A Report to the North Carolina Department of Public Instruction On the Alignment Characteristics of the *NCEXTEND1* Alternate Assessment Instruments

Grades 3-8 English Language Arts, Math and Science and
High School Grade 10 English II, Math I and Biology

Published April, 2016

Prepared by John L. Smithson, Ph.D., Researcher,
University of Wisconsin-Madison

TABLE OF CONTENTS

North Carolina 2015 *NCEXTEND1* Alignment Study Report

Introduction

The report below examines the alignment characteristics of mathematics, English language arts/reading (ELAR) and science assessments being administered as part of North Carolina's ***NCEXTEND1*** alternate assessment program. The ***NCEXTEND1*** alternate assessment is a performance-based alternate assessment designed to assess students with significant cognitive disabilities. The alignment targets for these assessments are based on the North Carolina Extended Content Standards.

The current report follows an earlier alignment study commissioned in January 2015 for the general population assessment program in mathematics, ELAR and science. As with the previous alignment report, the North Carolina Department of Public Instruction (NCDPI) commissioned the Wisconsin Center for Education Research to conduct an in-depth study into the alignment of the ***NCEXTEND1*** alternate assessments for mathematics, reading and science to the state's extended content standards for significantly cognitively disabled students.

Specifically, this report focuses on describing the alignment characteristics of the ***NCEXTEND1*** alternate assessment program in North Carolina based upon analyses of 17 assessment forms covering mathematics and reading ***NCEXTEND1*** alternate assessments for grades 3, 4, 5, 6, 7, 8, high school Math I and high school English II, as well as the ***NCEXTEND1*** alternate assessments for science at grades 5, 8, and high school Biology.

Content analysis of the ***NCEXTEND1*** alternate assessments were conducted using remote coding procedures. Remote coding procedures are used when copies of the assessments can be distributed under secure procedures to veteran analysts located at different locations. Analysts proceed with the independent stage of the analysis using the same general process as a workshop-based approach, beginning with independent analyses followed by group discussion and the opportunity for each team member to re-consider their codes. For more information on the data collection procedures used for present study see the supplement to Appendix A provided at the end of this report.

Structure of the Report

This report should be considered a companion document to the September 2015 North Carolina Alignment Study Report, which will be referenced below as the 'main report'. In order to avoid over-duplication of text, short summaries of some of the details associated with conducting and interpreting alignment analyses are provided in this report, with appropriate references to the main report for the reader interested in a more detailed discussion of The Surveys of Enacted Curriculum (SEC) alignment methodology.

The initial section of the report addresses the alignment targets selected for this analysis, and the rationale for reporting other alignment targets that take into consideration specific content not indicated for assessment in the test blueprint. Further discussion on the selection of alignment targets can be found in Section II: Conducting & Interpreting Alignment Analysis of the main report.

Part II of this report provides a brief overview of the SEC alignment process and the summary indicator measures used to describe an assessment's alignment characteristics. The summary results of the alignment analysis are presented in this section, organized by each dimension and reported across subject and alignment targets. For the reader interested in better understanding how the summary measures

presented in Section II are derived, or the justification for the 0.50 threshold measure amidst a broader discussion of determining 'good' alignment may also refer to Section II: Conducting & Interpreting Alignment Analysis of the main report.

Part III provides a discussion of the findings, and a more detailed review of assessments reporting relatively low alignment results, in order to better understand the alignment challenges represented by those assessment instruments. Additional information is provided in the various appendices to the main report.

Appendix A from the main report is intended to prepare the reader for exploring the underlying descriptive data behind the alignment results. Here the reader will learn more about content descriptions, how they are collected, processed and reported. These descriptive results further provide the reader with a detailed 'picture' of the content embedded in each of the documents analyzed for the study, and used to conduct the type of diagnostic analyses described in Appendix B. Appendix C provides a detailed description of the content alignment process and the materials used in the process. Appendices D-F have been updated to include the ***NCEXTEND1*** assessment descriptions and serve to provide subject-specific content viewers for generating the charts and maps introduced in Appendix A, and used to conduct fine grain analyses as described in Appendix B.

SEC Alignment Analysis Methodology

To understand how alignment is calculated using the SEC approach it is important to understand that alignment measures in SEC are *derived* from *content descriptions*. That is alignment analyses report on the relationship between two separate descriptions of academic content. In SEC these descriptions are based upon a three dimensional construct consisting of topics, cognitive demand and relative emphasis. Each dimension of the two selected descriptions can then be compared, using procedures described in the main report, to derive a set of alignment indicator measures that summarize the quantitative relationship between any two content descriptions. These alignment indicator measures are calculated on each of the dimensions used for describing academic content.

In addition to examination of each dimension independently, the method allows for examination of alignment characteristics at the intersection of all three dimensions employed, producing a summative 'overall' alignment indicator. This alignment indicator demonstrates a predictive capacity in explaining the variation in student achievement based upon a students' opportunity to learn the assessed content.

These *content descriptions* are systematically compared to determine the alignment characteristics existing between the two descriptions, using a simple iterative algorithm that generates an alignment measure or index based on the relevant dimension(s) of the content being considered. Content descriptions are described in more detail in Appendix A: Describing Academic Content of the main report.

As mentioned, there are three dimensions to the content descriptions collected, and hence three dimensions upon which to look at the degree of alignment the analyses indicate. These indicator measures can be distilled further to a single overall alignment index (OAI) that summarizes the alignment characteristics of any two content descriptions at the intersection of the three dimensions of content embedded in the SEC approach.

Alignment Targets

Typically the alignment target for a state assessment is a set of academic content standards selected by the state. In some cases the test framework or blueprint may be substituted as an alignment target, but this is not common as such documents often lack detailed content descriptions to support SEC style alignment

analysis. States may decide to not assess some content areas based on logistical, economic or other factors. In such cases an assessment can look dramatically out of alignment due to the un-assessed content area(s) purposely omitted from the test. To better describe the alignment characteristics of such assessments there are some instances where a more targeted selection of subject matter content is warranted.

Consider the case of reading assessments in the state. Reading encompasses only one portion of the complete set of content standards for ELAR. Yet the state has for example, explicitly chosen to not assess writing and other language skills as part of their reading assessment program. Even at the secondary level, where the state does include open-ended response items in its end of course assessments, those items are not scored for writing content. As a result, holding the reading assessment accountable to the writing content emphasized in ELAR standards would inevitably result in low alignment results relative to the larger scope of ELAR content, providing a misleading sense of the alignment of the reading assessments to content associated with language arts reading skills.

