Article

Assessing deeper learning of high school civics

Sheila W. Valencia[a], **Walter C. Parker**[b], **Jane C. Lo**[c]

[a]University of Washington, Seattle, USA [b]University of Washington, Seattle, USA, [c]Michigan State University, East Lansing, USA

– Deeper learning in civics is conceptually rich and facilitates learning in the future.

– We developed and conducted research on an assessment model and a test of deeper learning in high school civics.

– We used construct-driven assessment design to develop the assessment.

– We conducted research on the assessment using Design-Based Implementation Research across 13 schools.

– Core concepts and reasoning strategies for the course provided the framework for assessment alignment.

Purpose: Civic education is a central mission of public schools, and *deeper* learning of civics—learning that is complex and adaptive—is the goal. However, *assessment* of deeper civic learning is limited. Therefore, we aimed to develop an assessment model and test of deeper learning in the common high school civics course taught across the U.S.

Design/methodology/approach: Using Design-Based Implementation Research (DBIR), the assessment model and test were iteratively researched and revised by a team of researchers and teachers across seven years and multiple settings.

Findings: Results of validity and reliability studies show that the model and test are promising tools for assessing deeper civic learning.

Research limitations/implications: Additional research is warranted to refine the test-development process, design alternative test forms, and adapt the model to other social studies courses.

Practical implications: We suggest ways to use this assessment model to assess learning in civics and other social studies subjects.

Corresponding author:
Sheila W. Valencia, University of Washington, 2012 Skagit Lane, Seattle, Washington 98195-3600, USA. E-Mail: valencia@uw.edu

Declaration of conflicts of interests: No potential conflict of interest was reported by the authors.

1   INTRODUCTION

Initiatives to renew social studies curriculum and instruction often embrace the goal of learning that is deep, engaging, disciplined, and connected to life outside school. In the U.S., the inquiry arc of the *College, Career, and Civic Life* framework (National Council for the Social Studies, 2013) is an exemplar as was Newmann's project (1996) and numerous earlier projects of the New Social Studies (e.g., Oliver & Shaver, 1967). A similar emphasis can be found in Europe (EuroClio, n.d.; Gessner, 2017; Reinhardt, 2016). Assessing this kind of learning, however, has received only sporadic attention, and this is the gap we address in this article.

We present a model for assessing deeper learning that asks students to demonstrate their understanding in the context of a meaningful task and to explain the reasoning they used to do so. The model was developed for the ubiquitous high school U.S. government course, but a version of the course that had been designed by our team to achieve deeper learning (Parker, Valencia, & Lo., 2018). The course generally remains a high-enrollment site for education about constitutional democracy—an urgent subject today. An increasing number of states in the U.S. are making laws requiring students to take a civics course and/or pass a civics test (CivXNow, n. d.; Hansen et al., 2018). Successful completion of the U.S. Citizenship and Immigration Services' (USCIS*)* citizenship test, for example, is now required for high school graduation in over half the states. The growing popularity of this particular test, which only skims the surface of civic knowledge and does not touch on reasoning, demonstrates the need for tests that go deeper into the subject.

We proceed in an unusual way for a research article. Because assessment does not stand alone but is tethered to curriculum, instruction, and learning theory, we begin with a brief description of the conceptual (Section 2) and practical (Section 3) contexts in which this assessment was developed. This provides an overview of the redesigned high school civics curriculum on which the assessment was based. (A full description of that curriculum and its implementation can be found in Parker & Lo, 2016.) After this, in Section 4, we describe methods used to develop the test *model* and the *measure* itself, called the Complex Scenario Test (CST), which is the focus of this research. This measure, and the model it exemplifies, lead the Findings Section (5), as they were the result of seven years of iterative research. There, we report the research and related decision making surrounding our goal of building a valid and reliable assessment tool for deeper learning. In Section 6, Discussion, we reflect on the challenges and the lessons learned as we developed the CST with teachers of the course. We believe the process and the model will be useful to assess deeper learning in other social studies/social science courses as well, a point to which we return directly at the end.

2   DEEPER LEARNING AND ITS ASSESSMENT

Deeper learning, usually contrasted with superficial or rote learning, is promoted across numerous initiatives. It was a somewhat novel idea with Resnick and Klopfer's publication

of *Toward the Thinking Curriculum* in 1989 but then took hold and has become an expectation in educational scholarship, curriculum standards, and assessment of student learning (e.g. Hewlett Foundation's *Deeper Learning initiative* [2013]; National Research Council's *Education for Life and Work: Developing Transferable Knowledge and Skills in 21st Century Skills* [2012]; OECD [2018]). Although time-on-topic has been the unit of analysis in some research on deeper learning (Schwartz et al., 2008), we concentrate our work on the *kind of understanding* that obtains and how to assess that understanding.

We characterize deeper learning as complex and adaptive (Bransford & Schwartz, 2000). These are related attributes. To say that an understanding is complex means that it is nuanced and differentiated or shaded and varied. To say that it is adaptive means that it can be applied flexibly, bending as needed to novel problems. In this way, a deep understanding is a conceptual model that facilitates learning and action in the future when conditions have changed. Hatano and Inagaki (1986) call this "adaptive expertise" (p. 266).

We narrow now to deeper *civic* learning, the curricular aim of a redesigned government course and the target of our assessment. By deeper civic learning, we mean complex and adaptive knowledge of the concepts, structures, and functions of national and local government and politics, as well as the rights and responsibilities of citizens. Although more ambiguous than some school subjects (Davies et al., 2017), civic education is similar to other subjects in its specialization and boundedness. Just as a biology course is not a chemistry course, a civics course is not a geography course. It has its disciplinary subject matter—its focal concepts, issues, and reasoning strategies. Civic education is unique among school subjects, however, because it is so thoroughly situated in the national political context. An assessment of civic knowledge has to be pertinent to the discipline while also pertinent to the political context (Galston,1989; Löfström & Grammes, 2020).

