Flexibly Using the Surveys of Enacted Curriculum to Study Alignment[1]

Morgan S. Polikoff*
Hovanes Gasparian
Shira Korn
Martin Gamboa
University of Southern California

Andrew C. Porter
University of Pennsylvania

Toni Smith
Michael S. Garet
American Institutes for Research

*corresponding authors
3470 Trousdale Parkway
WPH 904A
Los Angeles, CA 90089
(630) 430-9416
polikoff@usc.edu

Published online first in *Educational Measurement: Issues and Practice*
September 6, 2019

[1] This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C150007 to the University of Pennsylvania, Graduate School of Education. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Almost thirty years in, the standards-based reform era in U.S. education shows no sign of abating. Federal law requires states to have content standards in mathematics, English language arts (ELA), and science, and states generally have standards in many other subjects as well (for example, the Massachusetts Department of Elementary and Secondary Education currently lists curriculum frameworks in those three subjects, and also in history and social science; digital literacy and computer science; vocational technical education; English language development; arts; comprehensive health; and foreign language). These standards are intended to guide the content of teachers' instruction, laying out a map for what students should know and be able to do. In core subjects, students are also assessed on state tests that are intended to reflect their mastery of those standards. Alignment—among standards, assessments, and instruction—therefore remains a pressing measurement and policy issue.

There are several main alignment methodologies, and these are summarized and compared in several recent reviews (Cizek, Kosh, & Toutkoushin, 2018; Martone & Sireci, 2009; Polikoff, in press). One of the most widely used and studied alignment methodologies is the Surveys of Enacted Curriculum (SEC) approach (Porter, 2002). The SEC approach involves coding a document against a neutral content language that defines content at the intersection of specific topics and levels of cognitive demand. Then, these coded documents are compared using alignment indices (which we describe below). SEC content languages exist in mathematics, ELA, science, and social studies, and the SEC has been used in dozens of policy studies over the last decade-plus (for example, see Polikoff, 2012b, 2015; Porter et al., 2011).

Typically, the SEC is thought of as having three characteristics: 1) a set of content languages that have been established by content experts and validated over time, 2) a data collection approach relying on content analysis of content standards, tests, and curriculum

materials and on end-of-semester or end-of-year surveys of teacher instructional practices (though in fact the existing SEC can be applied to time periods of any length), and 3), standard approaches to analyzing the data to produce indices of alignment and other indices of content coverage. The purpose of this article is to examine and expand the ways that alignment can be thought of and analyzed using the SEC. To do this, we report on three portions of the development and validation work we have done on a revised version of SEC content languages in the context of a national study of the implementation of new standards. The examples are specific, but we draw more general implications. Based on our experiences, we make recommendations for researchers, policymakers, and instructional leaders who might want to use the SEC or SEC-like surveys (by which we mean surveys or content analysis tools with a similar general format to the SEC) for their own work.

Background

Alignment in Research and Practice

The concept of alignment is at the heart of standards-based reform policy efforts (Smith & O'Day, 1990). Alignment plays several important roles in the theory of action underlying these policy reforms. First, aligning teachers' instruction with state standards is one of the core goals of standards-based reform policy—it is through this instructional alignment that student opportunity to learn is intended to improve and, concomitantly, student achievement. Second, alignment among policy instruments—standards, assessments, curriculum materials, professional learning opportunities, etc.—is thought to be one of the main mechanisms through which instructional alignment, and therefore improved student achievement, are attained. Thus, measuring alignment among instruction, standards, assessments, and other policy documents is essential to understanding the implementation of standards-based reforms (in the U.S. or abroad).

Research over the past decade has both supported and undermined elements of this theory of action. For example, there has indeed been a great deal of effort to improve the alignment of teachers' instruction with state standards, and a variety of studies offer evidence that this has happened to some extent (e.g., Hamilton & Berends, 2006; Polikoff, 2012). Survey and interview studies, as well as longitudinal analyses, find evidence that teachers are improving the alignment of their instruction with standards (e.g., Hamilton & Berends, 2006; Pedulla et al., 2003, Polikoff, 2012a, 2012b, 2013). These longitudinal analyses, which use the SEC, find associations of alignment with teacher and policy variables in ways that are predicted by theory. On the other hand, research has also raised troubling questions about the theory of action. For instance, the relationship of alignment to test score inflation and teachers' unintended instructional responses has been a source of study and critique (e.g., Holcombe, Jennings, & Koretz, 2013; Koretz, 2003, 2005). Taking this evidence together, there is clear evidence in favor of the theory of action for standards-based reforms, but also a number of serious design concerns and unintended consequences.

The Surveys of Enacted Curriculum

Content languages. There are several main approaches to measuring alignment (for recent reviews on the topic see Cizek, et al., 2018; Martone & Sireci, 2009; Polikoff, in press). Among the most widely used of these is the Surveys of Enacted Curriculum. This approach emerged out of a stream of research that was focused on understanding how teachers make decisions about what content to teach (see for example Schwille et al., 1988). The SEC approach is based on a set of content languages that are intended to cover all the content a teacher might teach in a given subject in any grade, k-12. These content languages were developed over time with the input of content area experts, and they continue to be revised on a periodic basis (the

most recent topic revisions occurred in 2013). The content languages can be used by teachers to report on the content of their instruction or by trained coders to content analyze the content of any document—standards, assessments, curriculum materials, etc.

There are currently SEC content languages in mathematics, English language arts, science, and social studies. Each content language defines content at the intersection of specific topics and levels of cognitive demand (also sometimes called "student expectation"). For instance, prior to the work we describe below, in mathematics there were 185 topics and five levels of cognitive demand, meaning there were 925 possible content distinctions. All four subjects used five levels of cognitive demand[2], generally ranging from memorization and procedures to application and generalization, though the names of the levels differed across subjects. In each subject, the topics were nested under broader clusters—for instance, the 185 topics in mathematics were nested under 16 clusters. These are commonly referred to as fine-grained and coarse-grained topics, respectively. Figure 1 shows an excerpt from this prior version of the SEC – this excerpt shows the five levels of cognitive demand and the fine-grained and coarse-grained topics (in this case only two of the coarse-grained topics are shown).