In order to make the alignment measures for the reading assessment more appropriate given the intended focus of those assessments, the results reported for reading below are based on alignment across the following ELAR content areas represented in the SEC taxonomy: Vocabulary; Text and Print Features; Fluency; Comprehension; Critical Reasoning; Author's Craft; Language Study; and Listening & Viewing. Excluded from the alignment analysis were content associated with the areas of Phonemic Awareness, Phonics, Writing Processes, Elements of Presentation, Writing Applications and Speaking & Presenting.

In mathematics the curriculum tends to be more stratified. That is basic mathematical operations and topics such as fractions, decimals and measurement units are typically handled at lower grades, while topics like algebra, geometry, and trigonometry tend to be focused at higher grade levels. As a result, even though there may clearly be mathematics content not taught at one or another grade level, such content is typically excluded from the mathematics standards for that grade level. Thus a modification of the alignment target is not generally needed for mathematics.

Nonetheless specific circumstances can and do arise in mathematics, that make a modification in the definition or description of the alignment target a reasonable consideration when it helps to highlight the impact of specific mathematics content on alignment. An example of this is provided below for grade 7 mathematics.

The Dimensions of Alignment

SEC content descriptions are collected on three dimensions: (1) topic coverage (2) performance expectation and (3) relative emphasis. The alignment indices used to summarize alignment characteristics parallel these dimensions with summary indices for: (1) balance of representation (BR); (2) Topic Coverage (TC); and (3) performance expectations (PE). When considered in combination with one another a fourth summary measure of overall alignment (OAI) can be calculated that serves as the key summary alignment measure resulting from the analyses. The algorithm for calculating these alignment indices is discussed further in *Section I: What is Alignment Analysis?* in the main report, as is a discussion of what constitutes 'good' alignment using the SEC approach.

Each alignment indicator is expressed on a scale with a range of 0 to 1.0, with 1.0 representing identical content descriptions (perfect alignment) and 0 indicating no content in common between the two descriptions, or perfect misalignment. For reasons discussed further in the main report, a threshold measure of 0.5 is set for each of the four summary indicator measures. Above this threshold alignment is

considered to be at an acceptable level. Below, and the alignment indicator suggests a more detailed examination of the descriptive data underlying the result is warranted. This does not *necessarily* mean that the assessment is not appropriately aligned, only that an explanation for the relatively low result is prudent, it means more information is needed. That additional information is available in the form of fine grain alignment analyses which serve a diagnostic, as opposed to summative purpose. An example of the fine grain analysis process is provided in Appendix B. The data and reporting tools necessary to conduct fine grain analyses for any of the documents analyzed (including the ***NCEXTEND1*** assessments, are supplied in Appendices D-F).

Balance of Representation

Of the three content dimensions on which alignment measures are based, two are directly measured and one is derived. Two of the content dimensions are based upon observer/analyst reports of the occurrence of one or another content description. The derived measure concerns 'how much' and is based on the number of reported occurrences for a specific description of content relative to the total number of reports making up the full content description. This yields a proportional measure, summing to 1.00. The SEC refers to this 'how much' dimension as 'balance of representation' (BR).

As a summary indicator, (BR) is calculated as the product of two values: the portion of the assessment that targets standards-based content, multiplied times the portion of standards-based content represented in the assessment. For example, if 90% of an assessment (i.e. 10% of the assessment covers content not *explicitly* referenced in the standards) covered 40% of the standards for a particular grade level (i.e. 60% of the content reflected in the standards was not reflected in the assessment), the BR measure would be 0.36. As with all the summary indicator measures, reported here, the 'threshold' for an acceptable degree of alignment is placed at 0.50 or higher. This example would thus reflect an alignment measure that would bear further review, given this criterion. For a fuller discussion of the rationale for the 0.5 measure, the reader is referred to Section II: Conducting & Interpreting Alignment Analysis of the September 2015 North Carolina Alignment Study Report.

The influence of BR runs through all of the alignment indices, since the relative emphasis of content is the value used in making comparisons between content descriptions. In a very real sense the dimensions of topic and performance expectation provide the structure for looking at alignment, while the balance of representation provides the proportional *values* placed in that structure.

For assessments, relative emphasis is expressed in terms of the proportion of score points attributed to one or another topic and/or performance expectation. When talking about standards, relative emphasis refers to the number of times a particular topic and/or performance expectation is noted across all the strands of a standard presented for a given grade and subject.

Table 1
Balanceof Representation Index by Grade and Subject

Grade	3	4	5	6	7	8	HS
NCEXTEND1 ELAR	0.28	0.36	0.30	0.37	0.31	0.44	0.44
NCEXTEND1 MATH	0.29	0.50	0.28	0.48	0.23	0.20	0.34
NCEXTEND1 SCIENCE			0.43			0.53	0.27

Table 1 summarizes the balance of representation results for the assessed standards. With only a few exceptions (grade 4 math and grade 8 science) the BR summary measures tend to fall well below the 0.5 threshold.

The ***NCEXTEND1*** alternate assessments face a particularly difficult challenge in achieving a strong BR measure, as the assessment for each test is limited to fifteen items. Given the shortness of the assessment instrument and the psychometric need for multiple items to support assertions of proficiency for any one topic, it is not surprising that these numbers are quite low compared to the general assessment instruments which incorporate about five times the number of test items per instrument. The balance of representation issue is addressed further under the discussion of findings below.

This one measure provides only one piece of the alignment picture and it tells only part of the story. The other indicators provide other perspectives for viewing alignment that help to provide a more detailed picture of the alignment relationship between the ***NCEXTEND1*** alternate assessments and their standards.

Topic Coverage

Topic coverage (TC) examines how well the assessment matches the relevant standards in terms of the topics covered by each. The algorithm used to calculate topical concurrence provides a summary of the extent to which topics in the assessment match the topics embedded in the relevant standards. Table 2 provides the summary results for TC across the mathematics and reading assessments analyzed for this study.

Table 2
Topic Coverage Index by Grade and Subject
(0.50 or greater = well-aligned)

Grade	3	4	5	6	7	8	HS
NCEXTEND1 ELAR	0.53	0.75	0.71	0.79	0.72	0.72	0.67
NCEXTEND1 MATH	0.60	0.66	0.72	0.75	0.53	0.73	0.84
NCEXTEND1 SCIENCE			0.63			0.71	0.73

For the most part, the results presented in Table 2 suggest good to strong alignment for the English language arts/reading alignment and mathematics targets.

Performance Expectations

The SEC taxonomies enable descriptions of academic content based on two dimensions ubiquitous to the field of learning: knowledge and skills. When referencing standards this is frequently summarized with the statement "what students should know and be able to do." The '*what students should know*' part refers to topics, while the '*be able to do*' reference expectations for student performance, or performance expectations. The SEC taxonomies enable the collection of content descriptions on both of these dimensions, and together form the alignment 'target' for both assessments and instruction.

