The Quality of Content Analyses of State Student Achievement Tests and Content Standards

Andrew C. Porter and Morgan S. Polikoff, *University of Pennsylvania*
Tim Zeidner, *The Church of Jesus Christ of the Latter-Day Saints*
John Smithson, *University of Wisconsin-Madison*

This article examines the reliability of content analyses of state student achievement tests and state content standards. We use data from two states in three grades in mathematics and English language arts and reading to explore differences by state, content area, grade level, and document type. Using a generalizability framework, we find that reliabilities for four coders are generally greater than .80. For the two problematic reliabilities, they are partly explained by an odd rater out. We conclude that the content analysis procedures, when used with at least five raters, provide reliable information to researchers, policymakers, and practitioners about the content of assessments and standards.

Keywords: content analysis, alignment, reliability, achievement tests, content standards

The purpose of this article is to report on investigations of the quality of content analyses of state student achievement tests and state content standards in mathematics and English language arts and reading (ELAR). In particular, this article asks two questions: what is the overall reliability of content analyses, and does reliability vary with grade level, academic subject, instrument analyzed (i.e., test vs. standard), or state? Content analyses are used for a variety of purposes. A primary purpose is for investigating the extent to which a state's student achievement tests are aligned to their state content standards. The No Child Left Behind Act requires that states demonstrate the alignment of their assessments to their content standards. Such demonstrations rely on content analyses of each set of documents. For the demonstrations to be convincing and replicable, the content analyses must be reliable in the sense that independent content analyses yield similar results. For this use, reliability is particularly critical.

Content analyses are used together with graphic displays to make clear to policymakers, practitioners, and researchers the "content messages" of content standards and assessments. Information from content analyses of assessments, standards, and instruction has been used as the backbone of information for teacher professional development. An alignment index calculated on data from the content analyses has been used as a teacher-level dependent variable in a number of program evaluations and research studies (Porter, 2002). Over 30 states and several big city school districts are using data from the content analysis procedures. The success of the various uses of content analyses requires that the results of the content analyses are reliable in the sense that they are replicable and not simply an artifact of a particular person's judgments.

The procedures for content analysis have been adopted by the Surveys of Enacted Curriculum (SEC) project of the Council of Chief State School Officers (CCSSO). The SEC is one of 13 CCSSO State Collaboratives on Assessment and Student Standards projects (SCASS). Since SCASS began in 1991, 31 states and six urban districts have had standards or assessments coded in mathematics, and 17 states have had standards or assessments coded in English language arts (procedures for English language arts are a more recent development). The origins of the content analysis procedures come from the Content Determinants Group of the Institute for Research on Teaching at Michigan State University during the late 1970s and early 1980s (e.g., Floden, Porter, Schmidt, Freeman, & Schwille, 1981; Schwille et al., 1983). The content analyses begin with the concept of a common language for describing content in a school academic subject. While the content languages have evolved over time, they have always included at least two dimensions. One dimension might be called "topics," for example, in mathematics, linear equations. The other dimension might be called cognitive demand, for example, memorization or problem solving. Specific content is then defined at the intersection of a topic and cognitive demand, for example, use a linear equation to solve a novel problem.

Currently, the SEC K–12 has content languages for mathematics,

Andrew C. Porter is Dean and George and Diane Weiss Professor of Education at the Graduate School of Education, University of Pennsylvania, Philadelphia, PA 19104; andyp@gse.upenn.edu. Morgan S. Polikoff is a Ph.D. Candidate, Education Policy, Graduate School of Education, University of Pennsylvania. At the time of writing, Tim Zeidner was post-doctoral researcher at Peabody College of Education and Human Development, Vanderbilt University. John Smithson is a researcher at the Wisconsin Center for Education Research, University of Wisconsin-Madison.

English language arts and reading, and science. Here, the focus is on mathematics and English language arts and reading. For each subject, the list of topics is divided into general areas. For example, in mathematics there are 16 general areas (e.g., operations, measurement, basic algebra). In each general area there are more specific topics, ranging in number from 19 to four, for a total number of specific topics in mathematics of 183. There are also five levels of cognitive demand. In mathematics they are: memorize; perform procedures; demonstrate understanding; conjecture, generalize, prove; and solve novel problems, make connections. In English language arts and reading there are 14 general areas of topics (e.g., phonemic awareness, fluency, comprehension). Within the general areas there are varying numbers of specific topics, from as many as 16 to as few as five, for 133 specific topics in all. As in mathematics, there are five levels of cognitive demand for English language arts and reading: memorize, perform procedures, generate, analyze, and evaluate (see Appendices A and B for complete descriptions of the content languages for mathematics and English language arts and reading).

The SEC is not the only measure of alignment currently in use. Norman Webb's alignment procedure has been used at all K–12 levels in the four major content areas. Webb's alignment measure relies on expert ratings in four areas of content agreement (no overall summary measure is reported): categorical congruence, depth of knowledge consistency, range of knowledge correspondence, and balance of representation (Webb, 2002). Achieve Incorporated's measure of alignment between standards and assessments is based on six dimensions: content centrality, performance centrality, source of challenge, level of challenge, balance, and range (Rothman, Slattery, Vranek, & Resnick, 2002). These are just a few of the examples of alignment measures that have been proposed and are in use. Each of these alignment procedures uses experts to perform a content analysis of the content standards, the student achievement tests, or both. In that sense, the utility of each is dependent upon the reliability of the content analyses, as are the SEC procedures investigated here. Nevertheless, the results for SEC alignment reliability are not generalizable in any straightforward way to the reliability of other procedures because the procedures themselves differ one from another. The reliability of each procedure needs to be estimated and reported. Toward that end, this study was done.