New approaches to assessment began to take hold in the 1990s when shifts in subject-matter standards placed a priority on complex thinking about subject matter—the aforementioned "thinking curriculum." Such assessments, aimed to "capture not just the 'right answer,' but also the reasonableness of the procedure used to carry out the task" (Shavelson et al., 1992, p. 22). Students were asked to demonstrate knowledge of what they had learned, and apply it to a task, rather than simply recognize and select answers from a given list (Baker, 1998; Linn et al., 1991). Among the formats advanced for these tests of deeper learning—sometimes referred to as performance assessments—were projects and problem-solving tasks, open-ended responses such as document-based/evidence-based arguments, portfolio assessments (collections of student work over time), and even collaborative work.

At the time of our project, there had been a good deal of research on effective practices associated with classroom-based *formative* assessment of deeper learning, much of it in mathematics, language arts, and science (Black & Wiliam,1998; Heritage et al., 2009). The focus on formative assessment makes sense because many classroom assessments of

deeper learning involve time-intensive tasks and projects that are integrated with instruction. During teaching activities, teachers observe students and provide feedback based on what they see and, in this way, instruction and formative performance assessment are folded into one another so seamlessly that instruction and assessment can be difficult to distinguish. In fact, we had designed the government course curriculum to include many opportunities for formative assessment of student learning.

Much less research, however, has been conducted on classroom-based *summative* assessments of deeper learning. Most of the research on summative assessment of deeper learning has been conducted in the context of large-scale, high-stakes assessments that are used for accountability outside the classroom (e.g., Niemi et al., 2007). As with formative assessments of deeper learning, most efforts have occurred in language arts, mathematics, and science, although several innovative projects have targeted assessing historical thinking (Ercikan & Seixas, 2015). Still, many classroom teachers, especially high school teachers, depend on classroom-based summative assessments as an important indicator of student learning at the end of the course. Teachers who are focused on deeper learning are no exception.

This was certainly the case in our work. Our collaborating teachers needed an assessment of deeper learning that could be administered in one class period and easily scored at the end of the high school government course. The purpose was threefold: a) to assess individual students' learning as contrasted to collaborative projects in which group learning might be assessed; b) to assess students' long-term learning across the course; and c) to provide evidence of student learning that would inform revisions to the course as it was implemented across seven years and additional sites. Working alongside the teachers, we designed and tested a model for a classroom-based, summative assessment of deeper learning that was tried out, scored, analyzed, and revised over time.

Our work on a low-stakes, summative assessment of deeper learning fell at the intersection of existing research on large-scale performance assessment (typically summative and high stakes) and classroom-based performance assessment (typically formative and low stakes). Accordingly, we looked to both fields of research for insights. Following Miller and Linn (2000), we were most attuned to issues of validity (content, construct, consequential) and scoring (both reliability and practical efficiency).

3   PRACTICAL CONTEXTS

3.1   The high school government course

The research reported here focuses on assessing student learning in a durable site of civic education in the United States: the high school government course. Nearly all students in the U.S. take it in one form or another (Hansen et al., 2018). At the urging of our collaborating teachers, we situated our work in the Advanced Placement version of this course (hereafter APGOV). It had a well-respected curriculum already developed for the

College Board (2018, the association that creates course frameworks and tests) by a sample of professors who teach the course in colleges and universities, and a high-stakes summative test of breadth, but not depth. It is a platform where rapid test-prep instruction predominates—an approach that could be called "breadth-speed-test" (Parker et al., 2013). Teachers are committed to curriculum coverage so their students have a fair chance of earning a high score on the AP test and, potentially, earning college credit.[1] Working with APGOV teachers afforded us the opportunity to design and investigate an alternative, summative assessment while they were implementing an engaging, experiential mode of instruction, and an AP curriculum reorganized for deeper learning—all of this on a platform where one would least expect to find it.

3.2 Design for deeper learning

We provide a brief overview of the curriculum and instructional framework of the civics course for which the assessment was designed. We do so because aligning assessment with the concepts and strategies taught in a course is an essential step to examining the validity of any assessment results. We aimed to develop an instructional framework that could deepen student learning of the APGOV curriculum while engaging students in lively learning activities. Working with teachers of the course, we developed a pedagogical model on four design principles. First, projects are the spine of the course. Students learn the content through participation in five weeks-long political simulations, each essential to understanding the workings of U.S. government and politics (e.g., arguing a Supreme Court case), playing various key roles, and working collaboratively. Second, "engagement first." Students are placed in roles immediately, engaging them *before* they encounter the content knowledge and strategies they acquire through the simulation. This creates a need-to-know (Schwartz & Bransford, 1998). Third, deeper learning is further facilitated as students are provided with multiple opportunities to re-engage with core concepts and strategies across the five projects—a strategy we call "looping." Students try out their fledgling understandings in iterative cycles, building prior knowledge for later encounters, which may need to be remodeled to correct misconceptions and address new material, resulting in a more complex and more applicable understandings.

Figure 1. APGOV+ core disciplinary concepts and reasoning strategies

Core Disciplinary Concepts

1. Limited government

2. Separation of powers (Federalism, three branches, checks and balances)

3. Constitutionalism (rule of law, precedent)

4. Civil rights and liberties

5. Institutions linking citizens to government (elections, interest groups, political parties, media)

Core Disciplinary Reasoning Strategies

1. Constitutional reasoning (reasoning about policy on the basis of the Constitution)

2. Deliberation (discussion to decide among alternatives; evidentiary warrant)

3. Perspective taking (e.g. trying on diverse political ideologies and social positions)

4. Political autonomy (making uncoerced decisions, consenting to be governed, voting for candidate X)

5. Close, interpretive reading of subject-specific texts

Fourth, while looping through multiple learning cycles addresses the *how* of deeper learning, deliberative curriculum decision making drives curricular decisions about *which* content and strategies are worthy of looping (see Parker & Lo, 2016; Davies et al., 2017). Working from the existing APGOV curriculum and deliberating with APGOV teachers across multiple meetings, we settled on five disciplinary concepts and five disciplinary reasoning strategies (see Figure 1). These were looped throughout the revised course to build deeper understanding, facilitate transfer, and encourage coherent instruction. These core concepts and reasoning strategies became the deeper-learning targets of the revised course, hereafter APGOV+, and consequently, of the assessment we aimed to develop. Hence, they are the key to the whole enterprise of teaching, learning, and assessment.