Just as one can examine alignment with respect to topic coverage alone, it is possible to examine and compare the performance expectations embedded in the content descriptions of assessments and standards in a similar manner. This alignment indicator is referred to as performance expectations (PE), and is based on the five categories of expectations for student performance employed by the SEC. While the labels vary slightly from subject to subject, the general pattern of expectations follows this general division: 1)

Memorization/Recall, 2) Procedural Knowledge, 3) Conceptual Understanding, 4) Analysis, Conjecture and Proof, and 5) Synthesis, Integration and Novel Thinking.

Table 3 reports the performance expectation alignment measures across the assessed grades for mathematics and reading. As with topic coverage this dimension is expressed as an index with a range of 0.0 to 1.0, with 0.5 indicating acceptable alignment.

Table 3
Performance Expectations Index by Grade and Subject
(0.50 or greater = well-aligned)

Grade	3	4	5	6	7	8	HS
NCEXTEND1 ELAR	0.85	0.81	0.78	0.73	0.71	0.77	0.58
NCEXTEND1 MATH	0.85	0.88	0.55	0.81	0.91	0.84	0.88
NCEXTEND1 SCIENCE			0.77			0.77	0.78

As can be seen from Table 3, all ***NCEXTEND1*** assessments easily surpass the threshold measure. All of the assessments report good, and most quite strong, alignment results with respect to performance expectations.

Overall Alignment

While the SEC approach allows for reporting and consideration of the results in terms of each of these three characteristics of alignment, the most powerful alignment indicator results when content is considered, and alignment is measured at the intersection of all three dimensions. It is the most challenging or rigorous of the alignment indicators because for a bit of content to be considered aligned, it must match the target on all three characteristics or dimensions.

Table 4
Overall Alignment Index by Grade and Subject
(0.50 or greater = well-aligned)

Grade	3	4	5	6	7	8	HS
NCEXTEND1 ELAR	0.49	0.61	0.56	0.61	0.55	0.58	0.52
NCEXTEND1 MATH	0.52	0.57	0.54	0.71	0.47	0.53	0.81
NCEXTEND1 SCIENCE			0.52			0.52	0.61

The resulting overall alignment index (OAI) has a range of 0.00 to 1.00 with 0.50 or higher indicating an acceptable level of alignment (just like the other alignment indices). Overall alignment results are reported in Table 4. Only grade 3 language arts and grade 7 mathematics report OAI's below the 0.50 mark. Even at that, the language arts result is just shy of the 0.5 mark at 0.49 and the grade 7 math result is not far below at 0.47. Nonetheless, following the typical procedures for alignment analyses, these sub-threshold results are examined further to better understand the nature of the alignment issues causing the sub-threshold results.

Table 5

Overall Alignment Index by Grade and Subject

(0.50 or greater = well-aligned)

Grade	OAI	BR	TC	PE
Gr.3 *NCEXTEND1* Reading	0.49	0.28	0.53	0.85
Gr.7 *NCEXTEND1* Math	0.47	0.23	0.53	0.87

Table 5 summarizes the results on all four of the alignment indicators for the two sub-threshold assessments with respect to OAI. Based on the results reported there it seems clear that in each case the key alignment issue centers around balance of representation. Note that the TC and PE measures for both assessments handily exceed the 0.5 threshold while the balance of representation results are substantially below that threshold.

Discussion of Findings

Table 6

Summary of Alignment Indices

English, Language Arts & Reading	BR	TC	PE	OAI
Gr. 3 *NCEXTEND1* ELAR	0.28	0.53	0.85	0.49
Gr. 4 *NCEXTEND1* ELAR	0.36	0.75	0.81	0.61
Gr. 5 *NCEXTEND1* ELAR	0.30	0.71	0.78	0.56
Gr. 6 *NCEXTEND1* ELAR	0.37	0.79	0.73	0.61
Gr. 7 *NCEXTEND1* ELAR	0.31	0.72	0.71	0.55
Gr. 8 *NCEXTEND1* ELAR	0.44	0.72	0.77	0.58
Gr. 10 *NCEXTEND1* ELAR	0.44	0.67	0.58	0.52
Mathematics	**BR**	**TC**	**PE**	**OAI**
Gr. 3 *NCEXTEND1* MATH	0.29	0.60	0.85	0.52
Gr. 4 *NCEXTEND1* MATH	0.50	0.66	0.88	0.57
Gr. 5 *NCEXTEND1* MATH	0.28	0.72	0.55	0.54
Gr. 6 *NCEXTEND1* MATH	0.48	0.75	0.81	0.71
Gr. 7 *NCEXTEND1* MATH	0.23	0.53	0.91	0.47
Gr. 8 *NCEXTEND1* MATH	0.20	0.73	0.84	0.53
Gr. 10 *NCEXTEND1* MATH	0.34	0.84	0.88	0.81
Science	**BR**	**TC**	**PE**	**OAI**
Gr. 5 *NCEXTEND1* SCIENCE	0.43	0.63	0.77	0.51
Gr. 8 *NCEXTEND1* SCIENCE	0.53	0.71	0.77	0.52
Gr. 10 *NCEXTEND1* BIOLOGY	0.27	0.73	0.78	0.62

As indicated by the results presented above, with only a few exceptions the ***NCEXTEND1*** alternate assessments used by the state reveal strong levels of alignment to the North Carolina Extended Content Standards. The results make clear that the design of the alternate assessments attend to the content embedded in those standards, and the implementation of that design yielded in most cases assessment

instruments with good alignment characteristics across the board as measured by the SEC methodology. Appendices D, E, and F provide fine grain results.

Clearly the most challenging aspect of alignment for the ***NCEXTEND1*** alternate assessments as a whole is balance of representation. Even after discarding content related to writing, the Balance of Representation (BR) measures for English language arts/reading tend toward the 0.3–0.4 range, with an exception or two at each end of that range.

The ***NCEXTEND1*** alternate assessments are especially challenged to achieve a strong BR result as it is a very short test, limited to fifteen items. While the extended content standards are also relatively short, certainly shorter than the state's academic standards for the general population, they still cover a range of content areas, leaving the assessment with too few items to adequately assess every topic identified in the extended standards. This can be seen more clearly by considering the descriptive data underlying these low BR measures.

Table 6 provides a more detailed view of the descriptive results underlying the BR indicator. Recall that BR is calculated as the product of two proportions; the proportion of assessment score points attributed to standards-based content, and the proportion of standards content represented on the assessment. In other words, the proportion of the standards on the test multiplied by the proportion of the test on the standards equals the balance of representation.

