 16 year-old students (n=143) in the Netherlands to assess their ability to contextualize the actions of actors in the past. In follow-up interviews with a subset of participants (n=36) the authors explored students' reasoning. The authors found that students fell into classic traps that limited their ability to engage fully in *HPT* such as a focusing on a present-oriented perspective. The authors indicate that a baseline historical knowledge is important in preventing presentism and that *HPT* can be greatly improved when students make affective connections with actors from the past. Rantala and colleagues (2016) similarly found limitations in Finnish students' ability to engage in historical empathy after participation in a role-taking simulation of the Finnish Civil War. The authors observed the simulation across two 75-minute history classes and conducted follow-up interviews with 22 high school participants. Despite their participation, most students did not

reach the goals set for historical empathy, still considering the past as dark and unknowable and engaging in generalizations about the actions of individuals from the past. If the field of social studies education deems *HPT* and Historical Empathy as disciplinary goals worth pursuing, these studies push us to think how students might show success along these measures.

Constructed-Response Items

A broader category of assessment, constructed-response items include any measures of student learning that include a written response to a disciplinary question or prompt. These prompts might include a number of primary or secondary sources to elicit student thinking toward goals beyond content knowledge, such as interpreting and evaluating sources and historical thinking. Many of the studies on constructed response items focus on civics (Civic Online Reasoning) and history (History Assessments of Thinking) and come from ambitious projects pursued by a cohort of scholars at the Stanford History Education Group. The scholarship collected here provides some convincing evidence that short constructed response items may be useful in both classroom-based and large-scale assessment in evaluating student progress toward disciplinary goals.

McGrew and colleagues (2018) set out to develop a measure to assess students *Civic Online Reasoning*. The authors defined *Civic Online Reasoning* as a students' ability to search for, evaluate, and verify information online by asking the questions: Who is behind the information? What is the evidence? What do other sources say? The authors analyzed the ability of middle (n=405), high, (n=348), and college (n=141) students to complete short constructed-response tasks related to *Civic Online Reasoning*. Researchers found that students largely struggled with the tasks associated with evaluating sources, rarely asking who created a given source and often making judgements about the trustworthiness of a source based on its

appearance alone. Participants, when invited to do so, rarely sought resources outside of what they were provided (McGrew et al., 2018). These short assessments provided researchers with insights into student conceptions of civics that went beyond content knowledge and toward the disciplinary skill of evaluating sources.

A series of projects from Joel Breakstone, Mark Smith, and Sam Wineburg have explored the viability of short constructed response items in order to measure students' ability to think like a historian. Breakstone's (2013) dissertation undertook the process of design and implementation of these new formative assessments called *History Assessments of Thinking* (*HATs*). These items often include a primary source image or document and several open-ended constructed-response questions to elicit historical thinking in students. A sequence of studies have shown the validity of these measures in eliciting students' proficiency in historical thinking skills and shown their strength over traditional multiple-choice test items (Breakstone et al., 2013; Breakstone, 2014; Smith, 2018; Smith et al., 2019). *HATs* represent potential for quick, valid, formative assessments that go beyond measuring historical content knowledge.

Two studies, specifically regarding the discipline of history, have also explored student engagement with formal writing tasks as assessments, using the Document-Based Question (DBQ) in which students are given a disciplinary question and set of documents to read, interpret, and organize in an argumentative written response. Writing tasks, especially the DBQ, have been shown to require a high level of instructor feedback and student practice to be successful in the classroom. Students and teachers need to engage in activities that emphasize the interpretive nature of the discipline, that consider historical texts and interpretations, that practice supporting interpretations with evidence and engage in cycles of direction instruction, guided practice, and independent practice in order to be successful (Monte-Sano, 2008). Monte-Sano

(2010) provides a useful framework for assessing student responses to DBQs in her qualitative analysis of 56 high school juniors' responses to an essay question.

