Title	Refining methods for estimating critical values for an alignment index
Author(s)	Morgan S. Polikoff and Gavin W. Fulmer
Source	*Journal of Research on Educational Effectiveness, 6*(4), 380-395
Published by	Taylor & Francis (Routledge)

Refining Methods for Estimating Critical Values for an Alignment Index

Morgan S. Polikoff

Gavin W. Fulmer

In press at *Journal of Research on Educational Effectiveness*

Abstract

The alignment among standards, assessments, and teachers' instruction is an essential element of standards-based educational reforms. The Surveys of Enacted Curriculum (SEC) is the only common tool that can be used to measure the alignment among all three of these sources (Martone & Sireci, 2010). Prior SEC alignment work has been limited by not allowing for significance tests. A recent paper (Fulmer, 2011), provided a first attempt to address this shortcoming of the SEC, but that work was limited in several ways. We extend Fulmer's simulation approach by accounting for important elements of the SEC procedures, including the proper framework size, number of standards and assessment points, number of raters, rater cell-splitting rates, and rater agreement results. The results indicate that inferences about relative alignment may be heavily influenced by features of the alignment procedures. Thus, our method should be broadly applied to future SEC alignment investigations.

In the decade since the No Child Left Behind Act (NCLB) became law of the land, the concept of alignment has become part of the standard educational lexicon. Owing to the law's repeated emphasis on issues of alignment – the text of the law alone mentions alignment more than 40 times – researchers and practitioners in U.S. K-12 education have intensified their focus on understanding alignment among instruction, curriculum, standards, and assessments. While there are several well known procedures for examining test-standards alignment, only one of these procedures can be extended to examine alignment of instruction with standards and assessments (Martone & Sireci, 2009). This is the Surveys of Enacted Curriculum (SEC) approach (Porter, 2002), and it is the focus of this work.

A great deal of recent work has used the SEC to investigate alignment among instruction, standards, and assessments (e.g., Liu & Fulmer, 2008; Liu et al., 2009; Polikoff, 2012a, 2012b, in press; Porter, McMaken, Hwang, & Yang, 2011; Porter, Polikoff, & Smithson, 2009; Porter, Smithson, Blank, & Zeidner, 2007). The primary use of the SEC has been in estimating alignment indices on a 0 to 1 scale and comparing indices among individuals or among documents. Until recently, however, there was no way for researchers and practitioners to understand and make use of alignment indices other than norm-referenced approaches (e.g., the alignment of assessments with standards in State X is larger than that in State Y). This changed with a recent paper (Fulmer, 2011), which provided a first effort at estimating means and critical values for the SEC alignment index. Unfortunately, the approach used in Fulmer's (2011) paper, while instructive, relied on a number of assumptions about the SEC data that did not accurately characterize the SEC content analysis procedures. Thus, the results of that paper do not provide the information needed by SEC users to compare alignment indices with critical values.

In what follows, we expand the scope of Fulmer's (2011) approach to estimating critical values for alignment indices, in order to provide SEC users with the tools necessary to compare indices across studies and with critical values. First, we extend the Fulmer analysis by updating his results using the correct number of SEC content cells. The results indicate that the critical values in Fulmer's analysis are larger than would be expected under SEC frameworks. Next, we expand Fulmer's approach by using real data from recent content analyses and revising the assumptions of his analysis. Specifically, we revise assumptions about a) the number of content analysts, b) their rate of cell-splitting (i.e., dividing analyzed content into multiple SEC cells), and c) their rate of agreement. Third, we apply the procedures to six pairs of recently-analyzed standards and/or assessment documents and examine the extent to which these documents are more or less aligned with one another than would be expected due to chance. Finally, we discuss issues in estimating critical values for alignment indices involving teachers' instruction.

Background

The peer review criteria established under NCLB specify five necessary components of alignment that state assessments must meet: a) cover the full range of content specified in the standards, b) measure both the content and process components of the standards, c) reflect the same degree and pattern of emphasis in the standards, d) reflect the full range of cognitive complexity, level of difficulty, and depth of the standards, and e) yield results that represent all achievement levels specified in the performance standards. Based on these criteria, several alignment procedures are viewed as acceptable. However, the two most common alignment procedures in use by states today are the SEC approach (Blank, Porter, & Smithson, 2001; Porter, 2002) and the Webb alignment procedure (Webb, 1999, 2002). Of these, only the SEC

procedure also allows for the investigation of instructional alignment with standards and assessments (Martone & Sireci, 2009).

The Surveys of Enacted Curriculum

The SEC tools developed through two decades of research aimed at understanding teachers' decisions about what to teach and the effects of those decisions on student learning gains (e.g., Porter, 1989; Porter, Floden, Freeman, Schmidt, & Schwille, 1988; Porter, Kirst, Osthoff, Smithson, & Schneider, 1993; Schwille et al., 1982; Schwille, Porter, & Gant, 1980). The tools began as a set of instructional surveys and expanded to include techniques for content analyzing assessments, standards, and curriculum materials. The tools arrived in their present form in the early 2000s (Blank et al., 2001), with only slight modifications to the content languages since then. There are content languages for each of four subjects: mathematics, English language arts (ELA), science, and social studies. Each language defines content at the intersection of specific topics and levels of cognitive demand. The surveys are available online (http://seconline.wceruw.org/secwebhome.htm). We focus here on mathematics, ELA, and science, since these are the subjects for which the SEC is most commonly used and the only subjects required to be tested under NCLB.