Table 7

Balance of Representation Detail

Assessment	Prp. Of Test on Stnds.	Prp. Of Stnds. On Test	**BR**
Gr. 3 *NCEXTEND1* ELAR	0.90	0.32	0.28
Gr. 7 *NCEXTEND1* MATH	0.47	0.50	0.23

The results reported in Table 7 suggest that for grade 3 English language arts/reading, the challenge to BR is primarily a function of the representation of the standards in the assessments (0.32), since 0.9 of the assessed topics match topics found in the extended content standards. For grade 7 mathematics, the BR issue is contributed to by both the proportion of the test focused on standards *and* the proportion of the standards reflected on the test.

As reported in Table 7, nine tenths of the grade 3 English language arts/reading alternate assessment is directed toward content identified in the extended content standards, while only 32% of the grade 3 extended standards are actually addressed by the assessment. The 0.90 calculation is derived from the number of standards-based topics addressed by the assessment (9), divided by the total number of topics (standards-based or not) identified in the assessment's content description (10). The 0.32 value is based on the number of extended standards topics in the assessment (9), divided by the total number of topics identified in the grade 3 extended standards (29). Clearly, a 15 item test is going to find it difficult to address 29 topics; particularly if one wants to attend to test construction issues of reliability and validity.

The mathematics assessment reveals a somewhat different, though no less challenging situation for mathematics. The 0.23 BR result reported for grade 7 mathematics is derived by multiplying 0.47 (=7/15), that is 7 standard-based topics out of a total of 15 topics included in the assessment, by 0.50 (=7/14), referring to the 7 assessed topics among a total of 14 mathematics topics identified in the grade 7 extended

content math standards. While having all 14 topics on the assessment match all the topics identified in the standards would offer an outstanding BR measure (1.00), it would undoubtedly be a poor test from a psychometric perspective.

The total number of topics identified in each of these grade-specific extended content standards (29 for grade 3 English language arts/reading and 14 for grade 7 mathematics) provide stark evidence of the tension between breadth of test content and the limitations of assessment reliability and validity requirements, making the 0.5 benchmark for the ***NCEXTEND1*** alternate assessments possibly an unreasonably high expectation. Nonetheless, the BR results reported above do raise questions about the number of non-standards topics identified in the grade 7 math alternate assessment.

At first glance, the results seem quite egregious… "Why should any non-standards content show up on the state assessment, much less more than half the assessment apparently targeting non-standards content?" Despite appearances, the results are not quite as bad as they may seem; but to understand why, we must address issues related to semantics, grain size, and incidental content.

The list of topics identified in the SEC content taxonomies is organized into a two-level hierarchy consisting of 16 *topics* within which a total of 184 mathematics *subtopics* are organized. These 16 topics are used to report coarse-grain results, and report the sum across all subtopics listed within each topic.

All SEC content descriptions are collected at the level of *subtopic*. Alignment analyses are conducted primarily at the coarse-grain level of description. That is, results at the subtopic level are summed up to the coarse-grain *topic* for the purposes of summarizing the descriptive results and calculating the various alignment indices. The balance of representation measure deviates from this pattern insofar as its calculations are based upon results at the fine-grain, i.e. subtopic level of description. Were the BR measure based only on coarse grain, topic level results, in most cases the ***NCEXTEND1*** alternate assessments would have a BR measure of 1.00. That is because in most cases (with one notable exception discussed later) *all* of the *topics* in the extended content standards are assessed in the ***NCEXTEND1*** alternate assessments. However this would be both an uninformative and misleading description of the results. The match at the sub-topic level is more challenging both in terms of the specificity of the match and in terms of the sheer number of subtopics addressed by the standards. It is the more challenging *fine-grain* BR measure used for the alignment analysis, as it provides a more sensitive and informative summary measure. For assessments with a larger number of items the BR measure becomes somewhat less challenging, as there are more items available to increase subtopic representation.

With fewer items, 'incidental' content also becomes a bigger issue, having a greater impact due to the relatively greater proportion of the test each individual content description represents. For example, on a 15 item test a single content description can account for from about 2% to 6% of the test total, depending on how many content descriptions are used to describe the item (see Appendix A for more information on content analysis procedures and content descriptions). In some cases the distinction between one sub-topic and another can be somewhat negligible from the standpoint of the standards, but nonetheless yield a mismatch in comparing the results for the assessments to the standards. For example, the SEC taxonomy has two topics related to ratio and proportion. One is listed under Number Sense, where students are expected to work toward and understanding of ratio and proportion from a Number Sense perspective. The other is listed under Operations, where the student is expected to use ratios and proportions in mathematical operations. The description of the grade 7 extended content standards make reference to the Number Sense usage of proportion, but do not explicitly reference the use of ratio and proportion as part of a mathematical operation (i.e. to carry out an operation using ratio and proportion). The grade 7 math alternate assessment has two items related to ratio and proportion, and the content analysts described these

items under both Number Sense and Operations. Thus one of the eight 'non-standards-based' topics reducing the BR measure for grade 7 is this reference to ratio and proportion under Operations. Another example of this type of 'incidental' content is found in two other grade 7 items that ask the student about the area of a rectangle. 'Area' is a topic identified in the grade 7 math extended standards, yet 'quadrilaterals' or rectangles are not. While the item description did increase the tests BR measure with respect to 'area', it was offset by a reference to 'quadrilateral'. If these examples alone were adjusted to not reflect this incidental content, the BR measure would go from 0.23 to 0.27. Presumably other examples of incidental content would increase the BR measure further.

The other factor affecting the BR measure is standards content not represented on the test. The impact this un-assessed content on BR and the other alignment indicators can be significant if the omitted content reflects to a relatively large portion of the standards content. One example of this has already been discussed relative to writing content on the reading assessment; and the impact this has on alignment is clear from Table 1. For the most part mathematics does not have a similar problem. Grade 7 mathematics however does provide an example that is at least reminiscent of English language arts/reading, if only in its impact on the alignment results.

About 10% of the grade 7 extended mathematics content description addresses topics related to the topic area of Probability (simple probability and sampling, sample spaces). It is the only grade level across the math extended standards that addresses the topic of Probability. While the author does not know if the exclusion of probability content in the grade 7 math alternate assessment was deliberate or an oversight, to better illustrate the impact of this content not being included in the grade 7 assessment, Table 8 reports both the 'modified' and unmodified alignment indicator results.

Table 8

Grade 7 Mathematics Alignment Indices

Alignment Target	**BR**	**TC**	**PE**	**OAI**
Gr.7 *NCEXTEND1* Math	0.23	0.53	0.91	0.47
mod Gr.7 *NCEXTEND1* Math	0.45	0.63	0.92	0.62

Clearly, regardless of incidental content, grain size or other missing content, addressing the issue of probability in grade 7 mathematics has a significant impact on how one interprets the results. Whether that decision is to clearly state that probability is not a topic area that the state wishes to assess with this population of students, or to add questions on probability to future grade 7 ***NCEXTEND1*** alternate assessments, the result would in either case lead toward a significant improvement in the alignment characteristics of the grade 7 math alternate assessment.