Wright and Endacott (2016) expand on Monte-Sano's findings in their quasi-experimental mixed methods study of the impact of inquiry-based instruction on sixth grade students' written argumentation as measured by the common core state rubric. The comparison of the quantitative rubric results between the control group (n=46) that received more traditional, didactic history instruction and the treatment group (n=59) that engaged in document-based inquiry instruction throughout the semester revealed minimal differences between the groups. Qualitative data analysis, however, revealed that students in the treatment group had improved responses when judged through a disciplinary lens, including it in their written responses arguments that showed strength in historical perspective taking, understanding cause and consequence, and using historical evidence. Wright and Endacott's (2016) study is important as it forces us to draw distinctions between student learning, measuring student learning, and scoring assessments. How do the rubrics we use to quantify student learning on a high-quality performance assessment reflect our disciplinary goals?

These studies of assessment using constructed-response items are powerful in their ability to highlight ambitious ways to assess disciplinary goals including and beyond content knowledge. However, we know very little about the use of constructed-response items in context. How does school context influence teachers' decision-making when using these tools as formative and summative assessments? How do students experience these assessment items in school context? Further, when a division or state uses a constructed-response item as a part of its high-stakes standards-based assessment what are the impacts on teachers and students? A small

number of researchers have begun to ask and answer these questions in an effort to understand assessment in context.

Navigating and Sensemaking New Assessment Measures and Contexts

In this final subsection of the results section, I will focus on key empirical studies that present a more complex picture of assessment in secondary social studies classrooms. Each of these studies captures the nature of this moment in assessment policy and research. These studies focus on student and teacher experiences of shifting assessment contexts, tensions in the field between the goals of the discipline and the measures of those goals, and suggest directions for future research.

Miller's (2018) study is unique as a recent study that explores the shifting assessment context as states move away from solely using multiple-choice tests as the summative measure of student performance. Tennessee (TN) represents a state that has recently included a written response item on their end-of-course 11th grade US History exam. The author explores three years of data from the Tennessee Department of Education and used multi-level modeling to analyze the association between school percentage English Learning (EL) and achievement on the restructured TN US History state assessment. Miller found that among ELs the addition of a written component to the end-of-course exam had a statistically significant negative relationship with their achievement in comparison to earlier years when the constructed-response item was not present. Miller's study emphasizes that while constructed-response items and performance-based assessments are sometimes considered as more equitable measures of student learning, educators and policy makers must pay careful attention to the way these assessments are implemented to ensure equity (see Darling-Hammond, 1994).

The research on large-scale assessment in social studies confirms the continued existence of an achievement gap and the potential bias that exists in high-stakes multiple-choice tests. As states move away from these exams under *ESSA*, Miller (2018) indicates that new kinds of assessments can put new demands on students that may exacerbate rather than alleviate pre-existing disparities. Across the field of social studies education there is still a disconnect between the goals touted by researchers and the goals implied by high-stakes tests. As these tests change shape to include constructed-response items and other visions of performance-based assessments, continued work is necessary to understand their alignment with the disciplinary goals of history, civics, economics, and geography.

The literature review surfaced three studies of teachers navigating the tension of an assessment context that was changing, or under which they were receiving multiple and mixed messages. Mueller and Colley (2015) used a goal-free evaluation case study to explore teacher perceptions (n=5) of a shifting context in Kentucky, where a new state-mandated US History exam designed by ACT-QualityCore endorsed a balanced assessment approach including multiple-choice items and constructed-response items. Their analysis of interview data, observations, and document analysis surfaced that the case study teachers, while somewhat uneasy with the assessment policy change, were largely in support of the new assessment that asked students to engage in higher-order thinking. In fact, teachers seemed open to and aware of the need to change their instruction to incorporate more scaffolding of learning to support students through these changes.

Meuwissen (2013) conducted an instrumental case study of two teachers (one novice and one experienced secondary social studies teacher) and their assessment practices as they each taught a government curriculum tied to a standardized test alongside a more flexible elective

curriculum. Both case study teachers understood the purpose of classroom-based assessment and wanted to use formative assessments as a feedback loop to improve student learning. However, when faced with the high-stakes context of the state-tested and AP-tested government course they had to make some concessions to their philosophies. Both participants, to varying degrees, engaged in transparency with their students (i.e., openly discussing the processes of, purposes of, and consequences of a given high-stakes assessment), reluctant compliance with testing policies, and pragmatic divergence from state policy in order to create what they deemed was an effective classroom assessment policy. Experience appeared to mediate these teachers' ability to effectively deviate from state policy, corroborating the quantitative survey analysis of Hong and Hamot (2020).