The SEC surveys and content analysis procedures. When used by teachers to report on the content of their instruction, the content languages are generally presented on an end-of-semester or end-of-year survey, though they have also been applied in log form to cover shorter time periods. The teacher is asked to consider a target class and a given time period. As shown in Figure 1, they then report the content of their instruction by first indicating their coverage of particular topics (using a four-point Likert scale indicating the number of days of instruction) and

[2] Another possible dimension might be "difficulty," and we note that difficulty is distinct from cognitive demand. For example, items emphasizing memorization can be easy (e.g., "What is the capital of the United States?") or more difficult (e.g., "What are the capitals of all the countries in Europe?"). The SEC is silent on difficulty, and we do not emphasize difficulty on our modified instruments nor in our intervention work with teachers.

second indicating their coverage of each of the levels of cognitive demand [using an indicator for major vs. minor emphasis (in the past, a four-point likert scale was also sometimes used for this rating of emphasis)]. The teachers' responses are then turned into proportions, so that the value in each cell is the proportion of the total semester's or year's content on the given topic/cognitive demand combination (e.g., the percent of time spent covering "line graphs" at the level of "memorize/recall"). The left side of Figure 2 shows an example matrix of proportions for a single teacher's instruction for a simplified set of topics and cognitive demands.

The content languages are also used by trained coders to analyze content standards, assessments, and curriculum documents. First, each document is broken into fine-grained chunks to facilitate coding. For standards, the chunks are usually the objectives. For tests and curriculum materials, the chunks are usually items. For curriculum materials, the chunks also typically include sections of text or example problems. The coders then analyze each chunk using the content language, identifying up to six SEC cells that are covered by the given chunk. The content analysis for the document is determined by computing a weighted average of the codes for each chunk. The weight for a given chunk is evenly divided across the cells coded for that chunk. For standards and curriculum materials, each chunk is typically weighted equally. For tests, chunks may receive differential weights (e.g., the codes for a two-point item would be double-weighted as compared those for a one-point item). The content analysis for a given rater is then calculated by averaging across coded cells; examples of individual raters' content analyses are shown in the top right and middle right of Figure 2. Finally, the sets of ratings, one from each rater, are then averaged to arrive at the overall rating; the bottom right matrix in Figure 2 illustrates this, as it is the average of the two individual raters' matrices above it. This again takes the form of a matrix of proportions, with the proportion in each cell representing the

proportion of total content in that document that is on the given topic/cognitive demand combination.

Analyzing SEC data. The resulting content matrices are used to calculate various indices of alignment. The main alignment index, proposed by Porter (2002) is as follows:

$$\sum_i 1 - \frac{|x_i - y_i|}{2}$$

Here, x_i represents the proportion of document (or teacher survey) x's content that is located in cell i, and y_i is the same proportion for document y. In the case of Figure 2, document x would be the teacher survey, and document y would be the average content analysis in the bottom right. The alignment index for this particular comparison would be .40. The alignment index ranges from 0 to 1, and it can be thought of as representing the proportion of the content located in the same cells in the two documents. Various other alignment indices have been presented (see Polikoff, 2012b for examples).

The SEC in research. The SEC tools have been used in a variety of research applications. One set of studies has used the SEC to compare standards with each other (e.g., Porter et al., 2011; Porter, Polikoff, & Smithson, 2009), assessments with standards (e.g., Polikoff, Porter, & Smithson, 2011), or curriculum materials with each other or with standards (e.g., Polikoff, 2015). This work tends to find modest alignment among these documents (generally alignment indices in the .25 to .40 range, with the exception of comparisons of textbooks that produce indices closer to .70 to .80) and often identifies common areas of alignment or misalignment. For example, research finds that both assessments and curriculum materials tend to give more emphasis to lower levels of cognitive demand than do standards (Polikoff, 2015; Polikoff et al., 2011).

A second set of studies has used SEC content analyses and teacher survey data to examine teachers' instructional alignment, changes over time, and relationships with policy and descriptive variables of interest. For instance, this work has identified increases in teachers' instructional alignment during the NCLB. This work has also related alignment to teacher experience and education levels, as well as to policy attributes such as the coherence and use of rewards and consequences under standards-based reform policy (Polikoff, 2012a; 2012b; 2013). Finally, this work has related alignment indices to student achievement outcomes, finding evidence of strong associations in one study (Gamoran et al., 1997) but weaker associations in another (Polikoff & Porter, 2014).

A final set of studies has focused on methodological issues in the SEC content analysis or alignment calculation procedures. One area of this work has focused on identifying the number of content analysts needed to achieve highly reliably coding (four for standards and assessments, two for curriculum materials, see Polikoff, Zhou, & Campbell, 2015; Porter, Polikoff, Zeidner, & Smithson, 2008). Another area has focused on laying out techniques for statistical inference on alignment indices (recommending simulation studies and generalized linear models, see Fulmer, 2011; Fulmer & Polikoff, 2014; Polikoff & Fulmer, 2012). This work has determined whether alignment indices are significantly greater than chance and identified significant contributors to misalignment.