Background

Recognizing the importance of quality of content analyses for enhancing the applicability of alignment measures, several studies have examined the quality of raters' content analyses of standards and assessments based on various criteria. Webb, Herman, and Webb (2007) examined three alignment studies of state-level standards and assessments that employed the Webb alignment process (Herman, Webb, & Zuniga, 2005; Webb, 2005, 2006). The purpose was to examine how different approaches to accounting for rater agreement affected conclusions about alignment. The results revealed that, despite common training, rater agreement varied widely among the three studies. In one study, two-thirds of reviewers agreed on the content of just 15 of 43 test items, while in another study two-thirds of reviewers agreed on the content of all test items. The authors also calculated alignment indices using only each analyst's ratings for which they agreed with the majority of analysts. The argument for this exercise was that it only makes sense to calculate alignment based on the items for which analysts agreed on the content. The authors showed that alignment indices tended to be higher when all ratings were considered, not just the ratings for which analysts agreed with the majority. Finally, the authors used their findings to suggest that clarifying and specifying state standards, employing larger numbers of reviewers, averaging responses across reviewers, and improving training may reduce rater disagreement issues.

Additional research suggested that rater agreement may depend in large part on who the raters are. One study (Buckendahl, Plake, Impara, & Irwin, 2000) had both textbook publishers and Nebraska teachers rate the alignment of the textbook companies' omnibus tests to the state's standards. While the publishers often rated alignment highly—between 11 and 15 of 16 standards covered on the two tests over three grade levels—teachers rated alignment much lower—no more than eight standards covered on any given test. Another evaluation used high school teachers and college faculty as raters for the Webb alignment process (Herman et al., 2005). The results revealed that, while raters generally agreed on the topic covered by an item, high school teachers tended to rate items as more multidimensional and requiring deeper knowledge than did faculty. The findings in these articles suggest that training is important for improving rater agreement and that raters' backgrounds may be related to the ratings they give.

While the findings in all of these studies are important and reveal several details about the quality of raters' content analyses, the study described here is the first to extend Porter's (2002) work on rater agreement on the SEC. That study considered the content analyses of standards and assessments in four states in seventh-grade mathematics. Porter found that average reliabilities for four raters was relatively high (.82), but he did not report reliabilities at the individual state level. This study will extend Porter's earlier work by examining interrater reliability on the SEC across states, grade levels, and subjects. This may be particularly important because the SEC is the only content analysis tool that is not essentially limited to measuring alignment among tests and standards; it can also be used to measure alignment of instruction with either tests or content standards (Porter, Smithson, Blank, & Zeidner, 2007). Furthermore, this study employs generalizability theory to address the question of how many raters are needed to achieve reliable results, a result that has policy relevance for the states and agencies using the SEC as content analysis tools. Finally, the study investigates whether reliability depends on instrument analyzed (test vs. standard), academic subject, grade level, or state.

Content Analysis Procedures

For assessments of student achievement, each item is content analyzed by each content analyst. Typically, there are three to five content analysts for any document content analyzed. Before content analysis begins, the content analysts are given the content language. Each specific topic is represented by a three-digit number where the first digit indicates the general content area and the next two digits indicate the specific subtopic within that content area. The three-digit number is followed by a letter, A–E, to indicate the five

possible levels of cognitive demand. The procedure requires that every item be given at least one content code, but items can be given up to three content codes (i.e., specific subtopic by cognitive demand combination). Multiple codes are possible because some items are "fat" in the sense that they cover more than one specific subtopic by cognitive demand combination (cell in the two-dimensional content taxonomy language).

To begin, a sample set of assessment items is content analyzed individually by each coder, using the coding procedures described above. The sample items and their content codes are then discussed by each coding team in order to establish a common understanding and set of coding conventions for conducting the content analyses of the various documents. For each document to be content analyzed, each team begins by working individually, coding each item and flagging any items that cause confusion for the coding process. Before the coding is completed, the content analysis team of three to five members convenes to discuss flagged items. Following discussion, coders may decide to change their initial coding, but that is not required. At the end of the procedure, each analyst is tasked with making their best professional judgment on how to code each specific item. To the extent that this discussion makes ratings nonindependent, calculated reliabilities will be inflated. However, discussion is only done on flagged items, which comprise a small portion of total items, and raters are not asked to reach a common shared decision.

The data are ultimately put in the form of proportions with proportions of total content summing to one across the rows and columns of the content taxonomy. The proportions are calculated on the basis of score points from the assessment. If an item worth up to three points is placed in one cell in the content matrix, then all three points for that item go in that cell. If that three point item is placed in two separate cells in the content matrix, then 1.5 points is put in each cell. Multiple choice items are weighted one score point, but again, that score point is divided among the relevant cells as indicated by the content analyst. The data for each content analyst are first put in the form of proportions. Next, an average is taken, cell by cell, across content analysts. It is the agreement among content analysts in cell proportions that determines the reliability of the average across analysts.

The procedures for content standards are in every way identical, except that what is being content analyzed is not test items, but rather the most specific level of statement found in the content standards document (often called objectives). Further, because some objectives can be quite "fat," the convention is to allow up to six combinations of specific subtopics by cognitive demand (i.e., cells in the content taxonomy) to represent a single objective.

There are some conventions for content analyzing either tests or standards:

- If an item or standard cannot be associated with a specific topic in the taxonomy, then (a) if it fits a general content area it is coded with that number followed by two zeros, (b) if it is a specific topic not listed in the taxonomy it is coded by the general area number followed by a 90, (c) if there is no appropriate content code then it is coded 000, and (d) if it is judged to be not content in the subject area it is coded 999.
- If no appropriate cognitive demand applies, then it is given a Z.

The content analysts are recruited from the ranks of professors with the appropriate subject matter expertise, state content experts/consultants, and occasionally a graduate student in the relevant content area.

In recent years, it has been common for the SEC/SCASS to convene a content coding conference where each of several states bring the materials they want content analyzed to the conference. These conferences generally last for 2 to 3 days. Occasionally, when the content analyses are not completed by the end of the conference, content analysts take the materials with them and complete the content analyses at their home location. The repository for the data is the Wisconsin Center for Education Research at the University of Wisconsin-Madison. John Smithson is the director of the WCER SEC activity.