In a separate study, Meuwissen (2017), through a comparative case study, explored the impact of a professional development, *Teachers Doing History (TDH)*, on two teachers enacting this curriculum in a broader accountability context. Findings indicated that state testing policies largely obstructed teachers' ability to enact the aims of the professional development: to facilitate interpretive history teaching with a focus on investigation and inquiry. One case study teacher was more successful in her implementation of *TDH* due to her perceptions of a strong professional community and trust within her school. These three context-specific case studies outlined above provide a basis of empirical evidence that experience, some school- and department-level variables like trust and professional community, and a clearly defined teaching philosophy surrounding assessment may help teachers navigate complex contexts in ways that improve the assessment experiences of their students.

The case study conducted by Fitzpatrick and colleagues (2019) is unique in that it explored both instruction and student learning in a unit on Byzantium in a 9th grade World

History course. Researchers focused on Mr. Smith and four focal students while they experienced a district-mandated performance assessment (DBQ) while still preparing for a state-mandated end-of-course multiple-choice test. Through classroom observations, interviews, and the collection of artifacts researchers found that instruction was largely still centered on historical facts rather than historical interpretation or thinking. Students, in their experience of the DBQ, treated the primary source documents provided as sources of factual information rather than interpretations to be analyzed. This case study is instrumental as it highlights that new, ambitious assessments alone are not sufficient to enhance teaching or student learning toward the goals of thinking like a historian or source analysis. Further, this study highlights the potential power of studies that include both teacher and student data in elucidating more complex learning processes that take place at the classroom level by connecting assessment and instruction.

Discussion

The limited, disparate, and disconnected nature of the research on assessment in secondary social studies make it difficult to "add it up" (Wilson & Anagnostopoulos, 2021) and paint a larger picture of the state of the field. What current research leaves us with is the echoed and unsatisfying call from previous reviews of the literature that more empirical inquiry is needed. In sum, the current state of the field shows promise in the explorations of ambitious assessments at the classroom-level, but very little is known about the process of integrating these assessments into large-scale systems of assessment at the state-level. Research on the nature of high-stakes tests and its pressure on teachers draws largely similar conclusions to those in previous reviews, despite nationally shifting conversations around what is valuable in social studies and what pathways are available to assess student learning. Finally, this review

highlighted several significant gaps in the literature that future research projects should aim to fill.

The research base on projects, inquiries, and performance-based assessment shows potential in how these measures of student learning towards ambitious disciplinary goals may be used at the classroom level. However, the nature of studies of these assessments, mostly case studies at the small scale, means that more work needs to be done to understand how ambitious instruction and assessment might be applied across contexts. The National History Day (2011) study and studies by Parker and colleagues (2011; 2013) also suggest that in the pursuit of more ambitious disciplinary goals students are, indeed, able to attain relevant content knowledge and still perform well on high-stakes multiple-choice measures of knowledge. We know very little about how these kinds of ambitious assessments scale up from the classroom level. Future research should explore new state testing policies that ask districts and schools to employ projects, inquiries, and constructed-response items as a part of their large-scale assessment programs. In addition, more research should explore the feedback loops and formative assessment tasks necessary for student success and growth when engaging in these ambitious assessments, as studies have illuminated the importance of these mechanisms (Monte-Sano, 2009; Levy; 2011)

Large-scale assessments, and the artifacts of old accountability and assessment policy, still have an impact on teachers' decision-making surrounding classroom-based assessment. A teacher's perceived assessment policy context (i.e., the pressures a teacher feels from district and state level policies) and the assessment tools available to them play a key role in that teachers process of gatekeeping classroom learning experiences for students. Future studies should situate the teacher not just as the curricular-instructional gatekeeper (Thornton, 1989; 2006), but also

acknowledge the key role that assessment plays in this gatekeeping process. How a teacher conceptualizes assessment has an impact on student experiences of and preparation for classroom-based and large-scale assessment. Further, a teacher's relationship with classroom-based assessment is the only way that a given teacher can gain information about student learning to inform future instruction. This assessment for learning, rather than assessment of learning, should be recentered in the discourse of assessment in secondary social studies.