Gaps in the SEC methods. While the SEC is flexible and adaptable for a wide array of uses, there are some ways in which the instruments and techniques could be improved to make them more useful. For example, because the SEC is designed to cover all of the content a teacher might cover in any grade, k-12, the topics listed are required to be somewhat coarse. If this were not the case, then the instrument would be overwhelmingly long. So, for instance, there is a topic

on the mathematics SEC for "multiply whole numbers and integers," but there are multiple Common Core mathematics standards that offer more specifics about the multiplication concepts students are supposed to learn with respect to whole numbers. For example, a third grade Common Core standard calls for students to use strategies involving the relationship between multiplication and division and to know from memory all single-digit products, whereas a fourth grade standard expects students to use strategies based on place value and the properties of operations. In a study focusing on teachers in grades where students are learning whole number multiplication, more fine-grained topics in this area would be essential. The existing SEC procedures do allow for topics to be added for individual studies, but a more comprehensive rethinking (allowing existing categories to be deleted or changed and new categories to be added) is not in the traditional repertoire of the instrument.

For another example, in situations where the SEC has been used to give teachers instructional feedback, it has typically been at the end of the year (though, as stated above, it has also occasionally been applied in log form or over a shorter period of time), but it might be useful to give teachers more real-time feedback on the content of their instruction. Tweaking the design of the instrument or its administration instructions could help make the instrument more feasible for this use. Of course, there are other ways in which the existing instruments might need to be modified by researchers, and it would be useful to lay out a general process for implementing those modifications to create a version of the SEC that works for each individual study. We describe our intended uses of the revised SEC in the next section, and we discuss the tradeoffs of modifying the instrument to tailor it to individual studies in the discussion section.

Context for This Study

The research presented here was part of a large, five-year project to study standards implementation in U.S. schools. The broader project contained four main strands, two of which necessitated the use of the SEC. The first strand, which we report here, was a measurement study focused on refining and providing validity evidence for the SEC or a SEC-like set of tools. The end goal of the measurement study was to create a set of instruments to be used in the fourth strand, an experimental study designed to test the impact of a standards-focused coaching intervention on instruction and student achievement. Because of our intended use of the SEC tools in the intervention study, we needed an instrument that had several key features.

First, our resulting instrument had to be able to be used by teachers in the intervention study to report on their content coverage in a way that would allow us to estimate alignment with whatever college- and career-ready standards were in place in their states. This meant the content languages had to be updated to ensure that the content codes were exhaustive as to the content in those states' standards.

Second, our resulting instrument had to provide data at a fine-grained enough level of detail that our alignment ratings would be accurate and that our feedback to teachers would be useful. This meant working with both teachers and coaches to ensure that the alignment ratings made sense to them, passed the "smell test" for validity, and had the possibility of offering them evidence they could use to drive instructional change. As we learned during our study, this also meant that the content codes needed to be much finer-grained than we originally thought.

Third, our resulting instrument had to be feasible for teachers to complete in the context of the intervention study, which involved asking teachers to complete a log and obtaining feedback multiple times over a school year. That is, the instrument could not be so onerous that no one would agree to participate (or, in other words, the level of burden could not exceed the

level of utility teachers perceived they were getting from the intervention). This meant that we needed to figure out how often the survey was going to be administered and how teachers would complete it in such a way that they found feasible and not overly burdensome.

In the remainder of this paper, we address each of these features, describing how we carried out three phases of development and research to achieve the features. The results accomplish two main aims. First, they provide new evidence on the SEC, its procedures, and the quality of the data it produces. Second, and more importantly, the results offer a guide for researchers in other areas who would like to use SEC-like methods to study similar content-related issues. Based on our experience and the results we present below, we think there is broad applicability for researchers to create or modify SEC-like instruments to answer a wide variety of research questions related to standards, assessment, curriculum, and accountability policies and teachers' curricular enactment in the classroom.

Phase 1: Ensuring Adequate Content Coverage

Revising the SEC Content Languages

In our project proposal, we recognized that the SEC, while representing a solid foundation for our content analysis work, would need to be reviewed to ensure it would work well for our intervention study. In particular, we were worried about the number of topics, their grain size, and their coverage of the content standards in our five partner states (at the time of the proposal, those were four Common Core states plus Texas). We were also worried about the number of cognitive demand levels, their definitions, and the ability of teachers to meaningfully distinguish among them, based in part on feedback from content-area experts (e.g., Beach, 2011; Cobb & Jackson, 2011).

To address these concerns, we assembled teams of experts to revise both the mathematics and ELA SEC content languages. We recruited eight outside experts, four in each subject, to attend an in-person instrument review and development meeting over three days. The experts in ELA included one ELA teacher educator, one expert in literacy instruction, one expert in writing instruction, and one general expert in survey methods. The experts in mathematics included one mathematics teacher educator, one expert in mathematics education (who had formerly been an SEC content analyst), one mathematician, and one general survey expert. All had familiarity with content standards and with content analysis procedures. In addition, four of the project's principal investigators and four other project staff, including two researchers who study students with disabilities and one who studies English learners, attended the development meeting and were evenly split between the two subjects.

Prior to the meeting, participants were sent several documents to prepare them for the work. These included, 1) the existing SEC in their respective subject, 2) the Common Core standards and Texas Essential Knowledge and Skills (TEKS) in their respective subject, and 3) an excerpt from the project proposal that described the goals of the meeting and the intended uses of the revised SEC. The unstructured agenda was focused on allowing each team the time to review the existing SEC and propose whatever revisions they felt were necessary to ensure the content languages adequately covered the target standards. We placed no constraints on our experts' recommendations.