4 METHOD: ITERATIVE RESEARCH AND DEVELOPMENT

Both the curriculum for the APGOV+ course and the CST assessment were designed and implemented over seven years using Design-Based Implementation Research (DBIR) methodology (Fishman et al., 2013). DBIR emphasizes iterative cycles of collaborative design and testing to achieve a sustainable innovation in a practice field, and it seeks further to scale up the innovation to some extent, which requires attention to ecological factors. Consequently, we were continually in conversation with our teacher partners, and we observed in classrooms several times each semester as they adjusted instruction and provided feedback to us on both the course and the CST design. We also conducted focus groups and individual interviews with students. The DBIR context is important for two reasons: 1) Teachers were partners in the ongoing design and implementation of both the curriculum and assessment. This ensured there was alignment between the APGOV+ course and the end-of-course assessment as each of them morphed and matured. 2) The participating teachers and schools changed over time, which resulted in iterative versions of the CST that were investigated with different cohorts of teachers and students.

4.1 Research sites

We began this work in collaboration with a well-resourced suburban school district, and then we supported its migration to three poverty-impacted urban school districts. Across

the seven years of the project, we conducted research in more than 50 classrooms across 13 schools in three states. Teachers adapted the curriculum to fit their particular district and class situations while staying true to the four design principles and the identified core concepts and reasoning strategies shown in Figure 1. Although the curriculum included performance tasks embedded in the curriculum, which provided opportunities for formative assessment of deeper learning, these tasks were typically collaborative projects carried out over several days or weeks. As noted earlier, our collaborating teachers needed a summative assessment that could be administered in one class period at the end of the course and then scored efficiently and reliably. So, our research on the CST took place in these classrooms as well. CST scores would not be used to make high-stakes decisions about students or recommendations for college credit; instead, CST scores would complement the formative assessment opportunities during classwork *and* scores provided by the summative, breadth-oriented AP test. Together, these would help teachers obtain a fuller picture of student learning, including the depth of their learning, while informing their instruction.

4.2   Participants

Our DBIR collaboration included an ever-changing group of teachers, and schools. Overall, more than 1,200 students from primarily high-poverty schools[2] took versions of the CST. Our 13 teacher collaborators had a wide range of teaching experience, ranging from two years to more than 15. Some stayed with the project for more than five years, others for only a year. Some taught one section of the course, others more. Some knew teaching methods (e.g., seminar, groupwork), others did not. We did not personally select teachers. They were mostly volunteers, although some were encouraged by department heads. All of the teachers, however, participated in at least two days of professional development on the curriculum and the design principles. The final version of the CST, presented in the Findings section below, was administered to 385 students across 12 schools in four districts that had completed the APGOV+ course. Demographic information indicated that the sample was diverse ethnically, linguistically, and socioeconomically (35% Caucasian, 23% Hispanic, 17% African American, 17% Asian, 8% other; 44% home language other than English; 36% eligible for free or reduced-price lunch).

4.3   Assessment development procedure

We drew on features of principled or construct-driven assessment design to develop and test our summative measure of deeper learning (Messick, 1994; Pellegrino, et al., 2001). This approach begins by identifying the essential subject-matter concepts or content standards to be assessed as well as the cognitive demands of the learning process. Further, it specifies the assessment tasks that will elicit the targeted knowledge and thinking as well as the scoring approaches that will be used to evaluate student performance. This three-pronged approach to assessment design mitigates some of the challenges that have

plagued performance assessments, chief among which are the labor-intensiveness of designing the assessments and the variability across assessment tasks and scoring rubrics. Principled assessment design addresses these concerns by creating a conceptual blueprint that supports the design of prototypes or "shells" (Solano-Flores & Shavelson, 1997, p. 18) that can be applied to various courses and content topics.

We developed a shell that was specific to deeper learning in the domain of the government course; however, we maintained an interest in applicability to other social studies courses. Our approach began with reviewing the core disciplinary concepts and reasoning strategies (see Figure 1), which the assessment would measure, and then constructing assessment tasks that required students to engage in applications of those concepts and strategies. As with other summative assessments, the CST was designed to capture a representative sample from the universe of course objectives (Shavelson et al., 1993). We chose not to include two disciplinary reasoning strategies—*deliberation and political autonomy*—because they were discussion-based and better assessed during classroom activities rather than in this summative classroom assessment. To make the CST meaningful and authentic (Messick, 1994; Newman & Associates, 1996), we designed its tasks around contemporary issues that are relevant to high school students, both in school and out.

5   FINDINGS

After several iterations, we arrived at a shell and a test that could be implemented and scored reliably. We begin this section with a description of this assessment shell as well as the tasks and questions that make up the final test. Then, we describe four findings related to validity and reliability: (1) reading and writing difficulty, (2) tasks and questions, (3) test content, and (4) scoring procedure. We believe educators developing measures of deeper learning in other social studies courses, such as national history or economics, will find helpful both our process of assessment development and the shell and test that resulted.

5.1   CST shell and final test

The CST was designed to be administered in one class period toward the end of the APGOV+ course. In keeping with our definition of deeper learning, it requires students to apply content knowledge (factual and conceptual) and disciplinary reasoning strategies learned in the course (see Figure 1) to novel problems (scenarios) drawn from real-life situations that are similar to, but not the same as, subject matter students have studied in the course. A chief goal was to integrate assessment of core course content and reasoning strategies into each scenario.

5.1.1   CST shell

The test begins with an overview of the assessment tasks read aloud by the teacher. It introduces students to a contemporary political problem such as an immigrant fighting extradition, a GPS tracking device used to find drug dealers, or a high school locker search (targeting the core concepts *constitutionalism, civil rights and liberties, limited government*). Then students are put in the role of an advisor to a client: an interest group with a particular policy agenda on this problem (*perspective taking*). This concept (*interest group*), plus how an interest group differs from a *political party* (*linkage institutions*) and how both of these differ from the *three branches of government* (*separation of powers*), are core course content. Students are asked to give knowledgeable and well-reasoned advice (*constitutional reasoning*) to the interest group on how to advance its policy agenda, and the advice they give is to be based on information they are given in the test itself and concepts and reasoning strategies learned across the course. Students must set aside personal opinions and, instead, mobilize concepts, facts, and disciplinary reasoning relevant to the scenario and to the perspective and aims of their particular client. Information about the problem at hand is presented in the form of an actual newspaper or magazine article, abridged (*close interpretive reading*).