The gaps in the current research base suggest a number of potential directions for future inquiry in the landscape of assessment in secondary social studies. First, while a cohort of researchers has worked to generate empirical data on assessments measuring students' historical thinking and their political efficacy, no research projects surfaced in this review have explored measures that might capture what it means for a student to "think like an economist" or "think like a geographer." A number of skills related to these disciplines are outlined in the C3 framework, but no research has explored how to measure student growth towards success in these outcomes. In sum, while empirical research has explored summative and classroom-based measures of student progress towards particular goals, and the ways that summative large-scale assessments impact teacher decision-making, we know very little about formative classroom-based assessment in the secondary social studies classroom. Assessment and instruction, long conceived of as separate endeavors in empirical research, should be considered together in order to endorse and support more constructivist approaches to teaching social studies that align closely with the ambitious disciplinary goals of the field.

References

Abrams, L.M, Pedulla, J.J., & Madaus, G.F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice*, *42*(1), 18-29. https://doi.org/10.1207/s15430421tip4201_4

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., ... Wittrock, M. C. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Objectives (Abridged ed.). Longman.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, 36(5), 258-267. https://doi.org/10.3102/0013189X07306523

Au, W. (2009). Social studies, social justice: W(h)ither the social studies in high-stakes testing?. *Teacher Education Quarterly*, *36*(1), 43-58. https://www.jstor.org/stable/23479200

Au, W. (2010). *Unequal by design: High-stakes testing and the standardization of inequality*. Routledge.

Barton, K. C., & Avery, P. G. (2016). Research on social studies education: Diverse students, settings, and methods. In D. Gitomer & C. Bell (Eds,), *Handbook of research on teaching* (pp. 985-1038). American Educational Research Association.

Blevins, B., LeCompte, K., & Wells, S. (2016). Innovations in Civic Education: Developing Civic Agency Through Action Civics. *Theory & Research in Social Education*, *44*(3), 344–384. https://doi.org/10.1080/00933104.2016.1203853

Bonner, S.M (2013). Validity in classroom assessment: Purposes, properties, and principles. In J.H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 461-472). SAGE.

Breakstone, J. (2013). History assessments of thinking: Design, interpretation, and implementation. [Unpublished doctoral dissertation]. Stanford University.

Breakstone, J. (2014). Try, Try, Try Again: The Process of Designing New History Assessments. *Theory & Research in Social Education*, *42*(4), 453–485. https://doi.org/10.1080/00933104.2014.965860

Breakstone, J., Smith, M., & Wineburg, S. (2013). Beyond the bubble in history/social studies assessments. *Phi Delta Kappan*, *94*(5), 53-57. https://doi.org/10.1177/003172171309400512

Buckles, S., Schug, M. C., & Watts, M. (2001). A national survey of state assessment practices in the social studies. *The Social Studies*, *92*, 141-146. https://doi.org/10.1080/00377990109603992

Cizek, G. J. (2009). Assessment for accountability in education: Past, present, and challenges ahead. *Educational Researcher 38*(6), 467-473. https://doi.org/10.3102/0013189X09344428

Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, *64*(1), 5-31. https://doi.org/10.17763/haer.64.1.j57n353226536276

Darling-Hammond, L., Bae, S., Cook-Harvey, C. M., Lam, L., Mercer, C., Podolsky, A., & Stosich, E. L. (2016). Pathways to new accountability through the Every Student Succeeds Act. *Palo Alto, CA: Learning Policy Institute*.

DeWitt, S. W., Patterson, N., Blankenship, W., Blevins, B., DiCamillo, L., Gerwin, D., Gradwell, J. M., Gunn, J., Maddox, L., Salinas, C., Saye, 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 & Research in Social Education*, *41*(3), 382–427. https://doi.org/10.1080/00933104.2013.787031

Drake Brown, S. (2013). Preparing effective history teachers: The assessment gap. *Journal of Social Studies Research*, *37*(3), 167–177. https://doi.org/10.1016/j.jssr.2013.04.005

Ercikan, K., & Seixas, P. (2015). *New directions in assessing historical thinking*. Routledge.

Evans, R. W. (2004). The social studies wars: What should we teach the children?. Teachers College Press.

Fehn, B. R., & Schul, J. E. (2011). Teaching and Learning Competent Historical Documentary Making: Lessons from National History Day Winners. *The History Teacher*, *45*(1), 25–43.