The Recommended Revisions

The revision team in mathematics recommended four main types of revisions to ensure the SEC fit better with the content in the Common Core and TEKS standards that would be the target of our intervention study. First, they recommended specific topics be added to better

capture the content in the target standards. Specifically, they recommended more than forty topic additions, mostly focused on operations, but spread across the coarse-grained topics. These included topics that specified specific algorithms that were covered in the standards (e.g., place value algorithms vs. traditional algorithms for the four main operations). Second, they recommended edits to existing coarse- and fine-grained topics to clarify their meaning or simplify the organization of the instrument. These included combining two coarse-grained geometry sections and edits such as "Evaluate formulas, expressions, and equations" becoming "Formulas, expressions, and equations" (because the former conveyed a cognitive demand level in addition to a topic). Third, they recommended expanding the number of cognitive demand levels from five to six, replacing "demonstrate/communicate understanding" with "communicate/explain mathematical ideas" and "identify/interpret/explain mathematical relationships" and changing the names of two other categories. Fourth and finally, to address the Common Core's Standards for Mathematical Practice (and similar math practices found in the TEKS), the reviewers recommended adding math practices questions at the end of the survey.

The revision team in ELA recommended a similar set of four major categories of revisions. First, they recommended a number of topic additions, especially in the area of writing (for example, they proposed expanding the number of topics under the area of writing processes and strategies from 7 to 12, with additions like "How to select topics/genres" and "Gathering and organizing information with technology"). Second, they recommended edits and reorganization of existing topics throughout the instrument to better align with Common Core and TEKS standards (e.g., changing "Purpose" to "Author's purpose and how it shapes organization, format, and meaning in a text"). Third, they recommended collapsing the five cognitive demand levels to just three—recall/reproduce, skills/concepts, and strategic and extended thinking. And

fourth, they recommended a set of questions focused on text complexity (e.g., the author and title of the main text used, and the extent to which students read the text and understood it), which was an important dimension of Common Core standards that was not well covered by the existing two dimensions.

In addition, the two content area teams came together to make some joint decisions. These decisions were mainly around the ability of the instrument to accurately capture the extent to which teachers were differentiating their instruction for students with disabilities and English learners—two focal student groups for our work. One goal we stated in our proposal was to examine the alignment of the instruction received by these two groups as well as the alignment for typical students. To measure this differentiation, our reviewers and our team members with expertise in these areas recommended adding a third dimension for each topic that was aimed at teasing out differentiation efforts by first asking teachers whether they altered their coverage of each topic they indicated covering and second asking them to indicate their cognitive demand coverage for each group they indicated altering for. In other words, if a teacher said she altered instruction on "multiplying fractions" for SWDs and ELs, she would then have to re-allocate her cognitive demand coverage on that topic for each indicated group (e.g., a teacher could say she covered multiplying fractions at a conceptual level for the average student and as a procedure for the students with disabilities). The teams also changed the response scales to 0-to-6 Likert scales.

Assessing the Recommendations

We evaluated the recommendations in two rounds of cognitive interviews (Desimone & Le Floch, 2004), conducted with general education math and ELA teachers from multiple grade levels. There were five participants in each round; this number was not chosen in advance, but as we were conducting the interviews we realized we were no longer learning new things toward

the final few interviews in each round. The cognitive interviews focused on the new cognitive demand levels and revised topics, as well as the modifications for students with disabilities and English learners. Protocols (available upon request) generally included two types of probes. First, we asked specific probes about language we had added to or modified in the survey (e.g., "What do you think the authors mean by [insert topic]?"). Second, we asked general probes as is typical in cognitive interviews (e.g., "What were you thinking about when you responded X?").

Data from the cognitive interviews was analyzed by two project team members working independently. Their task was to read the transcribed cognitive interviews and identify revisions they felt were needed based on their holistic appraisal of the five interviews from a given round. The two team members then shared their proposed revisions and discussed differences until consensus was reached. At this point, proposed revisions were shared with all of the project co-PIs and agreement was reached before changes were made. Revisions were made after both the first and second sets of cognitive interviews, and revisions after the second set were checked with both the cognitive interview participants and the revision teams by emailing them the suggested changes and asking their opinions. These cognitive interviews generally revealed that teachers understood the topics and cognitive demands well, with minor wording changes recommended. However, respondents did not use the third dimension that was added for ELs and SWDs. Even when directly probed, every teacher we interviewed said they did not alter topic or cognitive demand coverage for students with disabilities or ELs. Thus, this portion was adding a time burden but not resulting in any usable data.

Based on this feedback, we revised our handling of SWDs and ELs on the survey. We removed the newly-added dimension where we asked how teachers altered their coverage of each individual topic for SWDs and Els. We replaced this with a set of survey questions to be

answered at the end of the survey. In other words, we added questions asking about the routine alterations teachers made across the content area, as opposed to asking about the alterations they made for each individual topic (which they said they did not do). We also made minor wording changes to specific topics and levels of cognitive demand to clarify misunderstandings.

Implications from Phase 1

One strength of the SEC is that it is well-suited to being modified for particular contexts or studies, or even for being used in new content areas. In modifying the SEC for specific studies, we recommend a similar set of procedures to the ones we used here. First, identify the target to be measured—in our case it was the Common Core and TEKS standards, but it might be a different set of standards or some more general content domain. Then, identify a group of experts in that area and bring them together to review the existing instruments and propose revisions. Finally, test the revisions with potential respondents and make additional edits as needed.

Phase 2: Testing Data Collection Approaches

After we had revised the instruments, we conducted a small-scale validation study that was intended to assess the extent to which the instruments could be useful for content analysis and teacher surveys in our intervention study. Specifically, we sought to examine the revised content languages against two questions:

1) To what extent are teachers' reports of content coverage on every-other-week logs aligned with teachers' reports on surveys covering a semester?
2) How reliable are content analyses of teachers' assignments over a week? This question extends the work of Porter et al. (2008) and Polikoff et al. (2015), who examined the reliability of assessments, standards, and entire year-long curricula.

Data and Methods

To carry out this work, we recruited a sample of teachers from across the country and across all grades, k-12. The only requirement for inclusion in the study was that the teacher had to teach at least one section of mathematics or ELA in a U.S. school—there were no other qualifications required. Most participants were recruited through snowball sampling techniques, as there was no intention for the sample to be representative along any particular dimensions. For teachers who taught multiple subjects, we assigned them to one subject. For teachers who taught multiple sections of a given subject, we allowed them to choose any section they preferred. Our initial sample contained 79 teachers—38 in mathematics and 41 in ELA.