Analyses of the Data

Again, the purpose is to investigate the quality of content analysis data for the SEC alignment procedures. The data of interest are the topics by cognitive demand proportions in a content matrix for a specific document, either a student achievement test or a state content standards document. Again, the proportions are found by taking the average across separate content analysts. As described above, content analysts operate more or less independently, one from another, in completing the content analysis task. The states chosen for this study were taken from the available SEC data. States A and B are both Midwestern states—they were selected for this study because they were the only states for which tests and standards were content analyzed at three grades and in two subjects. Again, a purpose of this study was to investigate whether reliability is a function of subject, instrument (test vs. standards), grade level, and state.

We used generalizability theory and GENOVA software to do the analyses. First, we rearranged the two-dimensional content matrix data into a single vector for each content analyst, where the vector is defined as the complete listing of cells in the content matrix across all specific combinations of topics and cognitive demands. For mathematics there were 915 entries and for English language arts and reading there were 665 entries. It may be helpful to make an analogy to student achievement testing for which generalizability theory might be used to estimate reliability of the total score summing across items. In the analogy, our content matrix cells are the students and our content analysts are the items. In short, we are measuring the relative content emphasis of each cell as established by the average proportion across content analysts. Our measures of the cell are from each content analyst. We wish to determine the reliability of the average proportion across content analysts just as in student achievement testing we wish to determine the reliability of the student total score summed across items. Generalizability theory allows us to estimate what the reliability would be as the number of analysts varies. Obviously, the more analysts the better the reliability, though the relationship between reliability and number of analysts approaches an asymptote after approximately eight or nine raters.

Results

Taking from data since 2005, this study explored the reliability of content analysis results for State A and State B in mathematics and English language arts and reading. For each state and subject,

Table 1. Generalizability Coefficients

State A					State B				
Math									
		grade 3	grade 6	Algebra			grade 4	grade 6	High school
Test	# of Raters	(3)	(3)	(3)	**Test**	# of Raters	(5)	(4)	(4)
	2	0.83	0.78	0.84		2	0.63	0.89	0.69
	3	0.88	0.85	0.89		3	0.72	0.93	0.77
	4	0.91	0.88	0.92		4	0.77	0.94	0.82
	5	0.93	0.90	0.93		5	0.81	0.95	0.85
	6	0.94	0.92	0.94		6	0.83	0.96	0.87
	8	0.95	0.94	0.96		8	0.87	0.97	0.90
	10					10			
Standard	# of Raters	(3)	(3)	(3)	**Standard**	# of Raters	(3)	(3)	(3)
	2	0.76	0.63	0.73		2	0.76	0.66	0.70
	3	0.83	0.72	0.80		3	0.83	0.74	0.77
	4	0.87	0.77	0.85		4	0.86	0.83	0.85
	5	0.89	0.81	0.87		5	0.89	0.79	0.82
	6	0.91	0.84	0.89		6	0.90	0.85	0.87
	8	0.93	0.87	0.92		8	0.93	0.89	0.90
	10	0.94	0.90	0.93		10	0.94	0.91	0.92
ELAR									
		grade 3	grade 6	High School			grade 4	grade 6	High School
Test	# of Raters	(4)	(2)	(4)	**Test**	# of Raters	(5)	(4)	(4)
	2	0.71	0.27	0.57		2	0.69	0.74	0.54
	3	0.79	0.36	0.67		3	0.77	0.81	0.64
	4	0.83	0.43	0.73		4	0.82	0.85	0.70
	5	0.86	0.48	0.73		5	0.85	0.88	0.74
	6	0.88	0.53	0.80		6	0.87	0.90	0.78
	8	0.91	0.60	0.84		8	0.90	0.92	0.82
	10	0.93	0.65	0.87		10	0.92	0.94	0.85
Standard	# of Raters	(2)	(4)	(4)	**Standard**	# of Raters	(4)	(3)	(4)
	2	0.47	0.70	0.81		2	0.68	0.78	0.83
	3	0.57	0.78	0.87		3	0.76	0.85	0.88
	4	0.64	0.82	0.90		4	0.81	0.88	0.91
	5	0.67	0.85	0.92		5	0.84	0.90	0.93
	6	0.72	0.87	0.93		6	0.87	0.92	0.94
	8	0.78	0.90	0.95		8	0.90	0.94	0.95
	10	0.81	0.92	0.96		10	0.91	0.95	0.96

Note: The shaded values represent generalizability coefficients for four raters. These values are summarized in Table 2.

there were data for content analyses of both the test and the standards at grades 3, 6, and high school. The 2 × 2 × 2 × 3 design allows the analysis to explore whether or not the quality of data varies by subject, by test versus standards, by state, and by grade level. The basic data are shown in Table 1. In the table, each entry represents a generalizability coefficient, indicating the reliability of the average cell proportion across content analysts (raters). Data are shown for two through six raters and additionally for eight and 10 raters. At the top of each column of generalizability coefficients is a number in parenthesis indicating the actual number of content analysts (raters) on which the data are based. The other entries are projections. The generalizability coefficients for the actual number of raters are circled.

As can be seen, with a few exceptions, the reliability of the content analyses are quite good and certainly good when there are four or more raters (generally above .8). There are, however, a few troublesome numbers for English language arts and reading in State A (more on that later).

Table 2 displays the data for generalizability coefficients based on four raters. Calculating the marginal averages, the average reliability for math is .86 and for English language arts and reading .78. For tests, the average reliability is .80 and for standards .83, a trivially small difference even if statistically reliable. For grade levels the average generalizability coefficient is .81 for grade 3, .80 for grade 6, and .84 for high school, again small differences. Clearly, the quality of the data is equally good regardless of whether one is content analyzing tests or standards and regardless of grade level. There does appear to be a slightly lower quality of data for English language arts and reading than in mathematics despite raters having to make only 665 distinctions rather than the 915 required for math. However, this finding holds in State A but not State B.