A primary purpose of the tools is in estimating alignment among instruction, standards, assessments, and/or curriculum materials. Instructional surveys are based on teacher self-report of content coverage over a given period of time. For instance, teachers might be surveyed twice a year and the results aggregated to represent a full year's instruction. On each survey, teachers indicate the number of lessons they spent on each topic in the list. For each topic covered, teachers then indicate the proportion of their content coverage focused on each of five levels of cognitive demand. The five levels of cognitive demand differ slightly across subjects but

generally range from memorization and procedures to application and generalization. Figure 1 shows a sample of the paper version of the SEC mathematics teacher survey; the tool can also be completed online. For more details on the teacher survey techniques, see (Porter et al., 2007).

Instruction can be compared with the content of other documents. Most commonly, these documents are state standards or assessments, but they could also be textbooks, pacing guides, or many other materials. The content analysis procedures involve the use of multiple, trained content analysts. Generally, four analysts are used, but the number can be more or less. Each content analyst examines the document at the finest-grained level of detail available: for assessments, items; for standards, objectives; for textbooks, lessons, pages, or section headers. For each piece of material analyzed, the analyst determines the SEC cells (i.e., topics and cognitive demands) that are covered. Multiple cells are allowed because items, objectives, and lessons often tap multiple skills. The weight for each piece of material is then evenly allocated across the chosen cells (e.g., if a two-point test item is allocated across three cells, each cell receives two-thirds of a point)[1]. Finally, each rater's overall content analysis is calculated and the raters' analyses are averaged.

Estimating Alignment

With the surveys or content analyses complete, the next step is to estimate alignment indices. The formula provided by Porter (2002) is

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

[1] Unless otherwise specified, each objective, item, or lesson is assumed to be equally weighted. While it is possible that some objectives might be more important than others, there is rarely any indication in the document itself that this is so. Thus, using a system that does not include equal weighting would require agreement among content analysts as to the relative importance of each objective in each document, a task that is likely to be quite unreliable and challenging. Equal weighting is therefore the most conservative and defensible approach.

where x_i is the proportion of points in the *i*th cell of one table (e.g., a standards document), and y_i is the proportion of points in the *i*th cell of the other table (e.g., an assessment). Here, *i* denotes the cell number ranging from 1 to N, the total number of cells in the framework (i.e., N = J × K for a framework with J rows and K columns). The alignment index ranges from 0 to 1. It can be shown by proof that the formula is equivalent to the sum of the cell-by-cell minima in the two content analyses. The resulting value indicates the proportion of content in exact, proportional, cell-by-cell agreement between the two sources. For instance, an alignment index of .50 between instruction and the standards indicates that 50% of instructional content is on SEC cells that are covered in the same proportions in the standards. Another interpretation is that 50% of the standards content is on SEC cells that are covered in the same proportions in instruction.

The alignment index has been used in a number of published articles that compare instruction, standards, or assessments. A sample of these results is discussed here, but the interested reader is referred to the original sources for more details. Several articles have compared the content of standards with standards and/or assessments using the alignment approach (Polikoff, Porter, & Smithson, 2011; Porter, 2002; Porter et al., 2011; Porter et al., 2009). For instance, Polikoff et al., (2011) examined the alignment of state assessments with state content standards in the NCLB era, finding average alignment indices of .19 in ELA, .27 in mathematics, and .26 in science. The authors interpreted these findings as indicating poor average test-to-standards alignment in the NCLB era. Another study (Porter et al., 2009) examined the alignment of content standards across states within grades, asking if there was evidence of a de facto national intended curriculum. The average alignment indices for that analysis were .20 to .27 depending on subject and grade. These results were taken to indicate the lack of a de facto national intended curriculum.

The alignment index has also been used in several studies involving the alignment of teachers' instruction with state standards and/or assessments (Polikoff, 2012a, 2012b; Porter et al., 2007) or of teachers' instruction with other teachers' instruction (Polikoff, in press; Porter, 2002). An experimental study involving an intervention that gave teachers information about their enacted curriculum was found to have significant impacts on instructional alignment to mathematics standards (Porter et al., 2007). Two studies of instructional alignment to standards and assessments under NCLB found that teachers have generally increased their alignment by one-quarter to one-half a standard deviation, and that certain state policy attributes were significant predictors of teachers' alignment (Polikoff, 2012a, 2012b). However, these studies also found that average alignment indices across the sample of more than 30,000 teachers were low, with means in the range of .14 to .28 depending on subject and grade. In contrast, a study that compared the "average" instruction of eighth grade mathematics teachers across states found alignment indices between .56 and .84; these findings were taken to indicate that eighth grade math instruction is mainly similar across states (Porter, 2002).

Giving Meaning to the Alignment Index

While these studies are instructive, there are clear limitations of the existing work (Fulmer, 2011). For one, there have been no available criteria for describing the strength of alignment based on the alignment index. For another, there has been limited opportunity to conduct hypothesis testing on alignment indices, owing to the lack of available information on critical values. As such, researchers using the alignment indices have often relied on normative comparisons only – for instance, by describing distributions of alignment indices and comparing results across subject areas (Polikoff et al., 2011). Having objective criteria for describing alignment indices would allow alignment researchers to classify alignment results in terms of

their magnitude and would also allow researchers to conduct significance tests comparing alignment indices. These results would increase the relevance and utility of alignment research and help make it more accessible to quantitatively-oriented researchers who are not conversant in the language of the alignment methodologies.