After the teacher reads the overview and directions aloud, students begin the assessment on their own. They read a short framing paragraph that orients them to the problem and their role as advisors and then introduces them to the first reading containing new information (document #1). Then, students answer a number of short constructed-response questions prompting them to apply content taught in the course, and then they write an extended response providing advice to the interest group.

Next, students are introduced to a "critical incident" in which they are given additional, often conflicting, information on the same scenario. This is provided, again, in the form of an abridged newspaper or magazine article (document #2). Students are asked now to consider the new case and to revise their original advice as needed. In this way, the CST elicits two samples of students' adaptive transfer of disciplinary concepts and reasoning. The first is their initial advice to the interest group based on the first source document; the second is their response to additional information provided in the second document.

5.1.2   Final iteration of the test

This task shell can be used with a range of scenarios, interest groups, and controversial policy issues that cross levels of government (national, state, local) as well as branches (legislative, executive, judicial). The final iteration of the CST (Form A), based on this shell, is presented in the Appendix and features a controversy related to the Constitution's protection against "unreasonable searches and seizures."

To test the generalizability of the shell, we created another form of the CST (Form B). The core concepts and strategies remained the same as in the original Form A. However, Form B presented students with a different scenario (a controversy related to the

Constitution's guarantee of a "speedy and public trial") and, consequently, different documents that, again, were related to, but not the same as, those studied in the course or contained in Form A. Using a systematic matching method within classrooms, the forms were administered to a total of 385 students from ten schools in three districts that had implemented the APGOV+ curriculum. There was no significant difference between scores on these two forms. This extra step, creating an additional form of the CST, was reassuring. We were able to demonstrate that the test shell could be applied to a different scenario, that our process of selecting and revising documents was robust, and that students' learning of core concepts and strategies as measured by their CST scores was comparable across forms.

While these scenarios use topics related to the Bill of Rights (the first ten amendments of the U.S. Constitution), other scenarios can be developed using the same shell. For example, students advising an interest group wanting to abolish the Electoral College would need not only to understand the constitutional basis of the Electoral College but also the reasons for its continued popularity. Still another scenario could test students' understanding of elections to Congress. Here, the interest group seeking advice wants to stop voter suppression laws and gerrymandering of districts in a particular state. Students advising this group would need to understand how elections function in the United States, the role of the fifty state governments in determining voting rules (*federalism*), and how different types of elections (winner take all, proportional voting) influence election outcomes. As these examples demonstrate, this task shell has sufficient versatility to accommodate the core disciplinary concepts and reasoning strategies of this course and, we believe, can be adapted to other social studies courses as well. While the scenarios and associated examples differ from one scenario to another, the shell consistently attends to students' understanding of the specified core concepts and strategies of the course. It also features contemporary topics and interest groups, which lends authenticity to the task.

5.2   Investigations of validity and reliability

5.2.1   Working with reading and writing difficulty

Many assessments of deeper learning include documents that students must read in order to respond to the assessment tasks. This is one way to assess application or transfer of content and reasoning strategies learned in a course; introducing new information provides a context for students to use existing knowledge to analyze the new and to demonstrate their thinking. The challenge for assessment development, however, is that reading and writing ability can confound assessment of the targeted outcomes and become a source of construct-irrelevant variance (Haladyna & Downing, 2004; Miller & Linn, 2000).

We were confronted with reading and writing challenges early in the development of the CST as we examined existing standardized reading and writing scores for students in

our samples. Students demonstrated an exceptionally wide range of reading abilities on a standardized test (5th - 98th percentile) with approximately 40% of students scoring in the lowest quartile. These data suggested that reading difficulty, and perhaps writing difficulty, on the CST would need to be carefully considered so that it would not threaten the validity of students' performance on the test.

We examined all the text contained in the CST (written directions, framing language, documents, and constructed-response questions) and applied findings from research on text difficulty and complexity (Valencia et al., 2014) to identify potential challenges. These included text length, density of concepts, uncommon figures of speech (e.g., "short end of the stick"), challenging vocabulary unrelated to course content (e.g., "pervasive"), academic language (e.g., write a "memo"), clarity of directions, test format, and other factors that might impact readability and, consequently, contribute to construct-irrelevant variability. It is important to note that many of these factors that influence difficulty are not included in automated reading difficulty estimators such as Lexiles® (Stenner et al., 2006) or Text Evaluator (Sheehan et al., 2014) that are often applied to texts included on disciplinary content tests. All of these text-based elements of the test needed careful review and possible revision.

We also conducted cognitive think-alouds (Pressley & Afflerbach, 1995) with a stratified sample of students (i.e., students scoring across all quartiles on the PSAT reading test) to assess students' abilities to both read and comprehend the revised CST. Analysis indicated that students were highly capable of decoding the texts and the test questions—that is, they were able to correctly read the words with a high degree of accuracy (>95%). However, the majority of students demonstrated difficulty comprehending some aspects of the texts. Taken together, these findings led to further revisions of the CST to reduce the reading load of the test documents and to require teachers to read directions aloud. Over the next two years, the reading demands of the CST were reviewed and adjusted to reduce construct-irrelevant variance and address fairness in assessment of the targeted learning outcomes.

5.2.2   Working with tasks and questions

Moving from the targets of assessment (deeper understanding of the core curricular objectives in Figure 1) to specific assessment tasks and questions (the CST) is particularly difficult when designing assessments of deeper learning. The goal here is to create a test that best elicits students' learning rather than obscures it.

Accordingly, we conducted qualitative analyses of student responses to questions on the CST. We wanted to determine if students were understanding the intent of the questions. Analysis of a subset of papers revealed that students' understandings of many course concepts were reasonably complex, but students were unable to elaborate their reasoning. They provided only vague explanations and few details in their written responses. This suggested that the structure and language of the questions and tasks on

early versions of the CST might be obscuring more valid information about students' understandings. Heeding Messick's (1994) caution about the need for balance between complex performance tasks and structured exercises, the CST was revised to add more structure, facilitating student performance.