For each participating teacher, there were three main data collection activities. First, we asked teachers to complete the content matrix portion of the revised SEC in their chosen subject every other week for a semester, describing the content of their instruction in the period since the last survey. Given the variation in academic calendars, we started data collection after winter break and continued it until the first teacher's school year ended, which was 16 weeks later. Thus, teachers completed a maximum of eight biweekly logs. Second, we asked those teachers to also complete the content matrix portion of the revised SEC at the end of those 16 weeks, covering the entire study period. Third, we asked teachers to submit one week's worth of assignments and assessment activities (anything they asked students to complete, either in class or out). Teachers received bi-weekly reminder emails about their logs, and they received follow-up emails each week if they had not completed a log or the end-of-semester survey. No training was provided, for two reasons. First, as a matter of practicality, we did not have training materials ready at the time of the study because we had finished the revisions just before recruiting the teachers. Second, one of the purposes of this portion of the study was that it would

help us identify areas where training was needed for our intervention study (which it did, for instance relating to cognitive demand coverage and coverage of high-frequency content). However, we did provide participants with documents describing the instruments and the broader project. Participating teachers received incremental incentives for each completed activity, with a total possible incentive of $325.

In the end, we obtained complete or nearly complete responses (which we define as four or more biweekly logs, including the final one, and the semester end survey) from a total of 26 mathematics and 26 ELA teachers, for a response rate of 66%. Of these 52 respondents, the average number of logs completed was 7.42. When respondents skipped a log, we asked them to complete the next log reporting on the four weeks since their previous log. To aggregate their logs, we weighted each log by the number of instructional days it covered. For assignments, we received responses, typically comprising two to eight documents, from 47 teachers (23 in mathematics and 24 in ELA), for a response rate of 59%.

To answer the first question, we compared teachers' aggregated logs to their end-of-semester surveys using the traditional alignment index and other measures of alignment. To answer the second question, we follow Porter et al. (2008) by content analyzing the assignments using four raters, averaging across the assignments within teachers, and then conducting generalizability theory D-studies to estimate generalizability coefficients for two, three, and four raters (in other words, predicting how reliably two, three, and four raters could analyze a week's worth of assignments for each teacher).

Results

The alignment of logs with surveys. The results for question 1 are described here. The average alignment between the teachers' aggregated logs and their end-of-semester surveys was

.44 in mathematics and .57 in ELA. While seemingly low, these are quite a bit higher than typical alignment indices of instruction with standards found in other work (Polikoff, 2012b), and they far exceed chance alignment given the average number of SEC cells reported (Polikoff & Fulmer, 2012). The range in mathematics was from .25 to .64; the range in ELA was .32 to .79. There is considerably more variance in the ELA alignment indices than in mathematics, with a standard deviation of .130 as compared to .085.

Examining the data in other ways gives us a window into common sources of disagreement between teacher logs and end-of-semester surveys. For example, teachers generally reported very similar numbers of topics between their aggregated logs and their end-of-semester surveys. Mathematics teachers reported on average 5% fewer total topics on the surveys than their aggregated logs, while ELA teachers reported on average 11% more total topics on the surveys than their aggregated logs. In contrast, the total number of cells reported tended to be much higher on the aggregated logs (22% higher in mathematics, 52% higher in ELA), suggesting that one common problem was teachers underreporting cognitive demand coverage on the end-of-semester surveys.

Examining individual cells that are often over- or under-represented on the surveys vs. the logs can also shed light on common problems. There are three topics in each subject where the average proportion on the logs differs from the average proportion on the surveys by more than 1% in either direction. These topics are, in mathematics, Add/subtract whole numbers and integers, Formulas, equations, and expressions, and Use of calculators; and in ELA, Comprehension of multi-paragraph texts, Drawing inferences and conclusions from texts, and Listening comprehension. In each of these six cases, the teachers indicated covering it more on the logs than on the survey. As these are all high-frequency topics that are covered often in class,

it may be that it is difficult for teachers to accurately report the extent to which they cover high-frequency activities on the end-of-semester survey (perhaps because the response scale does not make enough distinctions).

The reliability of content analyses of assignments. In mathematics, content analyses for a week's worth of assignments were highly reliable. The average generalizability coefficient was .73 and .79 for two and three raters, respectively. The coefficients would be even larger if assignments from longer time periods were analyzed, as longer documents produce more reliable analyses (Polikoff et al., 2015). In contrast, the content analysis of ELA materials was not especially reliable—just .49 for two raters and .56 for three raters, on average. A few ELA teachers' assignments were especially poorly coded—six teachers had generalizability coefficients under .40. Examining these teachers' assignments suggests reasons for the poor content analysis—these assignments often were more difficult to code, because the actual task was not always clear from what was provided. For instance, one teacher turned in two writing prompts with two rubrics and a checklist. ELA assignments also were sometimes quite short in terms of the number of chunks to be coded—one teacher turned in two quizzes about Romeo and Juliet, which had a total of 16 questions.

Implications of Phase 2

The specific data collection approaches we pursued were those that were most relevant to our research. Of course, other researchers will be interested in different data collection methods and will need different validation approaches. What is important is that the plan for collecting validity evidence be spelled out in advance, that it aligns with what is needed for the desired uses of the instruments, and that it is analyzed in a nuanced way that sheds light on areas of strength and weakness in terms of the functioning of the instruments. In our case, since there was already

relatively strong evidence about the quality of content analysis data, we were less concerned about that (though our findings do suggest that content analysis of assignments in ELA may be an area of challenge). But our findings about the quality of teacher survey data offered us some useful ideas for how to help teachers understand and respond to the survey that we were able to incorporate into our intervention design. For instance, to reduce the burden of completing the logs and survey we decided to present to teachers a much narrower set of topics they might plausibly cover in elementary mathematics, removing from the survey topics in areas such as calculus and trigonometry that we knew they would not cover.