As a first step in addressing this gap in the literature, Fulmer (2011) estimated critical values for the alignment index and used his results to re-analyze the extant literature on alignment. He identified two features that affected the magnitude of typical alignment values – the number of cells in the content language and the number of "points" in the standards and assessment documents. Next, he conducted a series of simulations that varied along these dimensions, estimating a distribution of alignment indices for each variation. Given these distributions, he estimated means and critical values. For the number of cells in the framework, Fulmer used values between 10 and 114. For the number of points in the standards, he varied from 30 to 120. For the sake of simplicity, he assumed 100 points on the assessment. Based on the means and critical values Fulmer obtained, he discussed results of the prior alignment literature. He found that most of the studies of standards-assessment alignment had alignment values well below the means in his estimated distributions.

While Fulmer's analysis provided an important first step at developing a deeper understanding of the alignment index and its properties, it was limited in a number of important ways that reduce the utility of the results he described. One important limitation was that the frameworks in the simulations contained just 10 to 114 cells, while the SEC frameworks in ELA, mathematics, and science contain 665, 915, and 1055 cells, respectively. A second limitation was that the simulations did not account for multiple content analysts per document (generally three or four). A third limitation was that the simulations did not allow for the fact that content

analysts can split their codes among multiple cells in the SEC taxonomy. A fourth limitation was that the simulations did not account for the agreement of multiple raters, which research indicates can vary from document to document (Porter, Polikoff, Zeidner, & Smithson, 2008). A final limitation was that the simulations did not use actual SEC data; thus, the previous results are not of practical value to SEC users.

Clearly, addressing these limitations makes the estimation of means and distributions for alignment indices dramatically more complicated than was the case in Fulmer's (2011) analysis. Thus, the purpose of this article is to present a general method for using actual SEC data to estimate means and critical values for alignment indices. This article's contributions are 1) for the first time in published research, to describe each of the design features related to the magnitude of the Porter alignment index and 2) to present a method for estimating and interpreting strength of alignment that more accurately reflects these design features. Thus, the findings can be more readily applicable to use in analyses of alignment among instruction, standards, and assessments than any previously produced. Accompanying this method is an annotated Stata do file that can be edited based on the actual SEC data under consideration. The results can be used by SEC researchers during their analyses to examine the extent to which identified alignment indices are high, average, or low relative to what would be expected given content analysis data like theirs.

Data and Methods

The majority of data for the analysis is created by simulation. Specifically, we use the *bsample* (sampling with replacement) command in Stata SE version 12. Stata uses a pseudo-random number generator that has passed the DIEHARD tests used for judging the quality of pseudo-random number generators (http://www.stata.com/support/cert/diehard/index.html). Let

$N = J \times K$ represent the number of cells in a framework with J rows and K columns. Using the bsample command, we randomly allocate the given number of points to the N cells and calculate the alignment index (using the equation produced above, from Porter [2002]). We replicate this 2000 times and estimate the mean, standard deviation, and percentiles of the resulting distribution of alignment indices. The choice of 2000 replications is based on preliminary comparisons of estimations with 2000, 5000, and 10000 iterations, which yielded equivalent results to 3 decimal places. Finally, we replicate all of the above steps 10 times using different random number generator seeds to ensure the results are not merely reflective of the particular seed chosen for the random number generator. Given the 10 means, standard deviations, and percentile ranks (each based on 2000 random draws), we average these to determine the overall mean, standard deviation, and percentile ranks of estimated alignment indices. As this approach presents an application of hypothesis testing of an observed alignment versus the range of possible but unobserved values, the methodology is non-parametric in nature and does not meet the conditioning assumptions underlying the likelihood principle (cf. Edwards, 1974; Robins & Wasserman, 2000). Even so, testing the hypothesis that an observed alignment index is higher or lower than could be obtained by chance has important implications for policy analyses that examine the extent to which standards, assessments, and instruction are aligned.

Four small studies are reported. In all estimations, the number of cells used is based on previous SEC studies: 655, 915, or 1055, for ELA, math, and science, respectively. Similarly, the estimations use the SEC's typical range of numbers of test items (40 to 80) and objectives (30 to 60) in content analyzed documents.

In Study 1, we replicate the method used in Fulmer (2011), correcting only for the appropriate number of cells in the SEC frameworks and the typical range of numbers of test

items (40 to 80) and objectives (30 to 60) in content analyzed documents. This study assumes a single content analyst and no cell-splitting.

In Study 2, we extend and improve on Fulmer's estimated means and standard deviations by illustrating the dependence of critical values on the number of content analysts involved in coding. Because we are still assuming that raters operate independently of one another, this analysis is straightforward: using average-sized documents we simply double, triple or quadruple (for two, three, or four raters) the number of points on the standards and assessment documents in the simulation. For instance, if we are comparing a 45-objective standards document to a 60-item test and we assume four raters, we simply randomly select (with replacement) 180 SEC cells to represent the standards and 240 cells to represent the assessment. The resulting simulated content analyses are equivalent to having randomly selected 45 and 60 cells four times and averaging the four simulations, as would be done in actual content analysis.