For example, early versions of the CST required students to read two abridged news articles, one immediately following the other (the initial case and the critical incident case) and then write a "memo" to the interest group. Although students were reminded to include four specific points and explanations as part of the memo, the entire format proved too dense. Additionally, many students were unsure what was meant by a "memo" and what type of information and evidentiary warrant they needed to include. This resulted in great variability across student responses that may have been unrelated to content knowledge and disciplinary reasoning.

Subsequent CST revisions and the final shell produced more consistent representations of the assessed constructs. First, the task was divided into two-parts (described above under CST shell) so that students considered each case/document sequentially rather than all at once—students answered specific questions about each case before considering their similarities and differences. Second, several explicit, short-answer questions were added that targeted key factual information judged essential to the course objectives but often omitted in students' explanations. Third, additional scaffolding was provided to support students' extended written responses. Overall, these revisions were intended to reduce the construct-irrelevant cognitive load associated with responding to the complex scenario tasks. At the same time, they provided more structure so that students could demonstrate knowledge they had learned.

5.2.3   Working with content and constructs

 Aspects of content and construct validity are often examined using expert review. Following Miller and Linn (2000), we convened two different panels of expert teachers to examine the final version of the CST. This being a summative test, we asked them to judge how well the tasks represented the course content and constructs—its core concepts and disciplinary reasoning strategies—and how well it measured the goal of deeper understanding.

The first review panel was composed of three teachers who had implemented the APGOV+ course and participated in prior CST scoring. This provided an "insider" check that the assessment aligned with the curriculum being taught. The second panel was composed of four APGOV teachers who were not familiar with the revised course or the CST but who had been recognized by the College Board as accomplished APGOV teachers. This "outsider" panel provided a check on how the CST measured up to the core content and reasoning strategies defined for the course. Neither group of teachers had participated in the development of the CST.

The panels evaluated the CST on three criteria: disciplinary content, deeper learning,

and authenticity. Panelists independently assigned a rating of 1 (strongly disagree) to 5 (strongly agree) on each of the three criteria. The panels concluded, with a high degree of agreement ($\bar{x}$ > 4.4, SD = .53-.70), that the CST, overall, (a) covered the core disciplinary concepts and reasoning skills from the course, (b) measured deeper learning, and (c) assessed meaningful content and reasoning applied to authentic, meaningful contexts. This was one indication that we had met our goals for this summative test of deeper learning.

Concurrent evidence for validity was difficult to establish because there were no existing assessments that aligned with the goal of assessing deeper learning of the specific disciplinary concepts and strategies of the APGOV+ course. Therefore, following Shavelson et al. (1992), we examined correlations of students' CST scores with their performance on other standardized measures that were more and less closely related to the core course outcomes. Specifically, because the course content is drawn from and aligned with the College Board's APGOV curriculum and its APGOV test, the correlation of the CST with the APGOV test was examined. As with other studies of correlations between performance assessments and standardized measures (National Research Council, 2001), we anticipated a moderate—neither high nor low—correlation between the CST and the APGOV test. This was to be expected because the AP test emphasizes breadth over depth and because the CST is focused on students' deep disciplinary reasoning as well as their content knowledge.

We also examined the correlation between the CST and the ACT Composite score, an overall measure of achievement in English, science, mathematics, and reading that is highly correlated with IQ scores (Koenig et al., 2008). Here we posited a lower correlation than with the APGOV test, as the CST is a measure of deeper learning in a specific subject area rather than general academic ability. This lower correlation was of special interest to our work and to the validity of the CST because the APGOV+ course had been designed to support all students, including those in urban, poverty-impacted schools who until recently had rarely been admitted to or succeeded in AP courses. We reasoned that students should be able to perform well on the CST because they had learned the content and disciplinary reasoning in the course, not simply because they had general school ability (Miller & Linn, 2000). These hypotheses were borne out: the correlation of the final version of the CST with the College Board APGOV test was moderate (n=241, *Pearson's r* = .60) and the correlation with ACT was substantially lower (n=225, *Pearson's r* = .33). These comparisons with other tests provided a helpful check on the validity of the CST—that it was measuring the course content rather than general school ability.

## 5.2.4   Working with scoring rubrics and procedures

 A major hurdle for classroom assessments of deeper learning is designing scoring rubrics and procedures that are valid and reliable but also efficient. In the case of a classroom assessment administered and scored by the teacher, such as the CST, this can make the difference between assessments that are used and valued by teachers and those that are not.

We developed and tried out a number of different scoring rubrics with the aim of achieving acceptable levels of scoring reliability and validity. We experimented with rubrics that had 5-point scales for each question and rubrics that were organized by the two dimensions of the course outcomes—deep content knowledge and disciplinary reasoning. However, the scales proved unreliable across raters, and a factor analysis did not support the two factors; content knowledge and disciplinary reasoning were highly correlated (*Pearson's r* = .68). The final scoring rubrics were designed to measure students' disciplinary reasoning *with* content (RWC).[3] The scoring rules require scorers to consider content and reasoning simultaneously as they rate each response. Adapted from a model developed at Tennessee Technological University for the Critical-thinking Assessment Test,[4] a dichotomous tree rubric and exemplar responses take scorers through a series of yes/no questions that lead through the rubric via various pathways. The simplicity of yes/no choices along the scoring path increased the consistency of scoring. Because some pathways are better than others (there are more or less informed ways of arriving at the same conclusion), these pathways are rated as high, mid, and low in disciplinary reasoning. Similarly, content is assigned points based on accuracy and completeness as well as the point value of a particular question.

The scores were combined to get a total score for each question. What is unique about this scoring scheme is that it prioritizes disciplinary reasoning: a strong reasoning path enhances the content score and a weak path diminishes it. Conversely, a weak content score combined with a stronger disciplinary reasoning path earns more points than the reverse. All item scores were summed to create a total RWC CST score for each student. For sample student responses to one of the test questions see Figure 2.