Fitchett, P. G., Heafner, T. L., & Lambert, R. G. (2017). An Analysis of Predictors of History Content Knowledge: Implications for Policy and Practice. *Education Policy Analysis Archives*, *25*(65). https://doi.org/10.14507/epaa.25.2761

Fitchett, P.G., & VanFossen, P.J. (2013). Survey on the status of social studies: Development and analysis. *Social Studies Research and Practice*, *8*(1), 1-23. https://doi.org/10.1108/SSRP-01-2013-B0001

Fitzpatrick, C., van Hover, S., Cornett, A., & Hicks, D. (2019). A DBQ in a multiple-choice world: A tale of two assessments in a unit on the Byzantine Empire. *Journal of Social Studies Research, 43*(3). https://doi.org/10.1016/j.jssr.2018.09.004

Garcia, G. A., Nunez, A-M., & Sansone, V. A. (2019). Toward a multidimensional conceptual framework for understanding "servingness" in Hispanic-serving institutions: A synthesis of research. Review of Educational Research, 89(5), 745–784. https://doi.org/10.3102/0034654319864591

Gay, G. (2000). Culturally responsive teaching: Theory, research, and practice. Teachers College Press

Gay, G. (2002). Preparing for Culturally Responsive Teaching. Journal of Teacher Education, 53(2), 106–116. https://doi.org/10.1177/0022487102053002003

Girard, B., Harris, L. M., Mayger, L. K., Kessner, T. M., & Reid, S. (2021). "There's no way we can teach all of this": Factors that influence secondary history teachers' content choices. *Theory & Research in Social Education*, *49*(2), 227–261. https://doi.org/10.1080/00933104.2020.1855280

Goodlad, J. (1984). *A place called school: Prospects for the future*. McGraw-Hill.

Grant, S.G. (2017). The problem of knowing what students know: Classroom-based and large-scale assessment in social studies. In M.M. Manfra & C.M. Bolick (Eds.), *The Wiley handbook of social studies research* (1st ed., pp. 461-476). John Wiley & Sons, Inc.

Grant, S.G., & Horn, C. (2006). The state of state-level history tests. In S.G. Grant (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 9-27). Information Age Publishing.

Grant, S.G., & Salinas, C. (2008). Assessment and accountability in the social studies. In L.S. Levstik & C.A. Tyson (Eds.), *Handbook of research in social studies education* (pp. 219–238). Routledge.

Graue, M.E. (1993) Integrating theory and practice through instructional assessment. *Educational Assessment, 1*(4), 283-309. https://doi.org/10.1207/s15326977ea0104_1

Haladyna, T.M. (1999). *Developing and validating multiple-choice test items* (2nd ed.). Erlbaum.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Erlbaum.

Hawley, T. S., & Whitman, G. M. (2020). Fear and learning in student teaching: Accountability as gatekeeper in social studies. *The Journal of Social Studies Research*, *44*(1), 105–115. https://doi.org/10.1016/j.jssr.2019.04.003

Heafner, T.L & Fitchett, P.G. (2015). An opportunity to learn US history: What NAEP data suggest regarding the opportunity gap. *The High School Journal 98*(3), 226-249. https://doi.org/10.1353/hsj.2015.0006

Heafner, T. L., & Fitchett, P. G. (2017). US history content knowledge and associated effects of race, gender, wealth, and urbanity: Item Response Theory (IRT) modeling of NAEP-USH achievement. *The Journal of Social Studies Research*, *42*(1), 11–25. https://doi.org/10.1016/j.jssr.2017.01.001

Heafner, T. L., VanFossen, P. J., & Fitchett, P. G. (2019). Predictors of students′ achievement on NAEP-Economics: A multilevel model. *The Journal of Social Studies Research*, *43*(4), 327–341. https://doi.org/10.1016/j.jssr.2019.01.003

Hertzberg, H.W. (1988). Foundations, the 1892 committee of ten. *Social Education, 52*(2), 144-145.

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.