There is one final factor to consider with respect to interpreting the balance of representation results for these assessments. Summative assessments are generally not expected to cover the entire standards-based curriculum in any one school year. Indeed many states rotate items from year to year to expand the breadth of the standards-based content assessed over time. Looking across multiple administrations of an assessment program over several years would be necessary to fairly judge the breadth of standards-based content a particular state assessment program measures achievement on.

Summary and Conclusion

This study collected content descriptions of North Carolina ***NCEXTEND1*** alternate assessments covering grades 3-8 for mathematics and English language arts/reading, high school English II and Math I, as well as science assessments for grades 5, 8 and high school biology. The resulting content descriptions were then compared to content descriptions of the North Carolina Extended Content Standards either previously collected (grades 4 -8 mathematics and English language arts/reading extended content standards), or collected as part of this study (grade 3, high school Math I and English II; grades 5 & 8 science, and high school biology extended content standards) using the SEC content analysis and alignment procedures.

The alignment analyses reported above reveal a largely well-aligned set of ***NCEXTEND1*** alternate assessments compared to the extended content standards. Once the English language arts/reading alignment target content is adjusted to remove content related to writing the alignment indicators for topic coverage and performance expectations for all grades show levels of alignment exceeding the 0.5 threshold. Only one English language arts/reading alternate assessment (grade 3) reported an overall alignment index (OAI) less than 0.5, but just barely below 0.49.

The alignment story is almost identical for mathematics, with all of the ***NCEXTEND1*** mathematics alternate assessments reporting topic coverage and performance expectation results above 0.5, and only one assessment below the threshold for OAI (grade 7 OAI = 0.47). Even then, as discussed above, if the grade 7 alignment target is modified to exclude probability content, the OAI measure for grade 7 mathematics increases to 0.62. Alternately, including probability content in future grade 7 ***NCEXTEND1*** alternate assessment would have a similar impact on the alignment result.

Nearly all of the assessments had trouble meeting the 0.5 threshold for balance of representation. However, as discussed, with only a 15 item assessment it is very difficult to assess a sufficient range of content to reach the 0.5 threshold while maintaining test reliability and validity. Given these constraints the ***NCEXTEND1*** alternate assessments all reveal very satisfactory alignment results.

References

Smithson, J.L. (2015). North Carolina Alignment Study Report. Report to the North Carolina Department of Public Instruction. Raleigh, NC: North Carolina Department of Public Instruction.

Porter, A.C. & Smithson, J.L. (2001). Defining, developing, and using curriculum indicators. CPRE Research Report Series RR-048. Philadelphia, PA: University of Pennsylvania, Consortium for Policy Research in Education.

Webb, N.L. (1997). Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education. Research Monograph No.6. Washington, DC: Council of Chief State School Officers.

Appendices (to the Main Report)

Appendix A: Describing the Academic Content (Addendum)

Appendix B: Diagnostic Analyses of Selected Assessments

Appendix C: Procedures and Materials for Conducting Content Analyses (external file)

Appendix D: Mathematics Content Viewer (external file *content added to reflect* ***NCEXTEND1*** *data***)**

Appendix E: Reading Content Viewer (external file *content added to reflect* ***NCEXTEND1*** *data***)**

Appendix F: Science Content Viewer (external file *content added to reflect* ***NCEXTEND1*** *data***)**

APPENDIX A (Addendum)

Describing the Academic Content of the North Carolina *NCEXTEND1* Alternate Assessment Program

Appendix A of the full alignment report describes the process for conducting content analyses in some detail, and the interested reader is directed there for a discussion of the content analysis procedures. This supplement to Appendix A describes the particulars involved in the collection of the content analysis data for the ***NCEXTEND1*** assessments that differed from the original study of the state's general assessment program.

Whereas the original data collection of assessments and standards data as described in Appendix A was conducted in conjunction with, and following a one-half-day content analysis training workshop. For the ***NCEXTEND1*** assessments, remote data collection procedures were followed. Remote procedures refers to the analysts of each content analysis team working remotely from one another, using web, phone and other electronic media (e.g. email) for team discussions. This approach can be more cost efficient when large numbers of analysts are not required and the materials for analysis can be securely disseminated among team members to the satisfaction of the state's assessment director. For these analyses, the testing materials were delivered by express courier to the analysis team leader and then distributed to analysis team members as relevant to their subject area focus.

Each team received an initial introduction to the task and materials, and provided the opportunity to discuss and become familiar with the testing materials and procedures used for testing. All team members were veteran analysts, and did not require training. Most team members had worked together previously on content analysis tasks. All team members had the relevant content expertise, and the majority of members had experience as curriculum specialists with one or another state education agency

The independent results for each team member's analysis were sent to the project lead for entry into the SEC data system. Once the results from each team member were received and entered for a given assessment, the team was informed that their analysis results were ready for review. Analysts were then able to access through a password protected login the results for each team member for each assessment item for the assessments analyzed. Team members discussed items either by conference call or email, depending upon team members' schedules and preferences. Each analyst had the ability to add, subtract or revise their results based on the review and group discussion.

Once the reviews were completed, the data was collected and processed in preparation for alignment analysis just as done for the earlier alignment study for the general statewide assessments. From that point on, the analysis procedures followed were identical to those utilized in the prior study.

APPENDIX B: Diagnostic Analyses Using SEC Data

Purpose

Just as with assessment data, SEC results can be used for either summative or formative purposes. The primary goal of the main report was to provide summative results of our alignment analyses and determine how well state assessments meet the alignment requirements of a standards-based accountability system. The purpose of this section is to illustrate how the SEC data can be used for formative purposes, i.e. to *improve* the alignment of one or another component of the state's assessment program. This diagnostic analysis utilizes fine-grain descriptive and alignment results in order to identify areas of weak alignment and suggest strategies for possible improvement.

Selection

As noted in Table 5 of the study report, two grade levels of mathematics assessments and one reading assessment reported relatively weak, overall alignment (OAI) measures. This is the most demanding of the alignment measures, as it incorporates all three dimensions together. Among the sub-measures based on individual alignment dimensions, only one assessment reported a below-threshold measure. That assessment, grade 3 mathematics will serve as the focus of this example of conducting a more fine-grain 'diagnostic' analysis designed to identify specific areas of weak alignment in order to inform strategies for improving alignment in future versions of the assessments.

Note that the necessary tools and information needed to conduct a diagnostic analysis of any grade, or grade-specific form is possible using Appendices D – F (requires Excel 2010 or newer with macros enabled running on a PC platform). The following analyses provide a step-by-step approach that could be replicated for any given selection of alignment targets and assessment descriptions (either grade-level or form-specific) using the data sets supplied in the Appendices.