Phase 3: Learning Analytic Lessons During a Pilot

After our initial validation exercise, we used the revised SEC instrument in a six-month pilot with 11 teachers from our target grades and subjects (six fourth grade mathematics and five fifth grade ELA). Broadly, the purpose of this pilot was to ensure the instruments would be suitable for use in the intervention study—there were no specific research questions addressed. Participating teachers filled out the logs every few weeks as was planned for the intervention study, and they were given the kinds of feedback we anticipated giving them in the intervention study. However, no measures of outcomes were taken. In the process of this pilot, we learned several lessons about the design of the instrument and our analytic techniques.

Lesson 1: Analyzing Content Analysis Data

One of the earliest lessons we learned in our pilot was that our method of analyzing content standards resulted in content analyses that weren't consistent with the way our teachers or instructional coaches thought of the standards. Recall that in analyzing standards, each standard is assumed to be equally weighted, and the weight for each standard is then equally divided among the content codes indicated by the coders. The result of this approach is that each

element of content in "larger" standards—those that cover more SEC cells—receives less emphasis in the final content analysis, because the emphasis for the standard is divided among more cells. However, our coaches and teachers tended to interpret standards differently—they tended to see the larger standards as requiring more time to cover, since they contained more content. Thus, rather than allocate each standard one point and dividing that point across the SEC cells represented, we instead decided to equally weight the content by allocating each covered SEC cell one point—that is, weighting each standard by the number of cells it covers. This subtle analytical change resulted in content analyses that were much more consistent with how teachers think of standards, which we think will improve their buy-in to the intervention (not to mention send a better instructional signal to them). Of course, we cannot know what the "true" alignment to standards is without knowing what the standards' authors intended in terms of relative emphasis of the objectives. Since we cannot know this, we think that the best strategy is to use an approach that aligns with potential users' (i.e., teachers') interpretations of what the relative emphasis on the various objectives should be.

Another challenge with the content analysis had to do with how the results were being used by our instructional coaches in the intervention study. In the intervention study, the coaches were expected to take content analysis data from teacher logs, compare it to content analyses of the standards, and make recommendations for areas of instructional misalignment that could be shored up (in addition to offering resources to do so). The typical SEC procedures involve asking multiple raters to code the standards, and using the average of their ratings as the measure of content targeted by the standards. This doesn't work well from a coaching standpoint, however, because the expert raters may disagree, and coaches can hardly tell teachers "some experts think this standard means X and others think it means Y." Both coaches and teachers wanted a

definitive set of codes for their target standards. Thus, to define the instructional target, we engaged two new experts to work together to create a set of master codes that our coaches could use to guide their coaching.

Lesson 2: Improving the Feasibility of Logs

Teachers in our pilot overwhelmingly felt that the instructional logs were too burdensome because of their length and the overlap they saw in certain topics. In conversation with them and with the coaches, we devised three fixes to address this issue. First, we revised the list of topics that were presented to teachers, removing topics we were certain would not be covered by elementary teachers (e.g., trigonometry) and expanding some topics that teachers were still having difficulty parsing (e.g., add/subtract fractions became add fractions with like denominators, add fractions with unlike denominators, subtract fractions with like denominators and subtract fractions with unlike denominators). Second, we reduced the number of times a teacher would have to click by pre-setting all coverage values to zero (whereas before they were blank and teachers had to click zero). This means that a teacher would have to click on a topic only if they covered it at all during a given period, though they must still scroll through each section of the survey and confirm they did not cover any of the content. This is also how topic coverage is treated on the main version of the existing SEC. This immediately reduced the log burden by over 75% in mathematics and over 50% in ELA (given the number of topics teachers generally indicated covering). Third, rather than requiring teachers to first indicate content coverage and then indicate cognitive demand emphasis, we changed the response scale so that teachers just indicated their coverage at the intersection of topic and cognitive demand (keeping the 0-to-6 response scale we already had). We also made these changes to the end-of-year survey.

Implications of Phase 3

It may be that the specific changes we made in response to our pilot are ones that could apply to the SEC more generally—this is an important question worthy of future investigation. At a minimum, our experience suggests that it is important for researchers using the SEC or SEC-like instruments to pilot their tools with prospective users before going into the field. In this way, specific issues, which may differ from context to context, can arise and be dealt with. In our case, the two most important issues were that our coaches needed actionable content analysis data to coach from and that teachers were overwhelmed by the burden of completing the log every two weeks. In each chosen application of the SEC, researchers should investigate these issues but also other issues that may be specific to their particular study or population.

Discussion

The purpose of this paper was to describe our approaches to modifying the SEC in the context of a large study of teachers' instructional responses to college- and career-ready standards. We began with a purposeful rethinking of the content of the SEC content languages in our focal subjects. We organized this rethinking around our instructional targets—the standards under study—and engaged experts to help us think through the most essential revisions to our instruments. We tested these revisions qualitatively with a series of cognitive interviews to ensure that potential users could understand our revised instruments.

Next, we conducted several planned validation studies, one to test the quality of content analyses of assignments using the revised languages and one to test teachers' ability to report their instruction on logs vs. an end-of-semester survey. In those, we found that analyses of math assignments and assessments were highly reliable, but some teachers' ELA assignments were much harder to code. On the other hand, we found that teachers' end-of-semester surveys agreed

more with their aggregated logs in ELA than in mathematics. Our analysis of the survey data suggested that teacher reports on cognitive demand and coverage of highly emphasized topics might have contributed to misalignment between the surveys and the logs. One solution to this problem might be to lengthen the response scale (for instance, make it a 0-to-10 scale), which might allow teachers to more accurately represent their coverage of heavily emphasized content.