In Study 3, we use data on the typical rate of cell-splitting by content analysts to properly account for the fact that objectives and test items often tap multiple topics or levels of cognitive demand. First, using actual SEC data, we characterize high, medium, and low rates of cell splitting for standards and assessment documents—that is, the number of items or objectives placed into 1, 2, 3, 4, 5, and 6 cells. Next, we use these typical rates to weight the randomly sampled SEC cells accordingly. Thus, for the n items placed into one cell each, we randomly select n cells and weight them 1 each; for the m items placed into two cells each, we randomly select 2m cells and weight them ½ each; and so on. For instance, suppose there is a 10 item assessment with one content analyst and we determine that 50% of items are typically split between two cells. In this case, the simulation procedure would randomly select (with replacement) five SEC cells and allocate them 10% each to represent the non-split items. Then

the simulation would randomly select 10 SEC cells and allocate them 5% each to represent the split items. The procedure is analogous for the more complicated, four-rater system used by the SEC.

In Study 4, we illustrate the dependence of the estimated distributions and critical values on the agreement rate of content analysts. First, using actual SEC content analysis data, we characterize high and low rates of rater agreement for standards and assessments. Next, we use these typical rates to adjust the simulation procedure to account for agreement. Given an n-item test with r raters, there are nr SEC cells to be randomly selected. However, if there were perfect agreement across raters on the content of each item, only n cells would be randomly selected. Thus, if there is y% agreement, only nr – (nr – n)*y/100 cells would be randomly selected, and these would be weighted appropriately. For instance, suppose again there is a ten-item test, this time with two analysts and no cell-splitting, and we find that half the items have perfect agreement (analysts place items in exactly the same SEC cells) and half have perfect disagreement. In this case, the simulation procedure would randomly select five SEC cells and allocate them 10% each to represent the perfect agreement items. Then the simulation would randomly select 10 SEC cells and allocate them 5% each to represent the perfect disagreement items. The procedure is analogous for the more complicated, four-rater system used by the SEC with varying rates of cell-splitting.

For each of these improvements, we present a table to indicate how accounting for these aspects of the SEC alignment coding process affects the resulting estimated distributions. We also discuss how our results are different for the comparison of content standards or assessments with teachers' instruction. To conclude, we use real content analysis data from recently-analyzed state and district assessments to apply our approach.

Results

The results are presented for each of the four studies conducted. Throughout these, we adapt Fulmer's analysis to the appropriate number of cells in the SEC frameworks for English, mathematics, and science. We also use more accurate values for the number of points in the typical standards and assessment document. We present the means and standard deviations of the estimated alignment distributions, as well as several critical percentile ranks.

Study 1 -- Sensitivity to Cell Size

Study 1 replicated Fulmer's (2011) analysis with corrected values that reflected the appropriate SEC tables in terms of size, as well as number of standards and assessment points. Table 1 presents the results of these analyses, which assume one content analyst and no cell-splitting. In Fulmer's (2011) analysis, the largest table size was 114 cells. The expected mean values of the alignment index for a table of this size ranged from .390 to .585, depending on the number of standards and assessment points. In contrast, when using the larger SEC table sizes and assuming one content analyst, we find that the estimated mean alignment values are between .027 and .083, depending on the number of points in the standards and assessment documents. This concurs with Fulmer's finding that increasing table size would decrease the expected mean alignment, when other factors are held constant. Also as expected, the estimated mean alignment values are smaller when there are fewer standards and assessment points. However, when there are few standards points (30) it does not appear to matter how many assessment points there are – mean alignments are low regardless. The estimated means are also smaller for larger tables – approximately two-thirds as large for science (1055 cells) as for English (665 cells).

The distributions of the predicted alignment indices are also presented in Table 1. The range from the 0.5^{th} percentile to the 99.5^{th} percentile is between .09 and .16 for each standards-

assessment pair. There is no apparent pattern to the spread of the distributions – the standard deviation is approximately as large for the standards and assessments with the most points as it is for those with the least points. Even the largest value in the table – the 99.5th percentile for English standards and assessments with 60 and 80 points, respectively – is substantially smaller than the smallest value that was found in Fulmer's analysis. Clearly, after adjusting for the appropriate number of cells in the SEC alignment frameworks, the means and critical values are shifted much closer to zero.

Study 2 -- Number of Raters

Study 2 focuses on the inclusion of additional raters, which is common in applications of the SEC alignment framework. Using an average number of standards (45) and assessment (60) points, we allow the number of raters to vary from one to four. We assume that raters operate completely independently, but that they place each objective or assessment item into no more than one SEC cell (i.e., no cell splitting). The results of these analyses are shown in Table 2. As expected, more raters are associated with greater mean alignment indices. For instance, in mathematics the estimated mean for one rater is .047; two raters, .090; three raters, .129; and four raters, .165. For the other two subjects as well, the estimated means increase by a factor of roughly 3.5 in moving from one to four raters. The standard deviations, in contrast, decrease by approximately one-sixth in moving from one rater to four raters. Thus, these results indicate that more raters are associated with greater and more homogeneous estimated alignment indices. Still, even with four raters the estimated means are substantially lower than those reported in Fulmer's analysis.

Study 3 -- Cell Splitting

Study 3 adjusts the simulation to account for cell-splitting during the rating process, which occurs when content analysts rate objectives or test items into more than one cell. Table 3 shows the low, average, and high rates of cell-splitting in content analysis based on the documents used for this paper. As is evident, the rates of cell-splitting vary considerably from document to document. On average, test items are placed into just one cell more often (40% to 82%) than are objectives (25% to 62%). The table is read as follows: a target assessment with a low rate of cell splitting will result in 82% of items placed into one SEC cell, 16% placed into two cells, and 2% placed into three cells. Given the wide disparities in cell splitting rate and their potential effects on alignment index distributions, it is important to account for cell splitting in our simulations.