It is worth noting that the procedures for conducting formative analyses of assessment alignment and instructional alignment are similar in that both use alignment indicators and content descriptions to identify strategies for improvement. Due to similarities in the diagnostic process, the formative analysis of assessment alignment presented in this report will have many parallels to the structure of a teacher personal alignment report.

Formative Analyses

Formative and summative analyses alike combine descriptive and analytic results. The primary difference between summative (evaluative) and formative (diagnostic) analyses of SEC data relate to the level of detail at which the analyses are conducted. In SEC terminology this distinction is referred to as coarse grain and fine grain. Coarse grain results provide an efficient summary of alignment characteristics that offer a reasoned, evidence-based approach to assessing the adequacy of state efforts to implement an aligned standards-based system.

Behind these summative results lies a large body of empirical, descriptive data concerning the specific assessments and forms through which the state assessment program is delivered. These descriptive data provide the basis for the detailed analyses of the alignment characteristics of the state's mathematics and reading assessments. These data can support formative analyses designed to identify strategies most likely to lead to improved alignment between any two content descriptions.

Summative alignment results generally provide the starting point for narrowing the focus in searching for areas of misalignment as is done in formative analyses. While the alignment indicators at both coarse grain and fine grain levels assist in targeting those areas where improvements in alignment are needed,

the ultimate goal is to identify or target the specific descriptive results that will provide the most information for informing alignment improvement efforts.

Thus we begin with Figure 1, which contains the summative alignment results for grade 3 mathematics from the body of this report. As can be seen in Figure 1, of the four indicators presented, two are below the 0.50 threshold and two above. The focus here will be on the two below the threshold because these weak summary measures suggest where the majority of the misalignment occurs.

Figure 1

Grade 3 Mathematics
Coarse Grain Alignment Indices Summary
(0.50 or greater = well-aligned)

	BR	TC	CC	OAI
AI	0.68	0.68	0.41	0.40

BR Balance of Representation
TC Topical Concurrence
CC Cognitive Complexity
OAI Overall Alignment Indicator

As discussed in the main report, the OAI is a composite measure, sensitive to variation on three dimensions of alignment: 1) balance of representation (BR), 2) topic coverage (TC), and 3) performance expectations (PE). Setting OAI aside for the moment, we see that only one of the three underlying alignment measures is low - the indicator for performance expectations. This strongly suggests that the fundamental alignment problem for grade 3 mathematics is going to relate to non-alignment in the area of performance expectations.

This is borne out in examining the descriptive results underlying the CC indicator. Figure 2 shows the marginal) values for the five performance expectations (i.e., the overall emphasis of the standards and assessments on cognitive demand categories without regard to topic emphasis) for the content descriptions of the state's grade 3 mathematics standards (in red) and assessments (in blue)

The chart in Figure 2 indicates that the assessment lags one performance category behind the standards; as can be seen by imagining how similar the bars would look if the blue bars were simply moved one level down in the chart. The two descriptions would then be very highly aligned with regard to performance expectations. In fact under those circumstances the coarse grain cognitive demand indicator measure would be a remarkable 0.88.

Figure 2

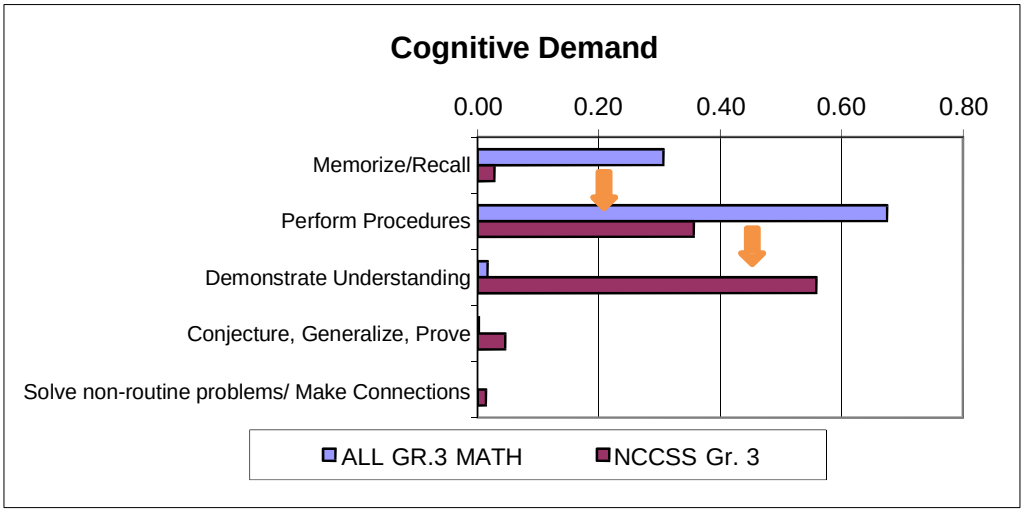


Figure 3

Cognitive Complexity	
Number Sense	0.27
Operations	0.53
Measurement	0.47
Consumer Applications	NA
Basic Algebra	0.46
Advanced Algebra	NA
Geometric Concepts	0.25
Advanced Geometry	NA
Data Displays	0.69
Statistics	NA
Probability	NA
Analysis	NA
Trigonometry	NA
Special Topics	NA
Functions	NA
Instructional Tech.	NA