Table 2. Average Generalizability Coefficients for Four Raters

		Grade	State A	State B
Math	T[a]	3	0.91	0.77
		6	0.88	0.94
		High School	0.92	0.82
	S[b]	3	0.87	0.86
		6	0.77	0.83
		High School	0.85	0.85
ELAR	T	3	0.83	0.82
		6	0.43	0.85
		High School	0.73	0.70
	S	3	0.64	0.81
		6	0.82	0.88
		High School	0.90	0.91

[a]Test.
[b]Content standard.

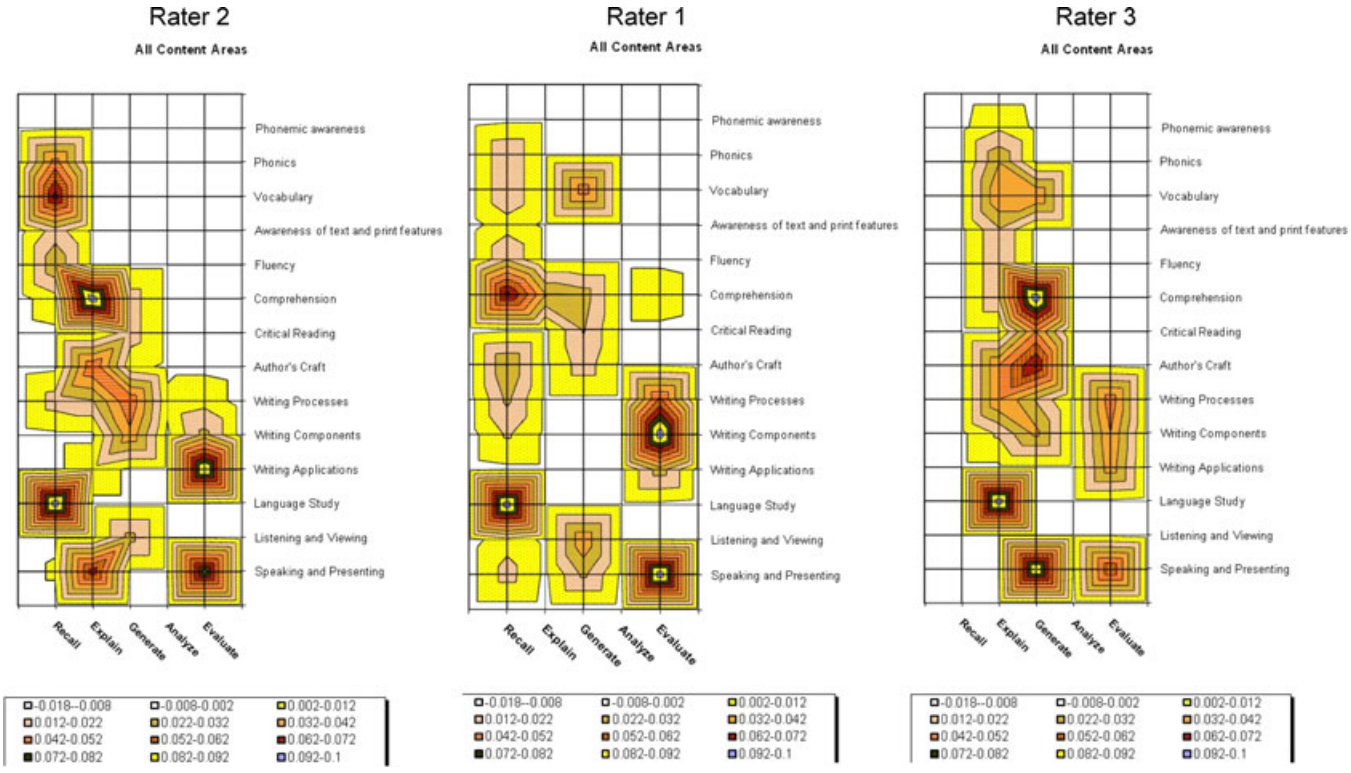

FIGURE 1. Rating of grade 3 ELAR standards.