Table 4 uses four content analysts and an average-sized standards and assessment document (i.e., 45 standards points and 60 assessment points) and varies the simulations based on low, average, and high levels of cell splitting. As in all previous simulations, the largest mean values are for the smallest framework—ELA. Within each subject, as the rate of cell-splitting increases, the mean estimated alignment indices increase. Moving from low to medium or medium to high cell splitting is associated with a .03 to .04 increase in the estimated mean alignment indices. Furthermore, the distributions become tighter as cell splitting increases, with standard deviations decreasing by 15-16% depending on subject as cell splitting increases from low to high.

Study 4 -- Rater Agreement

Study 4 examines the effect of rater agreement on alignment. Up to this point, we have assumed that raters operate completely independently. That is, we have assumed that agreement among raters as to the placement of SEC items is random (and therefore quite unlikely, since

there are several hundred cells in the SEC frameworks). However, this assumption is not an accurate representation of real rater agreement – in practice, the reliability of content analysts across raters is moderate to high (Porter et al., 2008). The more agreement there is among raters, the fewer SEC cells appear to be covered by the document, and the more weight is allocated to each of these cells. For instance, in the extreme case of perfect agreement, the content analysis for any number of raters appears the same as the content analysis for a single rater.

To investigate this issue, we first calculated the agreement rates for the documents reanalyzed for this paper. For each item/objective analyzed, we determined the proportion of that item's content that is in agreement across raters using an approach similar to the alignment index. For instance, suppose an item was placed into cell x by Rater 1, cells x and y by Rater 2, and cells y, z, and w by Rater 3. In this case, the agreement for Raters 1 and 2 is 50%; for Raters 1 and 3, 0%; and for Raters 2 and 3, 33%. Thus the agreement rate for this item is the average of 50%, 0%, and 33%, which is 27%. Next, we simply averaged the agreement rates for each item to obtain a document-level agreement rate. For the documents in our sample, the agreement rates for standards range from approximately 30% to 90% with a mean of 40%, and the agreement rates for assessments range from 20% to 83% with a mean of 60%.

Table 5 uses four raters, average sized documents (i.e., 45 standards points, 60 assessment points), and an average level of cell-splitting (as defined in Table 3). We vary the simulations based on agreement rate (low and high agreement for standards, low and high agreement for assessments, as described above). The results indicate that expected alignment values decrease with increasing levels of rater agreement. Even for low levels of agreement, the mean estimated alignment indices are .05 to .07 lower than in Table 4 (where random agreement was assumed). Moving from low agreement on both documents to high agreement on both

documents decreases estimated mean alignment indices further still—by approximately .05 to .07, depending on subject. For documents where raters have high agreement, the estimated alignment distributions are similar to those in Table 2 with three raters and no cell-splitting, indicating that rater agreement substantially affects estimated alignment distributions.

Application to SEC Data

We now apply the improvements in estimated critical values to alignment values observed in actual SEC data. We present two examples per subject, chosen from the set of recently content-analyzed standards and assessment documents in the SEC database. These examples were chosen to illustrate the kinds of alignment analyses typically conducted using SEC data. In some cases the names of states are withheld due to confidentiality agreements with participating states regarding the content of their assessments. All results are displayed in Table 6. To conserve space, the figure in the "cell splitting" column represents the average number of SEC cells each item or objective is placed into by each rater.

In ELA we present two comparisons. The first row represents the alignment of State V's sixth grade standards (Document A) with its sixth grade assessment (Document B), similar to the alignment indices presented in Polikoff, Porter, & Smithson (2011). Given the characteristics of these two content analyses, the expected distribution of the alignment indices due to chance has a mean of 0.228 and a standard deviation of 0.022. We would expect 95% of the alignment indices from this chance distribution to fall in the range 0.185 to 0.272. Thus, the actual value of 0.251 is within the 95% confidence range, and we can say that the actual alignment of these two documents is not significantly higher than would be expected due to chance. The second row represents the alignment of Arizona's fourth grade ELA standards with the fourth grade Common Core State Standards in ELA, similar to the alignment indices presented in Porter et al. (2011).

Here, the actual alignment value of 0.188 is significantly below the mean of the expected distribution of alignment indices.

In mathematics, the first comparison is between New York's fourth grade mathematics standards and the fourth grade Common Core State Standards in mathematics. The expected alignment distribution based on the characteristics of these content analyses has a mean of 0.126 and a 95% range of 0.088 to 0.167. The actual alignment index of 0.204 is significantly higher than the mean expected by chance. The second comparison in mathematics is between the high school geometry standards and end-of-course test in State K. The expected mean of the distribution due to chance is 0.091, with a standard deviation of 0.020. The actual alignment value of 0.380 is substantially and significantly higher than the expected mean, suggesting this state assessment is particularly well aligned with its standards relative to what would be expected due to chance.

Finally, we present two comparisons in science. The top row is for the alignment of State M's eighth grade science standards with its science assessment. Again, the actual alignment index of 0.226 is significantly above the mean alignment index expected due to chance (0.177). The second row in science is for the alignment of Indiana's fourth grade science standards with the fourth grade NAEP assessment item bank. Here, the alignment index of 0.215 is significantly higher than the expected alignment index due to chance (0.137).