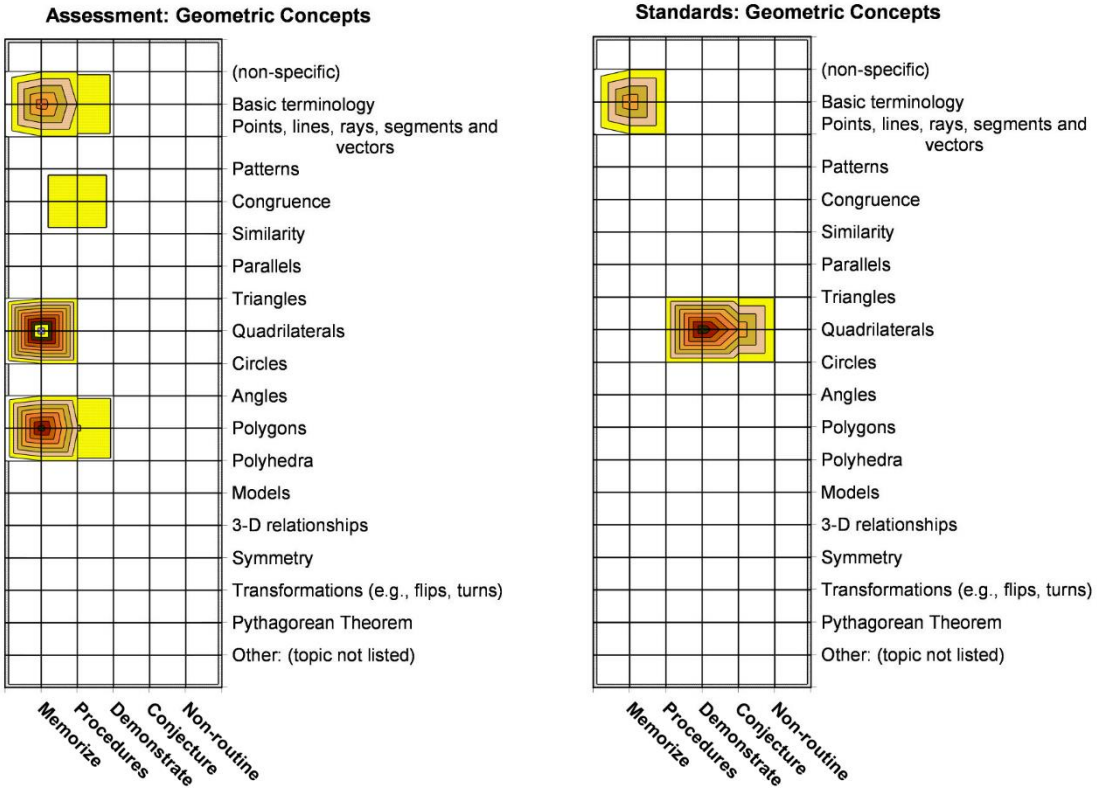
While the results reported in Figure 2 make clear that an insufficient number of items target the performance expectation 'demonstrate understanding', it does not give us any information about which topics would be best targeted for this shift in performance expectation.

It is possible to drill down on any one of the four summative indicators to get a more detailed look at the relevant alignment indicator measure for each topic area. Doing so allows one to identify those topic areas with low indicator measures to guide the next steps in the process.

Figure 3 reports the fine grain, topic area results for performance expectations.

As can be seen, many of the topic areas listed in figure 3 are not applicable (NA), indicating that neither the standards nor assessment descriptions contain content for that area. Of the six topic areas assessed, two exceed the 0.5 threshold, two additional measures are quite close, and two (Number Sense, and Geometric Concepts), reveal very low performance expectations measures. It is to these two topics that we turn next.

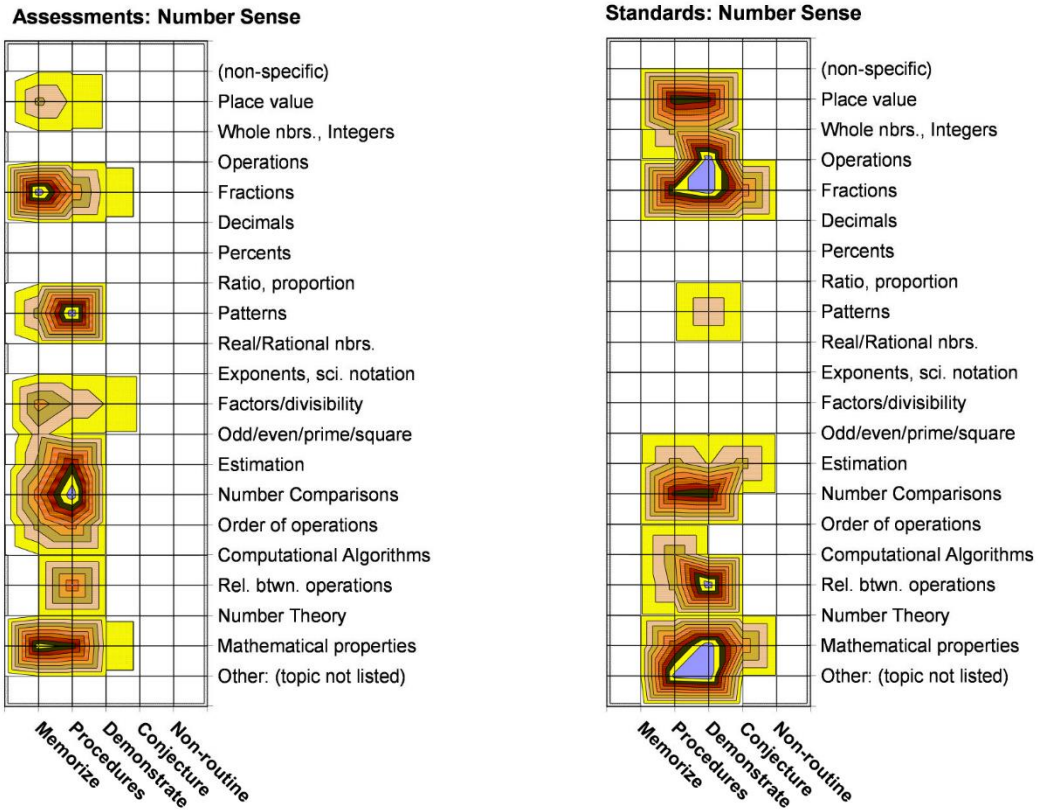
Figure 4



At this point in the analysis it is useful to return to content maps to get a visual display of the fine grain content embedded within the grade 3 standards and assessments with respect to Geometric Concepts and Number Sense. The maps displaying these results are reported in Figures 4 and 5 respectively.

As can be seen by the fine grain content maps for Geometric Concepts in Figure 4, only two sub-topics are addressed in the grade 3 standards; basic terminology and quadrilaterals. The assessment map on the left shows a good match on basic terminology at the recall level, but the assessment also touches on a couple of topics not in the standards (Congruence, and Polygons) and targets quadrilaterals at a different performance level than is addressed in the grade 3 standards (Recall instead of Demonstrate Understanding). Here is a good example of an assessment addressing the correct sub-topic at the wrong performance level. As demonstrated in this analysis, this is exactly the type of mismatch to which we were alerted by the SEC summary indicators presented in Figure 1.

Figure 5



The content descriptions for Number Sense, displayed in Figure 5, are more complex, as Number Sense is clearly an important topic area for grade 3. Once again we see a pretty good match up on sub-topics. One can identify an occasional topic that is reflected in one description but not the other, but generally topics covered in the standards are also covered in the grade 3 assessments. There are six sub-topics in the standards that show a moderate to strong degree of emphasis; place value, operations, fractions, number comparisons, relations between operations, and mathematical properties. By comparison, the assessment touches on content for five of these six sub-topics; operations being the lone sub-topic of the six not assessed to some degree.

The mismatch that occurs once again is with the performance expectations addressed for specific sub-topics. With the lone exception of Number Comparisons, the performance expectations patterns of emphases for assessments tend to be pitched one or two levels below that of the relevant standards.