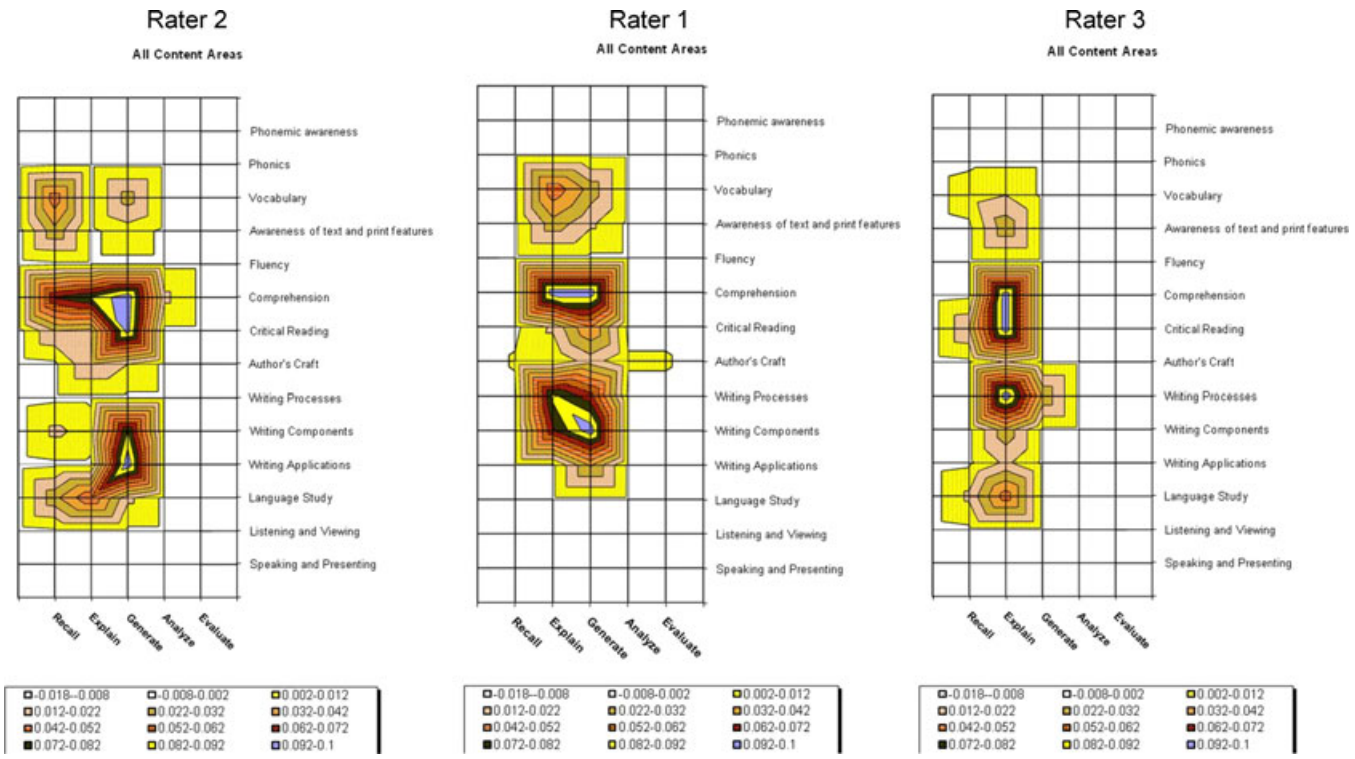

FIGURE 2. Rating of grade 6 ELAR test.

The average reliability by state for mathematics is .87 for State A and .85 for State B. For English language arts and reading, the average reliability is .73 for State A and .83 for State B.

For State A English language arts and reading, two reliabilities stand out as troublesome: the test at grade 6 and the standards at grade 3. In each case, there was some evidence of an "outlier" rater. For grade 3 standards, the "outlier" was reluctant to code anything as "recall." Figure 1 uses content maps to display the data for the three raters of grade 3 standards on 14 general categories of topics.[1] Across the rows are the different cognitive demands, and down the columns are the different content areas. At each intersection (cognitive demand-content area combination) the relative emphasis coded by the analyst is indicated by the shading. The darker the shading, the more emphasis was placed on that intersection by the analyst. As shown in Figure 1, while raters 1 and 2 coded several objectives as focused on recall (the leftmost column on the content map), rater 3 excluded recall altogether. In the case of the sixth-grade test, the coding team consisted of two experienced raters and one first-time rater. Figure 2 shows that the inexperienced rater 3 coded almost all test items under the cognitive demand "explain." Eliminating the odd rater out and recalculating the generalizability coefficients from these two cases raised the grade 6 test coefficient from .43 to .61 (still low) and the grade 3 test from .64 to .83. Eliminating the odd rater out also has the effect of diminishing the difference between mathematics and ELAR reliabilities by raising the average State A ELAR reliability to .79.

Summary and Conclusions

The purpose of this article was to investigate the quality of data from content analyses of student achievement assessments and state content standards using Surveys of Enacted Curriculum content analysis procedures. The data were from content analyses performed since 2005 for states A and B. The data represent content analyses of both tests and standards for both mathematics and English language arts and reading. For each, the data cover grades 3, 6, and high school. The study used generalizability coefficients to estimate the reliability of the content analyses when averaged across content analysts. The number of content analysts varied from as few as two in two cases to as many as five in two cases.

Using generalizability theory allowed the projection of reliabilities for varying numbers of raters. Settling on four raters for the bulk of the comparisons, the results showed the reliability of the sum across raters was typically .8 or better with some generalizability coefficients reaching .9 and above. It should be emphasized that these are reliabilities at the cell level; math content analysts had to decide among 915 cells and English language arts and reading had to decide among 665 cells. The reliability at the marginals of topics or cognitive demand would be much higher. Those reliabilities were not calculated because the SEC does not encourage aggregating to the marginals and because the cell-level reliabilities were more than satisfactory. Prior research has shown that if the goal is to explain differential gain in student achievement, the content of instruction must be aligned to the tests at the cell level, with correlations of alignment to gains of approximately .45. These correlations turn to essentially zero for just topics or just cognitive demand (Porter, 2002).

There was no evidence that the reliability of content analysts was better for tests than for standards. The average four-rater generalizability coefficient for standards was .83; for tests

it was .80. One might have hypothesized that the reliability would be better for tests since the unit being content analyzed is an item that is fairly self-contained, whereas in standards what is being content analyzed is an "objective," which differs in definition and size over documents. Neither did it appear to make any difference at what grade level the analyses were conducted; there was no indication of content data being more reliable at one grade level than another. There was some reason to believe that mathematics was content analyzed more reliably than English language arts and reading, but this result was entirely explained by an interaction between subject and state. The ratings for State A English language arts and reading were lower, .73, than for mathematics or for either subject in State B, above .8. A generalizability coefficient of .73 may be borderline acceptable for quality of data. Eliminating the odd rater out in these analyses increased the reliability of the data to .8 or above for four raters with the exception of the sixth-grade test in English language arts and reading, which remained a problem with a reliability of .61 estimated for four raters.

These results are encouraging. The procedures followed by SEC from recruitment of content analysts to training of content analysts to performing the task and archiving the data appear to result in reliable data, with reliability coefficients in the .8 to .9 range for four or more raters. Since only two states were included in this study, results may not generalize to all states. Almost certainly, states differ in the nature and format of their content standards. Some states may have quite specific content standards and others more general content standards, and this could result in content analyses that are more or less reliable. In this study, there was some evidence of between-state differences in the reliability for English language arts and reading. Further, no claims can be made about reliabilities using other measures of alignment.