The comparisons in Table 6 illustrate not only the alignment indices for particular document pairs, but also some important points about the need for examining critical values of the alignment indices. First, the expected distributions of the six alignment indices differ quite substantially. As discussed earlier, the expected means are higher in ELA than in the other two subjects. However, in these real-world applications, the expected means are lower in

mathematics than in science, owing to the smaller number of standards and assessment points in the mathematics documents and the greater rater agreement in mathematics. Second, the rank-ordering of the actual alignment values does not generally conform to their ordering after accounting for chance agreement. For instance, the second most aligned documents based on the raw alignment values are State V's ELA standards and assessment. However, this document pair is the fifth most aligned document pair after accounting for chance agreement, and it is not even significantly more aligned than would be expected due to chance. Both of these findings clearly illustrate the need to consider critical values of alignment distributions when making relative statements among document pairs regarding alignment.

Discussion

The present study extended and improved upon previous efforts to estimate mean and critical values of indices of alignment (Fulmer, 2011), by accounting for rating conditions used in application of the SEC content analysis process—large table sizes, multiple raters, cell-splitting, and rater agreement. Our results show that, for the larger table sizes used in the SEC, typical alignment values are indeed much lower than previously reported for smaller tables. On the other hand, given these large table sizes, the mean alignment index increases with each additional rater and when more items or objectives are placed into multiple SEC cells. However, the mean alignment index decreases with increasing rater agreement. Furthermore, when multiple raters and cell-splitting are included, the standard deviations in the simulated alignment indices decrease, indicating less variation in the simulated ratings.

We interpret these findings as an improvement upon Fulmer's (2011) simulations by accounting for the actual rating conditions used in the SEC; but, the results also show marked consistency with the earlier report. For example, Fulmer argues that the average alignment index

that could be obtained by chance will be lower when there are larger tables and higher when there are more points to be ascribed in each table. The present work is consistent with this prediction; the SEC tables are much larger than some of the studies Fulmer cited (e.g., Liu et al 2009; Liu & Fulmer, 2008), and so the simulated alignment indices observed here are noticeably lower than those Fulmer (2011) reported. Furthermore, the inclusion of additional raters or of cell-splitting is equivalent to including additional points for each respective table, so including more raters or having greater rates of cell-splitting increases the mean alignment index that could be obtained by chance.

These findings also show that, contrary to Fulmer's (2011) study, the values reported from previous SEC work are not necessarily lower than could be obtained by chance, and in many cases are higher. Furthermore, the findings show that interpretations about the relative levels of alignment across pairs of documents are influenced by the expected distributions. These expected distributions, in turn, are influenced by features of the standards or assessment documents and the content analyses—including the number of points in each document, the number of content analysts, their rate of cell splitting, and their agreement rate. Given these findings, it would be useful to reanalyze the data from recent alignment studies (e.g., Polikoff et al., 2010; Porter et al., 2009, 2011) to investigate the extent to which the reported results change after accounting for chance agreement.

Another important extension of the work is to the investigation of critical alignment values for the alignment of instruction with standards or assessments. The SEC is the only widely-used tool that can capture the alignment of instruction with standards or assessments (Martone & Sireci, 2010). These instructional alignment indices have been used in several recent studies (e.g., Polikoff, 2012a, 2012b; Porter et al., 2011). This extension should be

straightforward – for each teacher-standard pair, estimated alignment distributions could be simulated in much the same way as presented above. The teacher's instruction could first be characterized by the number of SEC cells they report teaching in the year and the concentrations of instructional time spent on these cells, and the simulation code could be easily changed to account for the difference. Indeed, the simulation script would be more straightforward in this case than for the standards-assessment alignment indices studies in this paper.

Alignment among standardized tests, content standards, and instruction is an important topic that influences the validity of inferences made about the ways that such tests measure students' attainment of educational standards, and in the evaluation and interpretation of teachers' instruction that can meet these standards and support students' achievement. Our work presents an approach to strengthen the quality and utility of alignment research that should be incorporated in future alignment work.

References

Blank, R., Porter, A. C., & Smithson, J. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science.* Washington, DC: Council of Chief State School Officers.

Edwards, A. W. F. (1974). The history of likelihood. *International Statistical Review, 42*(1), 9-15.

Fulmer, G. W. (2011). Estimating critical values for strength of alignment among curriculum, assessments, and instruction. *Journal of Educational and Behavioral Statistics, 36*(3), 381-402.

Liu, X., & Fulmer, G. W. (2008). Alignment between science curriculum and assessments in selected New York State Regents exams. *Journal of Science Education and Technology, 17*, 373–383.

Liu, X., Zhang, B. H., Liang, L. L., Fulmer, G. W., Kim, B., & Yuan, H. Q. (2009). Alignment between the physics content standards and standardized tests: A comparison among US-NY, Singapore, and China-Jiangsu. *Science Education, 93*, 777–797.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332-1361.

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

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

Polikoff, M. S. (in press). The redundancy of mathematics instruction in U.S. elementary and middle schools. *Elementary School Journal*.

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*(4), 965-995.

Porter, A. C. (1989). A curriculum out of balance: The case of elementary mathematics. *Educational Researcher, 18*(5), 9-15.