Finally, we conducted a pilot of our instrument with teachers like those who would be in the intervention study, where we learned several things about the way we were using content analysis data and the way the instrument could be constructed to reduce teacher burden. These findings led us to make subtle tweaks to the ways in which both teacher survey and content analysis data are collected and analyzed for our study. Though we cannot say if these changes should be adopted more broadly, in the context of our study they were essential.

All things considered, we draw several main conclusions from our research. The first is that the SEC content languages, which have historically been treated as fixed, should instead be treated as a general approach to measuring the content of instruction. Researchers interested in studying the content of instruction, standards, or assessments may find it useful to modify the SEC to better align with their research questions and their focal areas of study. This may mean expanding the number of topics, for instance, in a study focused on a narrower slice of the curriculum (e.g., just writing, just algebra). In addition, it may mean excising topics that are not germane to the particular grade level or area under study. It may also mean changing the cognitive demand levels and their definitions, or even adding other dimensions to the content language as called for by their study. Regardless of the changes that are made, they should be studied with content experts and potential respondents to ensure they make sense and are

adequate for the desired use of the tool. Of course, one potential downside of treating the SEC in this way is that data will not be able to be compared across different studies.

Our second conclusion is that the SEC analytical approaches are more flexible than is commonly thought, and that this flexibility can be useful when diagnosing alignment issues and when applying the SEC in particular contexts. For one, while the convention has been that chunks are equally weighted in the coding process, it may be appropriate to use alternative weighting schemes if that makes more sense in a given context. Similarly, though we did not do it here, there is no reason that content analysis must limit each chunk to a maximum of six SEC cells – this could be relaxed in a future study. For another, though the main alignment index is useful and has been by far the most studied, researchers should flexibly use SEC survey and content analysis data, reporting on areas of alignment and misalignment, in ways that advance the aims of their research. The alignment procedures should be thought of as a toolkit, rather than as a script, for how content analysis data should be analyzed.

Our third conclusion is that the SEC or SEC-like surveys can be used for data collection over varied units of time, from a few weeks to a full semester. However, there may be issues that arise depending on the length of time chosen. For example, we found that teachers underreport cognitive demand coverage and coverage of high frequency topics on the end-of-semester survey relative to the logs. Given that the end-of-semester survey has been the dominant data collection mode for the SEC, this is an area worthy of future study, both in our context and in other contexts. Regardless of the particular data collection time period chosen, researchers should investigate the extent to which respondents are able to reliably report on their instruction over that period.

Of course, the flexibility we have described comes with tradeoffs, and these are not inconsequential. The most obvious tradeoff is that modifying the SEC approach (especially the topics and cognitive demand levels, but even the response scales or the approach to weighting standards) makes the resulting data not comparable across contexts. In the vast majority of cases, where the SEC is used for a particular project in a particular context, this may be acceptable. In the rare cases where SEC data from multiple contexts and studies are used in a secondary data analysis (e.g., Polikoff 2012a, 2012b, 2013), this would be a fatal barrier. Regardless, these kinds of modifications would necessitate re-analyzing content standards with each new iteration of the SEC framework, something which may not be desirable. Given that the current primary usage of the SEC seems to be within-context, this concern may not be an issue. Allowing different numbers of cognitive demand levels in different subjects will limit the ability to compare results from one subject to another, but then the existing SEC also has different cognitive demand definitions across subjects (despite having the same number of levels), so it is not clear if this is a new tradeoff. Similarly, the existing SEC does sometimes allow adding or removing individual topics for particular studies, but even these modest modifications affect the comparability of data from study to study. Regardless, it is worth serious consideration from the researcher whether the loss of comparability of various kinds is too high a cost. In our case, it clearly was not, as we had no desire to compare the results of our intervention with those from any other study.

The research presented here also has important limitations. Perhaps the main limitation is that the SEC methodologies require many sets of assumptions about what the standards represent and how they will be interpreted by teachers; some experts may disagree with those assumptions. For instance, as discussed above, the methodology has typically treated the standards as equally weighted, but there is no "correct" weighting of the standards. The actual choice of one

weighting approach over another may matter more in terms of how it is explained to teachers than it may matter in terms of the alignment estimation or the resulting feedback given. For instance, if we compare a content analysis of the 4th grade math and 5th grade ELA Common Core standards analyzed using the old approach (each standard weighted equally) with an analysis using the new approach (each standard weighted by the number of cells it covers), the alignment values are quite high: .82 in mathematics, .80 in ELA. However, the new approach was much more understandable for both the coaches and the teachers, implying it may be a better choice even if it doesn't meaningfully change the alignment results.

It may not even be desirable for teachers to think about the standards' content topics nor cognitive demand levels in the discrete ways that are assumed. Importantly, the intervention we are creating does not ask teachers to only consider standards one-at-a-time. Quite the contrary, teachers receive coaching on how their instruction compares to the aggregated content emphasized across the whole standards document. The methodology may also be more challenging or less clear in cases where teachers teach to similar instructional targets day after day (e.g., reading comprehension over the course of a year), or in cases where earlier skills are repeatedly called up as newer skills are taught (e.g., the operations that are called up as students are solving algebra problems). Studying instructional content at scale is challenging and requires some degree of simplification if it is to be feasible, but there is always the risk that simplifying our measurement of instruction could lead to unintended negative consequences for the ways teachers think about and carry out instruction; these are important areas for new research.

There is a great deal more to know about teachers' coverage of standards-aligned content and about the content of important policy instruments like standards, assessments, and curriculum materials. A more flexible SEC could be an essential tool for carrying out this kind of

research. The strategies laid out here offer a path forward for the next generation of alignment research using the SEC.