Based on these analyses, it is recommended that all content analyses based on the SEC use at least five raters. Interrater correlations should be calculated in an attempt to identify a possible odd rater and this odd rater should be eliminated before calculating the average across raters. More broadly, any measure of alignment that is used should be studied to determine its reliability.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C050041–05 to the University of Pennsylvania. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education.

Note

[1]Surface area charts can be created using a variety of charting software, including Excel. Since the variables being graphed are nominal scale, the charts are only meaningful at the intersections of rows and columns.

References

Buckendahl, C. W., Plake, B. S., Impara, J. C., & Irwin, P. M. (2000). *Alignment of standardized achievement tests to state content standards: A comparison of publishers' and teachers' perspectives*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Floden, R. E., Porter, A. C., Schmidt, W. H., Freeman, D. J., & Schwille, J. R. (1981). Responses to curriculum pressures: A policy capturing study of teacher decisions about content. *Journal of Educational Psychology*, *73*, 129–141.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2005). *Measurement issues in the alignment of standards and assessments: A case study*. Paper presented at the Annual Conference of the American Educational Research Association, Montreal.

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

Porter, A. C., Smithson, J., Blank, R., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education*, *20*(1), 27–51.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing*. Los Angeles: University of California, National Center for Evaluation of Research on Evaluation, Standards, and Student Testing.

Schwille, J. R., Porter, A. C., Belli, G., Floden, R. E., Freeman, D. J., & Knappen, L. B. (1983). Teachers as policy brokers in the content of elementary school mathematics. In L. Shulman & G. Sykes (Eds.), *Handbook on teaching and policy* (pp. 370–391). New York: Longman.

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.

Webb, N. L. (2005). *Alignment analysis of mathematics standards and assessments, Michigan, high school*. Unpublished report. Madison, WI: Author.

Webb, N. L. (2006). *Alignment analysis of mathematics standards and assessments, Tennessee grades 3–9*. Unpublished report. Madison, WI: Author.

Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice*, *26*(2), 17–29.

Appendix A. K-12 Mathematics Taxonomy

100 Nbr. Sense/Properties/Relationships
200 Operations
300 Measurement
400 Consumer Applications
500 Basic Algebra
600 Advanced Algebra
700 Geometric Concepts
800 Advanced Geometry
900 Data Displays
1000 Statistics
1100 Probability
1200 Analysis
1300 Trigonometry
1400 Special Topics
1500 Functions
1600 Instructional Technology

K-12 Mathematics Taxonomy

100 Nbr. sense/Properties/Relationships
101 Place value
102 Whole numbers, integers
103 Operations
104 Fractions
105 Decimals
106 Percents
107 Ratio, proportion
108 Patterns
109 Real and/or rational numbers
110 Exponents, scientific notation
111 Factors, multiples, divisibility
112 Odds/evens/primes/composites/square numbers
113 Estimation
114 Number Comparisons (order, magnitude, relative size, inverse, opposites, equivalent forms, scale, number line)
115 Order of operations
116 Computational algorithms
117 Relationships between operations
118 Number theory (e.g. base-ten, non base-ten systems)
119 Mathematical properties (e.g., distributive property)
190 Other

200 Operations
201 Add, subtract whole numbers, integers
202 Multiplication whole numbers, integers
203 Division whole numbers, integers
204 Combinations of add, subtract, multiply, divide by whole numbers or integers
205 Equivalent/non-equivalent fractions
206 Add, subtract fractions
207 Multiply fractions
208 Divide fractions
209 Combinations of add, subtract, multiply, divide fractions
210 Ratio, proportion
211 Representations of fractions
212 Equivalence of decimals, fractions, %
213 Add, subtract decimals
214 Multiply decimals
215 Divide decimals
216 Combinations of add, subtract, multiply, divide decimals
217 Computing with percents
218 Computing with exponents, radicals
290 Other

300 Measurement
301 Use of measuring instruments
302 Theory (arbitrary, standard units, unit size)
303 Conversions
304 Metric (SI) system
305 Length, perimeter
306 Area, volume
307 Surface area
308 Direction, location, navigation
309 Angles
310 Circles (e.g., pi, radius, area)
311 Mass (weight)
312 Time, temperature
313 Money
314 Derived measures (e.g. rate/speed)
315 Calendar
316 Accuracy, precision
390 Other

400 Consumer Applications
401 Simple interest
402 Compound interest
403 Rates (e.g., discount, commission)
404 Spreadsheets
490 Other

500 Basic Algebra
501 Absolute value
502 Use of variables
503 Evaluation of formulas, expressions, equations
504 One-step equations
505 Coordinate plane
506 Patterns
507 Multi-step equations
508 Inequalities
509 Linear, non-linear relations
510 Rate of change/slope/line
511 Operations on polynomials
512 Factoring
513 Square roots & radicals
514 Operations on radicals
515 Rational expressions
516 Multiple representations
590 Other

600 Advanced Algebra
601 Quadratic equations
602 Systems of equations
603 Systems of inequalities
604 Compound inequalities
605 Matrices, determinants
606 Conic sections
607 Rational, negative exponents radicals
608 Rules for exponents
609 Complex numbers
610 Binomial theorem
611 Factor/remainder theorem
612 Field properties of real number system
613 Multiple representations
690 Other

700 Geometric Concepts
701 Basic terminology
702 Points, lines, rays, segments and vectors
703 Patterns
704 Congruence
705 Similarity
706 Parallels
707 Triangles
708 Quadrilaterals
709 Circles
710 Angles

K-12 Mathematics Taxonomy (*continued*)

711 Polygons
712 Polyhedra
713 Models
714 3-D relationships
715 Symmetry
716 Transformations (e.g., flips, turns)
717 Pythagorean theorem
790 Other