Hong, H., & Hamot, G. E. (2020). Differential effects of state testing policies and school characteristics on social studies educators' gate-keeping autonomy: A multilevel model. *Theory & Research in Social Education*, *48*(1), 74–100. https://doi.org/10.1080/00933104.2019.1655508

Huijgen, T., van Boxtel, C., van de Grift, W., & Holthuis, P. (2017). Toward Historical Perspective Taking: Students' Reasoning When Contextualizing the Actions of People in the Past. *Theory & Research in Social Education*, *45*(1), 110–144. https://doi.org/10.1080/00933104.2016.1208597

Journell, W. (2010). The influence of high-stakes testing on high school teachers' willingness to incorporate current political events into the curriculum. *The High School Journal, 93*(3), 11-125. https://www.jstor.org/stable/40864929

Kohlmeier, J., & Saye, J. (2019). Examining the relationship between teachers' discussion facilitation and their students' reasoning. *Theory & Research in Social Education*, *47*(2), 176–204. https://doi.org/10.1080/00933104.2018.1486765

LeCompte, K., Blevins, B., & Riggers-Piehl, T. (2020). Developing civic competence through action civics: A longitudinal look at the data. *The Journal of Social Studies Research*, *44*(1), 127–137. https://doi.org/10.1016/j.jssr.2019.03.002

Levy, B. L. M. (2011). Fostering Cautious Political Efficacy Through Civic Advocacy Projects: A Mixed Methods Case Study of an Innovative High School Class. *Theory & Research in Social Education*, *39*(2), 238–277. https://doi.org/10.1080/00933104.2011.10473454

Lin, A. R., Lawrence, J. F., Snow, C. E., & Taylor, K. S. (2016). Assessing Adolescents' Communicative Self-Efficacy to Discuss Controversial Issues: Findings From a Randomized Study of the Word Generation Program. *Theory & Research in Social Education*, *44*(3), 316–343. https://doi.org/10.1080/00933104.2016.1203852

Lin, A. R., Lawrence, J. F., & Snow, C. E. (2015). Teaching urban youth about controversial issues: Pathways to becoming active and informed citizens. Citizenship, Social and Economics Education, 14, 103–119. https://doi.org/10.1177/2047173415600606

Martin, D.M., Maldonado, S.I., Schneider, J., & M., Smith (2011). *A report on the state of history education: State policies and national programs.* https://doi.org/10.13140/RG.2.2.12801.40802

McGrew, S., Breakstone, J., Ortega, T., Smith, M., & Wineburg, S. (2018). Can Students Evaluate Online Sources? Learning From Assessments of Civic Online Reasoning. *Theory & Research in Social Education*, *46*(2), 165–193. https://doi.org/10.1080/00933104.2017.1416320

Meuwissen, K.W. (2013). Readin', writin', ready for testin'? Adaptive assessment in elective and standardized-tested social studies course contexts. *Theory and Research in Social Education*, *41*(3), 285–315. https://doi.org/10.1080/00933104.2013.812049

Meuwissen, K. W. (2017). "Happy Professional Development at an Unhappy Time": Learning to Teach for Historical Thinking in a High-Pressure Accountability Context. *Theory & Research in Social Education*, *45*(2), 248–285. https://doi.org/10.1080/00933104.2016.1232208

Miller, J. M. (2018). U.S. history state assessments, discourse demands, and English learners' achievement: Evidence for the importance of reading and writing instruction in U.S. history for English Learners. *The Journal of Social Studies Research*, *42*(4), 375–392. https://doi.org/10.1016/j.jssr.2017.12.001

Monte-Sano, C. (2008). Qualities of historical writing instruction: A comparative case study of two teachers' practices. *American Educational Research Journal*, *45*(4), 1045-1079. https://doi.org/10.3102/0002831208319733

Monte-Sano, C. (2010). Disciplinary Literacy in History: An Exploration of the Historical Nature of Adolescents' Writing. *Journal of the Learning Sciences*, *19*(4), 539–568. https://doi.org/10.1080/10508406.2010.481014

Mueller, R. G. W., & Colley, L. M. (2015). An evaluation of the impact of end-of-course exams and ACT-QualityCore on U.S. history instruction in a Kentucky high school. *Journal of Social Studies Research*, *39*(2), 95–106. https://doi.org/10.1016/j.jssr.2014.07.002

National Council for the Social Studies (1999). Authentic assessment in social studies [Special Issue]. *Social Education, 63*(6).