References

Beach, R. W. (2011). Issues in analyzing alignment of language arts Common Core Standards with state standards. *Educational Researcher*, *40*(4), 179-182.

Cobb, P., & Jackson, K. (2011). Assessing the quality of the Common Core State Standards for mathematics. *Educational Researcher*, *40*(4), 183-185.

Cizek, G. J., Kosh, A. E., & Toutkoushian, E. K. (2018). Gathering and evaluating validity evidence: The generalized assessment alignment tool. *Journal of Educational Measurement, 55*(4), 477-512.

Desimone, L. M., & Le Floch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis*, *26*(1), 1-22.

Fulmer, G. W. (2011). Estimating critical values for strength of alignment among curriculum, assessments, and instruction. *Journal of Educational and Behavioral Statistics, 36*, 381–402.

Fulmer, G. W., & Polikoff, M. S. (2014). Tests of alignment among assessment, standards, and instruction using generalized linear model regression. *Educational Assessment, Evaluation, and Accountability, 26*(3), 225-240.

Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis*, *19*(4), 325-338.

Hamilton, L. S., & Berends, M. (2006). *Instructional practices related to standards and assessments (RAND Working Paper WR–374–EDU)*. Santa Monica, CA: RAND.

Holcombe, R., Jennings, J., & Koretz, D. (2013). The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation* (pp. 163–189). Greenwich, CT: Information Age Publishing.

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational Measurement: Issues and Practice*, *22*(2), 18-26.

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Yearbook of the National Society for the Study of Education*, *104*(2), 99-118.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79,* 1332–1361.

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

Polikoff, M. (in press). Alignment. In S. Brookhart (Ed.), *Routledge encyclopedia of education.*

Polikoff, M. S. (2015). How well aligned are textbooks to the Common Core Standards in mathematics? *American Educational Research Journal, 52*(6), 1185-1211.

Polikoff, M. S. (2013). Teacher education, experience, and the practice of aligned instruction. *Journal of Teacher Education, 64*(3), 212-225.

Polikoff, M. S. (2012a). The association of state policy attributes with teachers' instructional alignment. *Educational Evaluation and Policy Analysis, 34*, 278–294.

Polikoff, M. S. (2012b). Instructional alignment under No Child Left Behind. *American Journal of Education, 118,* 341–368.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3–14.

Polikoff, M. S., & Fulmer, G. W. (2013). Refining methods for estimating critical values for an alignment index. *Journal for Research on Educational Effectiveness, 6*(4), 380-395.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*, *36*(4), 399-416.

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal, 48*, 965–995.

Polikoff, M. S., Zhou, N., & Campbell, S. E. (2015). Methodological choices in the content analysis of textbooks for measuring alignment with standards. *Educational Measurement: Issues and Practice, 34*(3), 10-17.

Porter, A. C., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The new U.S. intended curriculum. *Educational Researcher, 40,* 103–116.

Porter, A. C., Polikoff, M. S., & Smithson, J. (2009). Is there a de facto national intended curriculum? Evidence from state content standards. *Educational Evaluation and Policy Analysis, 31*, 238–268.

Porter, A. C., Polikoff, M. S., Zeidner, T., & Smithson, J. (2008). The quality of content analyses of state student achievement tests and state content standards. *Educational Measurement: Issues and Practice*, *27*(4), 2-14.

Schwille, J., Porter, A. C., Alford, L., Floden, R., Freeman, D., Irwin, S., et al. (1988). State policy and the control of curriculum decisions. *Educational Policy, 2*(1), 29–50.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233–267). Bristol, PA: Taylor & Francis.

Time on Topic		*Grades K-12 Mathematics Topics*	*Performance Expectations for Students*				
<none>	9	**Data Displays**	**Memorize/ Recall**	**Perform Procedures**	**Demonstrate/ Communicate Understndg.**	**Conjecture / Analyze / Generalize**	**Integrate / Synthesize / Critique**
⓪ ① ② ③	901	Summarize data in a table or graph	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	902	Bar graphs and histograms	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	903	Pie charts and circle graphs	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	904	Pictographs	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	905	Line graphs	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	906	Stem and leaf plots	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	907	Scatter plots	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	908	Box plots	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	909	Line plots	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	910	Classification and Venn diagrams	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	911	Tree diagrams	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
<none>	10	**Statistics**	**Memorize/ Recall**	**Perform Procedures**	**Demonstrate/ Communicate Understndg.**	**Conjecture / Analyze / Generalize**	**Integrate / Synthesize / Critique**
⓪ ① ② ③	1001	Mean, median, and mode	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤
⓪ ① ② ③	1002	Variability, standard deviation, and range	p ① s ①	p ② s ②	p ③ s ③	p ④ s ④	p ⑤ s ⑤

Figure 1. A snippet of the existing (prior) SEC survey in mathematics

Teacher survey	cd A	cd B	cd C	cd D	cd E
topic 1	.083	.083	0	0	0
topic 2	.167	.083	0	.083	0
topic 3	0	.20	.20	.10	0

Content standards rater 1	cd A	cd B	cd C	cd D	cd E
topic 1	.02	.14	.09	0	0
topic 2	0	.11	.29	0	0
topic 3	.10	.05	0	0	.20

Content standards rater 2	cd A	cd B	cd C	cd D	cd E
topic 1	.18	.06	.01	0	0
topic 2	0	.29	.11	0	0
topic 3	.10	.15	0	.10	0

Content standards overall	cd A	cd B	cd C	cd D	cd E
topic 1	.10	.10	.05	0	0
topic 2	0	.20	.20	0	0
topic 3	.10	.10	0	.05	.10

Figure 2. Example content analysis data from a comparison of a teacher survey to a set of content standards that have been coded by two raters