800 Advanced Geometry
801 Logic, reasoning, proof
802 Loci
803 Spheres, cones, cylinders
804 Coordinate geometry
805 Vectors
806 Analytic geometry
807 Non-Euclidean geometry
808 Topology
890 Other

900 Data Displays
901 Summarize data in a table or graph
902 Bar graph, histogram
903 Pie charts, circle graphs
904 Pictographs
905 Line graphs
906 Stem and leaf plots
907 Scatter plots
908 Box plots
909 Line plots
910 Classification, Venn diagrams
911 Tree diagrams
990 Other

1000 Statistics
1001 Mean, median, mode
1002 Variability, standard deviation, range
1003 Line of best fit
1004 Quartiles, percentiles
1005 Bivariate distribution
1006 Confidence intervals
1007 Correlation
1008 Hypothesis testing
1009 Chi square
1010 Data transformation
1011 Central limit theorem
1090 Other

1100 Probability
1101 Simple probability
1102 Compound probability
1103 Conditional probability
1104 Empirical probability
1105 Sampling, sample spaces
1106 Independent/dependent events
1107 Expected value
1108 Binomial distribution
1109 Normal curve
1190 Other

1200 Analysis
1201 Sequences and series
1202 Limits
1203 Continuity
1204 Rates of change
1205 Maxima, minima, range
1206 Differentiation
1207 Integration
1290 Other

1300 Trigonometry
1301 Basic ratios
1302 Radian measure
1303 Right triangle trigonometry
1304 Law of sines, cosines
1305 Identities
1306 Trigonometric equations
1307 Polar coordinates
1308 Periodicity
1309 Amplitude
1390 Other

1400 Special Topics
1401 Sets
1402 Logic
1403 Mathematical induction
1404 Linear programming
1405 Networks
1406 Iteration, recursion
1407 Permutations combinations
1408 Simulations
1409 Fractals
1490 Other

1500 Functions
1501 Notation
1502 Relations
1503 Linear
1504 Quadratic
1505 Polynomial
1506 Rational
1507 Logarithmic
1508 Exponential
1509 Trigonometric circular
1510 Inverse
1511 Composition
1590 Other

1600 Instructional Technology
1601 Use of calculators
1602 Use of graphing calculators
1603 Use of computers & internet
1604 Computer programming
1605 Use of spreadsheets
1690 Other

Cognitive Demand Categories for Mathematics

B Memorize Facts, Definitions, Formulas	C Perform Procedures	D Demonstrate Understanding of Mathematical Ideas	E Conjecture, Generalize, Prove	F Solve Non-Routine Problems/Make Connections
Recite basic mathematical facts	Use numbers to count. order, denote	Communicate mathematical ideas	Determine the truth of a mathematical pattern or proposition	Apply and adapt a variety of appropriate strategies to solve non-routine problems
Recall mathematics terms and definitions	Do computational procedures or algorithms	Use representations to model mathematical ideas	Write formal or informal proofs	Apply mathematics in contexts outside of mathematics
Recall formulas and computational procedures	Follow procedures/ instructions	Explain findings and results from data analysis strategies	Recognize, generate or create patterns	Analyze data, recognize patterns
————	Solve equations/ formulas/routine word problems	Develop/explain relationships between concepts	Find a mathematical rule to qenerate a pattern or number sequence	Synthesize content and ideas from several sources
————	Organize or display data	Show or explain relationships between models, diagrams, and/or other representations	Make and investigate mathematical conjectures	————
————	Read or produce graphs and tables	————	Identify faulty arguments or misrepresentations of data	————
————	Execute geometric constructions	————	Reason inductively or deductively	————

Appendix B. K-12 English Language Arts and Reading Content Areas

100	Phonemic awareness	800	Author's craft
200	Phonics	900	Writing processes
300	Vocabulary	1000	Writing components
400	Awareness of text and print features	1100	Writing applications
500	Fluency	1200	Language study
600	Comprehension	1300	Listening and viewing
700	Critical Reading	1400	Speaking and presenting

K-12 English Language Arts/Reading Taxonomy

100 Phonemic awareness
101 Phoneme isolation (e.g., the distinct sounds /c/ /a/ /t/)
102 Phoneme blending (e.g., c/a/ t = cat)
103 Phoneme segmentation
104 Onset-rime
105 Sound patterns
106 Rhyme recognition
107 Phoneme deletion/substitution/addition
108 Identify Syllables
190 Other

200 Phonics
201 Alphabetic principle (includes alphabet recognition, order)
202 Consonants
203 Consonant blends
204 Consonant digraphs (e.g., ch, sh, th)
205 Diphthongs (e.g., oi, ou, ow, oy (as in "boy"))
206 R-controlled vowels (e.g., farm, torn, turn)
207 Patterns within words
208 Vowel letters (a, e, I, o, u, y)
209 Vowel phonemes (15 sounds)
210 Sound/symbol relationships
211 Blending sounds
290 Other

300 Vocabulary
301 Compound words and contractions
302 Inflectional forms (e.g., -s, -ed, -ing)
303 Suffixes, prefixes, and root words
304 Word definitions (including new vocabulary)
305 Word origins
306 Synonyms, antonyms, homonyms
307 Word or phrase meaning from context
308 Denotation and connotation
309 Analogies
310 Sight words
311 Use of references
390 Other

400 Text and print features
401 Book handling
402 Directionality; sequence of text
403 Parts of a book (e.g., cover, title, front, back)
404 Letter, word, sentence distinctions
405 Structural elements (e.g., index, glossary, table of contents, subtitles, headings)
406 Graphical elements (e.g., graphs, charts, images, illustrations)
407 Technical elements (e.g., bullets, instructions, forms, sidebars)
408 Electronic elements (e.g., hypertext links, animations)
409 Environmental print, i.e. prints or symbols found in students' everyday environment
490 Other

500 Fluency
501 Prosody (e.g., phrasing, intonation, inflection)
502 Automaticity of words and phrases (e.g., sight and decodable words)
503 Speed/pace
504 Accuracy
505 Independent reading (e.g., repeated/silent reading for fluency)
590 Other