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

Porter, A. C., Floden, R., Freeman, D., Schmidt, W., & Schwille, J. (1988). Content determinants in elementary school mathematics. In D. A. Grouws & T. J. Cooney (Eds.), *Perspectives on research on effective mathematics teaching*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Porter, A. C., Kirst, M. W., Osthoff, E. J., Smithson, J., & Schneider, S. A. (1993). *Reform up close: An analysis of high school mathematics and science classrooms.* Madison, WI: Wisconsin Center for Education Research.

Porter, A. C., McMaken, J., Hwang, J., & Yang, R. (2011). Common Core Standards: The new U.S. intended curriculum. *Educational Researcher, 40*(3), 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*(3), 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.

Porter, A. C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*(1), 27-51.Robins, J., & Wasserman, L. (2000). Conditioning, likelihood, and coherence: A review of some foundational concepts. *Journal of the American Statistical Association, 95*(452), 1340-1346.

Schwille, J., Porter, A. C., Belli, G., Floden, R., Freeman, D., Knappen, L., et al. (1982). *Teachers as policy brokers in the content of elementary school mathematics: Research series no. 113*. East Lansing, MI: Institute for Research on Teaching, Michigan State University.

Schwille, J., Porter, A. C., & Gant, M. (1980). Content decision making and the politics of education. *Educational Administration Quarterly, 16*(2), 21-40.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states. Research Monograph No. 18.* Madison, WI: National Institute for Science Education.

Webb, N. L. (2002). *Alignment study of language arts, mathematics, science, and social studies of state standards and assessments in four states.* Washington, DC: Council of Chief State School Officers.

Time on Topic		*Grades K-12 Mathematics Topics*	*Expectations for Students in Mathematics*				
<none>	1	**Number Sense/Properties/Relationships**	**Memorize Facts/ Definitions/ Formulas**	**Perform Procedures**	**Demonstrate Understanding of Mathematical Ideas**	**Conjecture/ Generalize/ Prove**	**Solve Non-Routine Problems/Make Connections**
⓪ ① ② ③	101	Place value	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	102	Whole numbers and integers	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	103	Operations	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	104	Fractions	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	105	Decimals	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③
⓪ ① ② ③	106	Percents	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③	⓪ ① ② ③

Figure 1. Example of the SEC mathematics survey.

Table 1

Estimated Alignment Distributions Assuming One Content Analyst and No Cell-Splitting

				Percentiles								
Standards Points	Test Points	Mean	SD	0.5	2.5	5	16	50	84	95	97.5	99.5
English (665 Cells)												
30	40	0.043	0.031	0.000	0.000	0.000	0.005	0.050	0.075	0.100	0.109	0.143
30	80	0.044	0.023	0.000	0.002	0.013	0.025	0.038	0.063	0.088	0.095	0.111
60	40	0.057	0.029	0.000	0.008	0.017	0.033	0.050	0.083	0.109	0.118	0.143
60	80	0.083	0.030	0.015	0.026	0.038	0.051	0.080	0.113	0.137	0.145	0.166
Math (915 Cells)												
30	40	0.032	0.027	0.000	0.000	0.000	0.000	0.025	0.050	0.075	0.100	0.118
30	80	0.032	0.020	0.000	0.000	0.000	0.013	0.025	0.050	0.065	0.075	0.091
60	40	0.042	0.025	0.000	0.000	0.000	0.017	0.034	0.067	0.084	0.100	0.118
60	80	0.062	0.026	0.006	0.013	0.025	0.038	0.063	0.088	0.108	0.115	0.138
Science (1055 Cells)												
30	40	0.027	0.026	0.000	0.000	0.000	0.000	0.025	0.050	0.075	0.083	0.108
30	80	0.028	0.018	0.000	0.000	0.000	0.013	0.025	0.050	0.063	0.068	0.086
60	40	0.036	0.024	0.000	0.000	0.000	0.017	0.033	0.066	0.083	0.085	0.110
60	80	0.054	0.025	0.000	0.013	0.013	0.025	0.050	0.076	0.099	0.104	0.125

Table 2

The Effect of Number of Raters on Alignment Distributions for Typical Standards and Assessment Documents

			Percentiles								
Raters	Mean	SD	0.5	2.5	5	16	50	84	95	97.5	99.5
					English (665 Cells)						
1	0.064	0.030	0.000	0.017	0.017	0.033	0.067	0.098	0.117	0.130	0.151
2	0.120	0.028	0.053	0.068	0.075	0.092	0.119	0.147	0.168	0.176	0.196
3	0.170	0.026	0.105	0.121	0.128	0.144	0.170	0.195	0.212	0.221	0.238
4	0.214	0.024	0.155	0.168	0.175	0.190	0.214	0.238	0.253	0.261	0.278
					Math (915 Cells)						
1	0.047	0.027	0.000	0.000	0.005	0.017	0.050	0.069	0.098	0.102	0.125
2	0.090	0.025	0.031	0.043	0.050	0.067	0.091	0.115	0.132	0.141	0.160
3	0.129	0.023	0.072	0.085	0.092	0.106	0.129	0.152	0.169	0.176	0.191
4	0.165	0.022	0.110	0.123	0.129	0.143	0.165	0.188	0.202	0.209	0.225
					Science (1055 Cells)						
1	0.041	0.025	0.000	0.000	0.000	0.017	0.033	0.067	0.083	0.100	0.118
2	0.079	0.024	0.025	0.034	0.042	0.056	0.078	0.102	0.119	0.127	0.144
3	0.114	0.022	0.060	0.072	0.078	0.091	0.113	0.136	0.151	0.160	0.175
4	0.146	0.021	0.094	0.105	0.112	0.125	0.146	0.167	0.182	0.189	0.202