National Council for the Social Studies. (2013). The college, career, and civic life (C3) framework for social studies state standards: Guidance for enhancing the rigor of K-12 civics, economics, geography, and history. Silver Spring, MD

National History Day. (2011). *National history day works: National program evaluation.* Retrieved from https://www.nhd.org/why-nhd-works

Parker, W.C., & Lo, J.C. (2016). Reinventing the high school government course: Rigor, simulations, and learning from text. *Democracy and Education, 24*(1). https://democracyeducationjournal.org/home/vol24/iss1/6

Parker, W. C., Lo, J., Yeo, A. J., Valencia, S. W., Nguyen, D., Abbott, R. D., Nolen, S.B., Bransford, J.D., & Vye, N. J. (2013). Beyond breadth-speed-test: Toward deeper knowing and engagement in an Advanced Placement course. *American Educational Research Journal*, *50*(6), 1424-1459. https://doi.org/10.3102/0002831213504237

Parker, W., Mosborg, S., Bransford, J., Vye, N., Wilkerson, J., & Abbott, R. (2011). Rethinking advanced high school coursework: Tackling the depth/breadth tension in the AP US

government and politics course. *Journal of Curriculum Studies*, *43*(4), 533-559. https://doi.org/10.1080/00220272.2011.584561

Patterson, T., & Shuttleworth, J. M. (2019). The (mis) representation of enslavement in historical literature for elementary students. Teachers College Record, 121(4), 1-40. https://doi.org/10.1177/016146811912100403

Pedulla, J., Abrams, L., Madaus, G., Russell, M., Ramos, M., & Miao, J. (2003). *Perceived effects of state mandated testing programs on teaching and learning: Findings from a national survey of teachers.* National Board on Educational Testing and Public Policy.

Rantala, J., Manninen, M., & van den Berg, M. (2016). Stepping into other people's shoes proves to be a difficult task for high school students: Assessing historical empathy through simulation exercise. *Journal of Curriculum Studies*, *48*(3), 323–345. https://doi.org/10.1080/00220272.2015.1122092

Reich, G. A. (2009). Testing historical knowledge: Standards, multiple-choice questions and student reasoning. *Theory & Research in Social Education*, *37*(3), 325-360. https://doi.org/10.1080/00933104.2009.10473401

Reich, G. A. (2013). Imperfect models, imperfect conclusions: An exploratory study of multiple-choice tests and historical knowledge. *The Journal of Social Studies Research, 37*(1), 3-16. https://doi.org/10.1016/j.jssr.2012.12.004

Reisman, A. (2012). Reading Like a Historian: A Document-Based History Curriculum Intervention in Urban High Schools. *Cognition and Instruction*, *30*(1), 86–112. https://doi.org/10.1080/07370008.2011.634081

Salinas, C. (2006). Teaching in a high-stakes testing setting: What becomes of teacher knowledge? In S. G. Grant (Ed.), Measuring history: Cases of high-stakes testing across the U.S. (pp. 177–194). Information Age Publishing.

Saye, J. & Social Studies Inquiry Research Col. (2013). Authentic Pedagogy: Its Presence in Social Studies Classrooms and Relationship to Student Performance on State-Mandated Tests. *Theory & Research in Social Education*, *41*(1), 89–132. https://doi.org/10.1080/00933104.2013.756785

Shemilt, D. (2018). Assessment of learning in history education: Past, present, and possible futures. In S.A. Metzger & L.M. Harris (Eds.), *The Wiley international handbook of history teaching and learning* (1st ed., pp. 449-471). John Wiley & Sons, Inc.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14. https://doi.org/10.3102/0013189X029007004

Shepard, L.A. (2006). Classroom assessment. In R.L. Brennan (Ed.) *Educational measurement* (4th ed., pp. 623-646). American Council on Education.