Figure 2. Sample student responses

Rating = Low
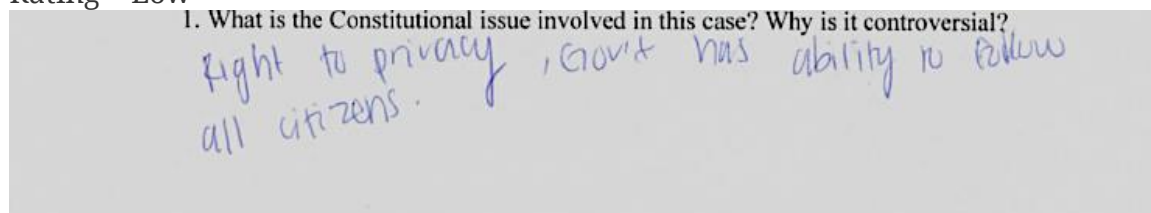
"Right to privacy. Gov't has ability to follow all citizens."

Rating = Mid

"The constitutional issue is whether entering Moreno's "curtilage" and attaching a GPS to his car, and subsequently monitoring his movements, violates the 4th Amendment or not. The 4th Amendment protects the right of citizens to be secure in their possession against unreasonable search & seizure."

Rating = High

"The Constitutional issue of government GPS tracking possibly violating 4th Amendment rights to privacy is involved in this case. It is controversial because there is uncertainly in how much the 4th amendment really covers."

Teachers scored the CST. After 2-3 hours of training and calibrating scores, raters were able score each test in approximately 10 minutes. Six raters each scored a random sample of the total papers (n=385), with approximately 10% of the papers double-scored. Interrater reliability across forms was estimated in two ways. First, interrater agreement was calculated for scores on individual items to learn more about the efficacy of the rubrics for scoring complex, open-ended items Here, results for Cohen's Kappa were generally in the "good – very good range" (70% – 93%)[5]. Results for exact agreement were in an acceptable range (82%–97%, Niemi et al., 2007) especially when considering the CST as a low-stakes assessment.

6   DISCUSSION

We designed and tested an assessment of deeper learning to measure students' content and reasoning strategies in civics and to apply them to new, meaningful problems found in their communities. A deeper understanding, we specified, facilitates learning and action in the future when conditions have changed. Over seven years and multiple school settings, we used DBIR to iteratively research and revise the assessment. Positioned as a summative, end-of-course, classroom-based assessment, the CST is a tool to assess individual students' long-term learning across the course *and* to provide evidence to us and our teacher collaborators that can inform revisions to the course across time, schools, and states. Through its iterations, the CST stayed true to its complex scenario format,

challenging students to think deeply about, and to apply, the concepts and reasoning strategies they had learned.

The foundational activity of any assessment of deeper learning is the identification of the core disciplinary concepts and reasoning strategies. Of course, these need to be sharply limited in number, as they are the objects of both instruction and assessment. Although this may seem straightforward or obvious, we found this curriculum decision making to be the most important and challenging undertaking in this whole endeavor. There were many debates as to which concepts and strategies would be deemed important enough to become the constructs targeted by the assessment, and others may have made different decisions. But certainly, we could not have developed a summative assessment of deeper learning without this curriculum deliberation. Even then, decisions about what to include or exclude in a summative test are still needed. In particular, issues of curriculum representation have to be considered.

We learned that early versions of the CST included features that were problematic: Reading and writing demands risked construct-irrelevant variance on the test, and complicated task and question structures risked the validity of students' scores. Similarly, early scoring rubrics were too complex and failed to support a two-factor reporting structure. Finally, we learned about the importance of beginning with deliberative curriculum decision making and, thereby, focusing on a limited number of concepts and strategies that are the objects of deeper learning and become the focus of the assessment.

6.1 Limitations

We conducted our work over several years and, as a result, our samples of students and teachers were neither representative nor stable over multiple studies. Furthermore, the CST itself was in flux every year as we refined it in response to data and teacher feedback. On the one hand, this shifting context made it difficult to compare findings over time or to generalize to other samples. On the other hand, working in the messy real-life DBIR context provided ecological validity to our work (field testing across contexts) and a continual check on the alignment between the assessment and on-the-ground instruction. It made us sensitive to the feasibility of using the CST to assess deeper learning, and we benefited from ongoing feedback and advice from our collaborating teachers.

We deployed a definition of deeper learning that should travel well across social studies courses, but the assessment task—giving competent advice to an interest group—may not travel as well. While it is widely applicable in government and politics courses, and civic education more broadly, and while we can easily imagine it in economics and some domains of geography, its use in history may be limited as it may encourage presentism. Conversely, it may be possible to direct students, when role-playing an advisor to clients in the past (e.g., Malcolm X, Napoleon) to contextualize their advice.

We were not able to conduct cognitive interviews with students who took the final iteration of the CST. Although we did conduct cognitive think-alouds in order to assess

students' ability to read and comprehend textual material on the revised CST, we were not able to verify the thinking students used to respond to the questions except by examining their written explanations. It would be good to follow up on this.

6.2   Implications

Several features of our assessment design present avenues for advancing assessment of deeper learning not only in civics but across social studies courses. First, we operationalized assessment of deeper learning by having students use, on the test, what they had learned in the course to solve a novel problem that is meaningful, that is, relevant to students' lives and communities. And by following the first problem with a critical incident, we were able to gather a second sample of student performance, all within a single assessment that can be completed in one class period. This model of bringing knowledge to bear, or what we call "knowledge in action," marks a critical balance between what students had learned in the course and their ability to apply it. In our experience with other project-based curricula and performance assessments, students too often can reason their way to a weak or poorly-reasoned solution without deeper understanding of the disciplinary content and reasoning they have learned. Instead, they simply apply everyday, experiential, or spontaneous knowledge (Vygotsky, 1986) to solve the problem. We worked hard to avoid this threat to the validity of the assessment scores.

Second is a rubric design that prioritizes reasoning *with* content. Our efforts to develop a rubric that integrates content knowledge and disciplinary reasoning grew from both our empirical data (the two were highly correlated) and the nature of deeper understanding (Bransford & Schwartz, 2000; Hitano & Inagaki, 1986). Flexible application by definition requires students to reason with knowledge, determining the relationship of old and new. Accordingly, the dichotomous rubric design articulated a clear path for scoring which ultimately improved reliability across raters. Although some might question our decision to weigh disciplinary reasoning more heavily than conceptual and factual knowledge, a major goal of the revised curriculum was, indeed, to teach reasoning with content. By elevating its importance in the rubric, we hoped to communicate a priority for instruction—one that is important to a course striving for deeper learning.