Table 3

Typical Rates of Cell-Splitting for SEC Content Analyses

	Rate of Cell-Splitting		
Cells	High	Average	Low
	Tests		
1	0.40	0.56	0.82
2	0.40	0.34	0.16
3	0.20	0.10	0.02
4	0.00	0.00	0.00
5	0.00	0.00	0.00
6	0.00	0.00	0.00
	Standards		
1	0.25	0.40	0.62
2	0.32	0.32	0.26
3	0.20	0.13	0.08
4	0.11	0.09	0.02
5	0.04	0.03	0.01
6	0.08	0.03	0.01

Table 4

The Effect of Cell Splitting on Alignment Distributions for Typical Standards and Assessment Documents

Cell Splitting Rate	Mean	SD	Percentiles								
			0.5	2.5	5	16	50	84	95	97.5	99.5
English (665 Cells)											
Low	0.245	0.022	0.191	0.203	0.210	0.224	0.245	0.267	0.282	0.289	0.303
Average	0.287	0.020	0.237	0.249	0.255	0.267	0.287	0.306	0.320	0.326	0.338
High	0.322	0.018	0.276	0.287	0.292	0.304	0.322	0.341	0.352	0.358	0.369
Math (915 Cells)											
Low	0.191	0.020	0.140	0.152	0.158	0.171	0.191	0.211	0.224	0.231	0.244
Average	0.227	0.018	0.180	0.191	0.197	0.208	0.226	0.245	0.257	0.264	0.276
High	0.258	0.017	0.214	0.225	0.231	0.241	0.258	0.275	0.287	0.292	0.303
Science (1055 Cells)											
Low	0.170	0.020	0.121	0.132	0.138	0.150	0.169	0.189	0.202	0.208	0.221
Average	0.203	0.018	0.158	0.169	0.174	0.185	0.203	0.221	0.232	0.238	0.248
High	0.233	0.017	0.191	0.200	0.205	0.216	0.232	0.249	0.260	0.266	0.276

Note. All analyses assume 45 standards points, 60 assessment points, and four raters.

Table 5

The Effect of Rater Agreement on Estimated Alignment Distributions

Standards Agreement	Test Agreement	Mean	SD	Percentiles 0.5	2.5	5	16	50	84	95	97.5	99.5
English (665 Cells)												
Low	Low	0.222	0.022	0.167	0.179	0.187	0.200	0.222	0.243	0.258	0.265	0.280
High	Low	0.201	0.022	0.145	0.157	0.165	0.179	0.201	0.223	0.238	0.245	0.260
Low	High	0.163	0.022	0.108	0.121	0.128	0.141	0.163	0.186	0.201	0.209	0.223
High	High	0.154	0.024	0.096	0.109	0.116	0.131	0.154	0.178	0.195	0.203	0.218
Math (915 Cells)												
Low	Low	0.172	0.020	0.123	0.133	0.139	0.151	0.171	0.192	0.206	0.212	0.226
High	Low	0.154	0.020	0.104	0.116	0.122	0.134	0.154	0.175	0.189	0.196	0.210
Low	High	0.123	0.020	0.075	0.086	0.091	0.104	0.123	0.144	0.157	0.164	0.179
High	High	0.117	0.021	0.065	0.077	0.083	0.095	0.116	0.138	0.153	0.160	0.174
Science (1055 Cells)												
Low	Low	0.152	0.019	0.106	0.117	0.122	0.133	0.152	0.171	0.185	0.191	0.203
High	Low	0.137	0.019	0.091	0.100	0.106	0.117	0.136	0.156	0.169	0.176	0.188
Low	High	0.109	0.019	0.063	0.073	0.078	0.090	0.108	0.128	0.141	0.148	0.162
High	High	0.103	0.020	0.055	0.065	0.070	0.082	0.102	0.123	0.137	0.145	0.160

Note. All analyses assume 45 standards points, 60 assessment points, four raters, and average levels of cell-splitting.

Table 6

Simulated and Actual Alignment Indices for Six Document Pairs in Three Subjects

Document A				Document B					Simulated Distributions						
Agreement	Cell-Splitting	Points	Raters	Agreement	Cell-Splitting	Points	Raters	Actual Alignment	Mean	SD	2.5	16	50	84	97.5
0.31	2.23	47	3	0.21	2.16	46	3	0.251	0.228	0.022	0.185	0.206	0.228	0.250	0.272
0.44	2.99	107	4	0.83	2.08	106	3	0.188	0.304	0.017	0.271	0.287	0.304	0.321	0.337
0.64	1.72	100	3	0.90	3.51	35	4	0.204	0.126	0.020	0.088	0.106	0.126	0.146	0.167
0.76	2.01	29	4	0.75	1.53	75	4	0.380	0.091	0.020	0.054	0.071	0.091	0.112	0.133
0.42	3.27	49	4	0.37	1.58	55	4	0.226	0.177	0.018	0.142	0.159	0.177	0.195	0.213
0.64	1.44	52	4	0.51	2.01	147	4	0.215	0.137	0.014	0.111	0.123	0.137	0.151	0.165