600 Comprehension
601 Word meaning from context
602 Phrase
603 Sentence
604 Paragraph
605 Main idea(s), key concepts, sequence of events
606 Descriptive elements (e.g., detail, color, condition)
607 Narrative elements (e.g., events, characters, setting, plot)
608 Persuasive elements (e.g. propaganda, advertisement, emotional appeal)
609 Expository or informational elements (e.g., explanation, lists, organizational patterns: description, cause-effect, compare-contrast)
610 Technical elements (e.g., bullets, instruction, form, sidebars)
611 Electronic elements (e.g., hypertext links, animations)
612 Strategies (e.g., activating prior knowledge, questioning; making connections, predictions; inference, imagery, summarization, re-telling)
613 Self-correction strategies (e.g., monitoring, cueing systems, and fix-up)
614 Metacognitive process (e.g., reflecting about one's thinking)
615 Interpret maps, graphs, charts
616 Test taking strategies
690 Other

700 Critical reading
701 Fact and opinion
702 Appealing to authority, reason, emotion
703 Validity and significance of assertion or argument
704 Relationships among purpose, organization, format, and meaning in text
705 Author's assumptions, bias
706 Comparison of topic, theme, treatment, scope, or organization across texts
707 Inductive/deductive approaches (e.g., making inference, drawing conclusions)
708 Logical reasoning in text (e.g., implications, authors' rationale, development of argument)
709 Textual evidence, use of references to support
710 Drawing meaning from allegory, myth
711 Distinguishing real from fantastical events in literature
790 Other

800 Author's craft
801 Theme/thesis
802 Purpose (e.g., inform, perform, critique, appreciate)

K-12 English Language Arts/Reading Taxonomy (*continued*)

803 Characteristics of genres or forms
804 Point of view (e.g., first or third person, multiple perspectives)
805 Literary devices (e.g., analogy, simile, metaphor, hyperbole, flashbacks, structure, archetypes)
806 Literary analysis (e.g., symbolism, voice, style, tone, mood)
807 Influence of time and place on authors and texts (e.g., historical era, culture)
808 Aesthetic aspects of text (e.g., dramatic, poetic elements)
890 Other

900 Writing processes
901 Printing, cursive writing, penmanship
902 Pre-writing (e.g., essential questions, topic selection, brainstorming)
903 Drafting and revising
904 Editing for conventions (e.g., usage, spelling, structure)
905 Manuscript conventions (e.g., indenting, margins, citations, references)
906 Final draft, publishing
907 Use of technology (e.g., word processing, multimedia)
990 Other

1000 Elements of presentation (verbal and written)
1001 Purpose, audience, context
1002 Main ideas
1003 Organization
1004 Word choice
1005 Support and elaboration
1006 Style, voice, technique, use of figurative language
1007 Writing Conventions (e.g., capitalization, punctuation, indentation, citation)
1008 Transitional Devices
1090 Other

1100 Writing applications
1101 Narrative (e.g., stories, fiction, plays)
1102 Poetry
1103 Expository (e.g., report, theme, essay)
1104 Critical/evaluative (e.g., review)
1105 Expressive (e.g., journals, reflections)
1106 Persuasive (e.g., editorial, advertisement, argumentative
1107 Procedural (e.g., instructions, brochures, lab report)
1108 Technical (e.g., manual, specification, research report)
1109 Real world applications of writing (e.g., résumés, letter to editor, note taking)
1190 Other

1200 Language study
1201 Syllabification
1202 Spelling
1203 Capitalization and punctuation
1204 Signs and symbols (e.g., semiotics)
1205 Syntax and sentence structure
1206 Grammatical analysis
1207 Standard and non-standard language usage
1208 Linguistic knowledge (including dialects and diverse forms)
1209 History of language
1210 Relationships of language forms, contexts, and purposes (e.g., rhetoric, semantics)
1211 Effects of race, gender, ethnicity on language & language use
1290 Other

1300 Listening and Viewing
1301 Listening
1302 Viewing
1303 Nonverbal communication
1304 Consideration of others' ideas
1305 Similarities/differences of print, graphic, and nonprint communications
1306 Literal/Connotative meaning
1307 Diction, tone, syntax, convention, rhetorical structure in speech
1308 Media supported communication
1390 Other

1400 Speaking and presenting
1401 Public speaking, oral presentation
1402 Diction, tone, syntax, conventions, rhetorical structure in speech
1403 Demonstrating confidence
1404 Effective nonverbal skills (e.g., gesture, eye contact)
1405 Knowledge of situational and cultural norms for expression
1406 Conversation and discussion (e.g., Socratic seminars, literature circles, peer discussion)
1407 Debate and structure of argument
1408 Dramatics, creative interpretation
1409 Media-supported communication
1410 Selecting presentation format
1411 Interviewing
1490 Other

Cognitive Demand Categories for Language Arts/Reading

B Memorize/ Recall	C Perform Procedures/Explain	D Generate/Create/ Demonstrate	E Analyze/ Investigate	F Evaluate
Reproduce sounds or words	Follow instructions	Create/develop connections among text, self, world	Categorize/ schematize information	Determine relevance, coherence, internal consistency, logic
Provide facts, terms, definitions, conventions	Give examples	Recognize relationships	Distinguish fact and opinion	Assess adequacy, appropriateness, credibility
Locate literal answers in text	Check consistency	Dramatize	Compare and contrast	Test conclusions, hypotheses
Identify relevant information	Summarize	Order, group, outline, organize ideas	Identify with another's point of view	Synthesize content and ideas from several sources
Describe	Identify purpose, main ideas, organizational patterns	Express new ideas (or express ideas newly)	Make inferences, draw conclusions	Generalize
_____	Gather information	Develop reasonable alternatives	Predict probable consequences	Critique
_____	_____	Integrate with other topics and subjects	_____	_____