Third, we developed this model of assessing deeper learning to serve as a classroom-based *summative* measure in contrast to others that are formative. We were driven by teachers' requests for a valid, easy-to-administer and score, end-of-course test that aligned with the course outcomes. Without something like this, many teachers resort to textbook tests or outside measures that are not aligned with the goal of teaching for deeper learning. We were driven, too, by our desire to have a common assessment we could use across teachers and classrooms and that would provide a concrete representation of the aim of teaching for deeper learning, one that could, and did, prompt rich discussions about the what and the how of teaching for deeper understanding. One, if not the most, important consequence of good assessment is to improve teaching and learning, and to that end, we

believe the CST model and the collaborative process of developing it is a step in the right direction.

We see the potential of using the CST model in other ways as well. We plan to construct a set of interim assessments instead of, or in addition to, a summative assessment (Perie et al., 2009). Interim assessments occur at strategic points in the curriculum such as the end of major units of study (there were five in APGOV+). Courses that use looping to deepen learning could use the same shell across units but focus the scenarios on the specifics of each unit. The assessments would be easy to administer (one class period) and score (dichotomous rubrics), provide measures of individual students' progress, and most important, would provide formative feedback to teachers and students to inform instruction. Performance could be aggregated for each student across the course. This aggregate score could be compared with a single summative CST. The aim here would be to embed assessments of deeper learning into the fabric of instruction by having both interim and summative assessment lead to a coherent course assessment system across the year or semester (National Research Council, 2001; Niemi et al., 2007).

We hope our assessment model and the CST contribute both substantive and procedural knowledge to new efforts to elevate the importance of deep learning in civic education and its assessment. Without vigorous research and development activity on this front, it is unlikely that our field will be able to support the goals of educating knowledgeable and responsible citizens.

REFERENCES

Baker, E. L. (1998). Model-based performance assessment. *CSE Technical Report 465*. CRESST, University of California, Los Angeles, CA. The Regents of the University of California.

Baker, E. L., Chung, G., & Cai, L. (2016). Assessment gaze, refraction, and blur: The course of achievement testing in the past 100 years. In P. Alexander, F. Levine, W. Tate (Eds.), *Review of research in education* (Vol. 40, pp. 94-142). Los Angeles, CA: Sage Publications.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Bransford, J. D., & Schwartz, D. L. (2000). Rethinking transfer: a simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education* (Vol. 24, pp. 61-100). Los Angeles, CA: Sage Publications.

CivXNow. (n.d.). *State legislation database tracker*. Cambridge, Mass: Author. Retrieved on December 1, 2021 from https:docs.google.comspreadsheetsd1gt98M4JU67YAVofHPYLhhXs40Qai1me5Y76YYlDN1j0edit#gid=0

College Board. (2018). *AP U.S. government and politics course framework.* New York, N.Y.: Author.

Davies, I., Grammes, T., & Kuno, H. (2017). Citizenship education and character education. *Journal of Social Science Education, 16*(3), 2-7.

Ercikan, K. & Seixas, P. (Eds.). (2015). *New direction in assessing historical thinking*. London, England: Routledge.

EuroClio. (n.d.). *The future of education and skills 2030.* OECD. Retrieved December 10, 2020 from https://www.euroclio.eu/ .

Fishman, B. J., Penuel, W. R., Allen, A.R., Cheng, B. H., & Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *National Society for the Study of Education, 112*(2), 136-156.

Galston, W. (1989). Civic education in the liberal state. In N. Rosenblum (Ed.), *Liberalism and the moral life* (pp. 89-102). Cambridge, MA: Harvard University Press.

Gessner, S. (2017). Teaching civic education in a migrating global community. *Journal of Social Science Education, 16*(2), 42-52.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hansen, M., Levesque, E., Valant, J., & Quintero, D. (2018). *The 2018 Brown Center report on American education.* Washington, DC: Brown Center on Education Policy at Brookings.

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), *Child development and education in Japan* (pp. 262-272). New York, NY: Freeman.

Heritage, M., Kim, J., Vendlinski, T., Herman, J. (2009). From evidence to action: A seamless process in formative assessment*? Educational Measurement: Issues & Practice, 28*(3), 24-31.

Hewlett Foundation (2013). *Deeper learning defined.* Author. Retrieved September 8, 2018 from http:www.hewlett.orglibrarydeeperlearningdefined

Koenig, K. A., Frey, M. C., & Detterman. D. K. (2008). ACT and general cognitive ability. *Intelligence, 36*(2), 153-160.

Linn, R. L., Baker, E.L., & Dunbar, S. B. (1991). Complex, performance based assessments: Expectations and validation criteria. *Educational Researcher, 20(*8), 15-21.

Löfström, J., & Grammes, T. (2020). Outlining similarities and differences in civics education in Europe. *Journal of Social Science Education, 19*(1), 1-4.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*(4), 367-378.

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: Author.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, D.C.: National Academies Press.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C: National Academies Press.

Newmann, F., & Associates. (1996). *Authentic achievement: Restructuring schools for intellectual quality*. San Francisco, CA: Jossey-Bass.

Niemi, D., Baker, E. L., & Sylvester, R. M. (2007). Scaling up, scaling down: Seven years of performance assessment development in the nation's second largest school district. *Educational Assessment, 12*(3 & 4), 195-214.

OECD (2018). *The future of education and skills: Education 2030*. Paris, France: OECD.

Oliver, D. W., & Shaver, J. P. (1966/1974). *Teaching public issues in the high school.* Logan, UT: Utah State University Press.

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

Parker, W. C., Lo, J., Yeo, A. J., Valencia, S., Nguyen, D., Abbott, R. D., 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.

Parker, W. C., Valencia, S. W., & Lo, J. C. (2018). Teaching for deeper political learning: A design experiment. *Journal of Curriculum Studies, 50*(2), 252-277.

Pellegrino, J., Chudowsky, N. & Glaser, R. (2001). *Knowing What Students Know*: *The Science and Design of Educational Assessment.* Washington, DC: National Research Council of the National Academy of Sciences.