Smith, M. (2017). Cognitive validity: Can multiple-choice items tap historical thinking processes? *American Educational Research Journal. 54*(6), 1256-1287. https://doi.org/10.3102/0002831217717949

Smith, M. D. (2018). New Multiple-Choice Measures of Historical Thinking: An Investigation of Cognitive Validity. *Theory & Research in Social Education*, *46*(1), 1–34. https://doi.org/10.1080/00933104.2017.1351412

Smith, M., Breakstone, J. & Wineburg, S. (2019). History assessments of thinking: A validity study. *Cognition and Instruction*. *37*(1), 118-144. https://doi.org/10.1080/07370008.2018.1499646

Swan, K., & Hofer, M. (2013). Examining Student-Created Documentaries as a Mechanism for Engaging Students in Authentic Intellectual Work. *Theory & Research in Social Education*, *41*(1), 133–175. https://doi.org/10.1080/00933104.2013.758018

Thornton, S.J. (1989, March 27-31). *Aspiration and practice. Teacher as curricular-instructional gatekeeper in social studies* [Paper Presentation]. AERA 1989 Conference, San Francisco, CA. https://files.eric.ed.gov/fulltext/ED315347.pdf

Thornton, S.J. (2006). What matters most for gatekeeping? A response to VanSledright. *Theory and Research in Social Education, 34*(4), 416-418. https://doi.org/10.1080/00933104.2006.10473316

Torrez, C.A., & Claunch-Lebsack, E.A. (2013). Research on assessment in the social studies classroom. In J.H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 461-472). SAGE.

van Hover, S. D. (2006). Teaching history in the old dominion: The impact of Virginia's accountability reform on seven secondary beginning history teachers. In S. G. Grant (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 195–219). Information Age Publishing.

van Hover, S. D., & Heinecke, W. (2005). The impact of accountability reform on the "'wise practice" of secondary history teachers: The Virginia experience. In E. A. Yeager & O. L. Davis Jr. (Eds.), *Wise social studies teaching in an age of high-stakes testing* (pp. 89–105). Information Age Publishing.

Volger, K. (2006). The impact of high school graduation examination on Mississippi social studies teachers' instructional practices. In S.G. Grant (Ed.), *Measuring history: Cases of state-level testing across the United States* (pp. 273-302). Information Age Publishing.

Wilson, S.M., & Anagnostopoulos, D. (2021) Methodological guidance paper: The craft of conducting a qualitative review. *Review of Educational Research, 91*(5), 651-670. https://doi.org/10.3102/00346543211012755

Wright, G. P., & Endacott, J. L. (2016). Historical inquiry and the limitations of the common core state standards. *Journal of Social Studies Research*, *40*(4), 309–324. https://doi.org/10.1016/j.jssr.2015.07.003

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of educational measurement*, *26*(1), 55-66. https://doi.org/10.1111/j.1745-3984.1989.tb00318.x

Appendix A

Articles Included in Review

Ambitious Classroom-Based Assessment	
Projects	Blevins et al., 2016; Fehn & Schul, 2011; LeCompte et al., 2020; Levy, 2011; National History Day [NHD], 2011; Parker & Lo, 2016; Parker et al., 2011; Parker et al., 2013; Swan & Hofer, 2013
Discussion	Kohlmeier & Saye, 2019; Lin et al., 2016
Constructed-Response Items	Breakstone, 2013; Breakstone, 2014; Breakstone et al., 2013; McGrew et al., 2018; Monte-Sano, 2008; Monte-Sano, 2010; Smith, 2017; Smith, 2018; Smith et al., 2019; Wright & Endacott, 2016
Historical Perspective Taking	Huijgen et al., 2017; Rantala et al., 2016
Multiple Choice Testing	
Multiple-Choice Measures	Reich 2009; Reich, 2013; Reisman, 2012; Smith, 2017; Smith, 2018
Large-Scale Analysis of NAEP Results	Fitchett et al., 2017; Heafner and Fitchett, 2015; Heafner and Fitchett, 2017; Heafner et al., 2019; Patterson & Shuttleworth, 2020
State Standards and State Tests	Dewitt et al., 2013; Saye & SSIRC, 2013
Teaching in Large-Scale Contexts	Drake Brown, 2013; Fitchett & VanFossen, 2013; Girard et al., 2021; Hawley & Whitman, 2020; Hong & Hamot, 2015; Hong and Hamot, 2020
Navigating and Sensemaking New Measures and New Contexts	
New Measures and New Contexts	Fitzpatrick et al., 2019; Miller, 2018; Mueller & Colley, 2015; Meuwissen, 2013; Meuwissen, 2017;