Perie, M., Marion, S., Gong, B. (2009). Moving toward a comprehensive assessment system. A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5-13.

Pressley, M., & Afflerbach, Peter. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Reinhardt, S. (2016). The Beutelsbach Consensus. *Journal of Social Science Education, 15*(2), 11-13.

Resnick, L. B., & Klopfer, L. E. (1989). *Toward the thinking curriculum: Current cognitive research*. Alexandria, VA: Association for Supervision and Curriculum Development.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475-522.

Schwartz, M. S., Sadler, P. M., Sonnert, G., & Tai, R. H. (2008). Depth versus breadth: How content coverage in high school science courses relates to later success in college science coursework. *Science Education, 93*(5), 798-826.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sample variability of performance assessments. *Journal of Educational Measurement, 30*(3), 215-232.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*(4), 22-27.

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal, 115*(2), 184–209.

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice, 16*(3), 13-22.

Stenner, A. J., Burdick, H., Sanford, E. & Burdick, D. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement, 7*(3), 307-322.

Valencia, S.W., Wixson, K.K., & Pearson, P.D. (2014). Putting text complexity in context: Refocusing on comprehension of complex text. *Elementary School Journal, 115*(2), 270-289.

Vygotsky, L. S. (1986). *Thought and language*. Cambridge, MA: MIT Press.

APPENDIX

Sample Complex Scenario Test – Form A

Dear AP+ Project students,

We are interested in what you have learned about government and politics in this course and how you can apply that information to a controversial issue.

Today, you will complete an activity in which you take on the role of an adviser to a new citizens group called "Stop Mobile Surveillance." This group wants to stop the government and police from monitoring people's actions by using information from GPS systems in their cars or from their cell phones. You will answer some questions that will help the group understand the issues and how they can try to stop mobile surveillance.

This task has two parts, please read the information on both parts carefully. Use what you have learned in this class and information from the readings to answer the questions in Parts A and B. You can use your copy of the Constitution and your class notes to help you with your answers. Feel free to mark up the text as you read.

Be sure to do the best you can to explain your answers. Don't worry about spelling or grammar. We are more interested in your ideas and how you think than spelling or grammar.

Mobile Tracking: A New Era of Government Surveillance?

Background

Recent court cases show how easy it has become for police and other government officials to monitor (watch) people's movements using mobile technology —information from cell phones and GPS tracking systems. Although surveillance by the police and other government agents is not new, these technologies make surveillance cheaper, easier, faster, and more accurate than ever. *The Stop Mobile Surveillance Group* formed in an effort to stop this type of GPS tracking and surveillance.

PART A

A recent court ruling says that the government doesn't even need a warrant to track suspects using GPS tracking systems. The case of an Oregon man named Juan Pineda-Moreno who was convicted in this manner is summarized in the newspaper article below:

[Document #1: Excerpt of a case featured in Time Magazine inserted here]

Answer these questions to help *Stop Mobile Surveillance Group* understand the issues and to begin their work to end mobile surveillance by the government and the police.

1. What is the Constitutional issue involved in this case? Why is it controversial?
2. Consider the powers granted to government in the Constitution. In the Pineda-Moreno case, which <u>level</u> of government and which <u>branch</u> would be the best place for the *Stop Mobile Surveillance Group* to begin their work to end mobile surveillance?
    a. Circle the <u>level</u>: local, state, federal
Explain why this level of government is the best choice:
    b. Circle the <u>branch</u>: executive, legislative, judicial

Explain why this branch of government is the best choice:
3. After the decision in the Pineda-Moreno case, *Stop Mobile Surveillance* decides to contact linkage institutions to support its actions. Identify two linkage institutions they should contact and explain why each would be helpful to its cause.

    1.) _____

    2.) _____

4. As an advisor to the *Stop Mobile Surveillance Group*, describe the three most important steps that this group should take to advance its cause, and then explain why each step is important.

PART B

After receiving the advice YOU gave in question 4, *Stop Mobile Surveillance* learned that another U.S. District Court of Appeals, this one in Washington D.C., issued a ruling on another GPS tracking case. This case resulted in a decision that is different from the case you read in PART A. This new case is summarized in the news article below:

[Document #2: Excerpt of case featured in The Washington Times inserted here.]

Answer the following questions with consideration to both court cases.

5. How does the Washington D.C. case help or hurt *Stop Mobile Surveillance* in its goal to stop GPS tracking? Explain your answer.
6. Consider the ruling in the Washington D.C. case. Now add one new step OR make one change to the steps you suggested in question 4. Explain your reason.

ENDNOTES

[1] There are nearly 40 AP courses, and they are offered in many high schools across the United States. Similar to the A-Levels in the United Kingdom, participation in these courses and passing the summative test can influence students' college admissions and, once admitted, they often can skip the matching introductory courses, hence "advanced placement."
[2] Across the seven years of this research, and across the 13 schools involved, the percentage of students eligible for free and reduced-price lunch ranged from 50-100%.
[3] See Ercikan & Sexias, 2015, for a discerning treatment of the role of content and reasoning in assessment.
[4] Our thanks for their advice to developers Barry Stein and Ada Hynes.
[5] See http://www.pmean.com/definitions/kappa.htm

AUTHOR BIOGRAPHIES

Sheila W. Valencia is Professor of Literacy Education Emeritus at the University of Washington, Seattle. Her research focuses on K-12 reading and writing instruction, assessment, and policy. She consults with national and state educational agencies in the USA to develop both large-scale and classroom-based assessments.

Walter C. Parker is Professor of Education and (by courtesy) Political Science Emeritus at the University of Washington, Seattle. His research focuses on civic education in elementary and secondary schools. His new book is *Education for Liberal Democracy: Using Classroom Discussion to Build Knowledge and Voice.*

Jane C. Lo is Associate Professor of Teacher Education at Michigan State University. Her research focuses on the political engagement of youth, social studies curriculum development, and developing measures of deep learning and collaboration. Her methodological expertise includes mixed-methods designs, design-based implementation research, interview and survey methods, and advanced correlational techniques. She teaches courses in social studies methods.