Based on this analysis, if the state wanted to achieve relatively large gains in alignment while making relatively few changes to the mathematics assessments, the most efficient strategy would be to replace Recall items with items that require students to provide evidence that they understand the concepts at issue in the assessment item.

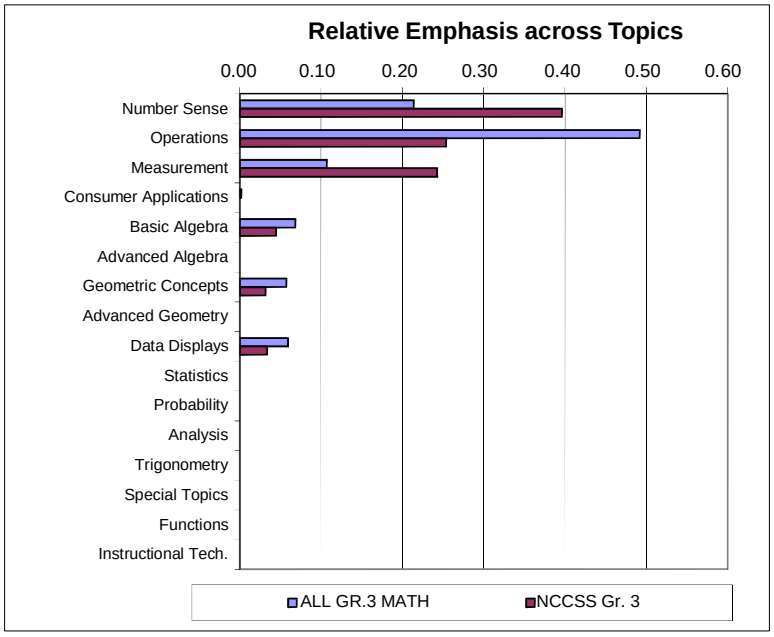
Balance of Representation

Thus far our analyses have focused on topic and cognitive demand in comparing the grade 3 assessments to the grade 3 standards. It will be recalled that BR refers to the relative emphasis placed on one or another topic. When examining BR at the fine grain level of detail it is important to understand that what constitutes relative emphasis changes, depending on whether one is talking about assessment, standards or instruction. For assessments relative emphasis is defined in terms of score points, i.e. what portion of a student's score is based on this content. For standards relative emphasis is more a function of the relative number of times a given bit of content as defined by SEC is addressed across the content strands of the standards. With instruction, relative emphasis is simply the proportion of instructional time spent on a given topic.

While the overall indicator measure for BR was good (0.68), it is worth looking at the fine grain results, topic by topic, to identify instances of especially low alignment. Figure 6 shows the relative emphasis given to each topic area for both the grade 3 assessment and standards..

As revealed in Figure 6, both the grade 3 mathematics assessments and standards show a strong emphasis on three primary topic areas: Number Sense, Operations, and Measurement. However Figure 6 also reveals that the relative emphases each description gives to these three topic areas is quite different. While the state grade 3 math standards appear to place the greatest emphasis on number sense (0.40), the assessment places the greatest emphasis on Operations (0.50). This over-emphasis on Operations (relatively speaking) comes largely at the expense of emphasis on Number Sense and Measurement.

Figure 6



While the emphasis on Operations may be based on a legitimate assessment design decision, these results indicate that practically any increase in the assessment of content related to Number Sense or Measurement would lead to improved alignment..

Discussion

The SEC data set provides a sound framework for collecting content descriptions that enable coarse- and fine-grain, quantified comparisons based upon a simple iterative algorithm that provides alignment indicators to summarize alignment characteristics and relationships. While thresholds can be set at reasonable levels based on rational argument (e.g. 0.50) or normative findings (i.e. above or below average), the implications for assessment design or classroom instructional practice remains more art than science. In other words, SEC data is designed and intended to *inform* policy decisions, not *make* them. In any analysis it is the reader, or the organization's decision-makers that must digest the results, and along with other factors, make decisions intended to improve the learning opportunities of those students under their purview. This is the case whether the data is being examined by a policy-maker, assessment director, district/school administrator, or teacher. In each case it is essentially the same analytic process.

While the analyses presented in this section make clear what strategies *could* be employed to move toward better alignment of the grade 3 assessments; nothing reported here can alone answer the policy questions of what strategies *should* be used to improve alignment of these or any other assessments; or even whether the alignment of these assessments need to be improved at all. The reader must assimilate the descriptive results and consider them within the contexts of practical feasibility.

Though not explicit, there is a logic and procedure underlying the analysis done in this section that can be generalized for use with any two content descriptions of interest. Those procedures are outlined below to enable relevant department staff to review the study results in more detail, using the interactive appendices A-C. Thus in describing the general procedures for conducting detailed alignment analyses reference will be made to particular charts, tables or content maps to be found in the relevant subject-based appendices.

Analyses begin with the summative coarse grain results for the selected target and assessment. It will be recalled that there are four summative measures; a general overall indicator (OAI) that is based upon, or sensitive to variation on each of three foundational 'dimensions' of alignment. These are Balance of Representation (BR), Topic Coverage (TC), and Performance Expectations (PE). By convention WCER researchers have set 0.50 as a threshold measure for adequate alignment; implying that measures below 0.50 deserve further attention, particularly if alignment improvement is a goal. The OAI measure serves to identify any assessments that might deserve a deeper look. For this study, alignment results indicated that the grade 3 math assessment had the weakest alignment scores among the assessments examined. For that reason, grade 3 math was selected to use as an example for this description of diagnostic analyses.

At whatever level of examination being conducted, analyses begin with the OAI measure, followed (for those cases that fall below the selected threshold) by review of the marginal measures for each alignment dimension, i.e., BR, TC, and CC. Here again the results are reviewed against a desired threshold (presumably, but not necessarily 0.50), with deeper examination guided by those indicators that fall below the threshold.

For each alignment indicator measure there is at least one, if not two descriptive displays available as an aid to understanding the data represented by that indicator. Depending on the measure, descriptive results are reported in bar charts, or content maps, or both. In the subject specific Excel-based content viewers that accompany this report, the worksheets display the results necessary for both coarse grain and fine grain analyses of alignment results. The 'Alignment Table' worksheet supplies all of the alignment indicators, both coarse grain and fine grain, organized into an alignment table that summarizes all of the alignment results for a given pair of content descriptions (consisting of an alignment target and the assessment (or instruction) being examined relative to that target).

Fine grain alignment indicators are reported by topic area in the alignment table. Both coarse and fine grain content maps content descriptions are reported in the 'ContentMap' worksheet. Finally, bar charts

detailing Balance of Representation and Performance Expectations are reported in the Marginals worksheet. The ContentMap worksheet is used to select which content descriptions are to be analyzed.

Once a selection is made from among the BR, TC, and PE indicator results, the relevant column for the selected indicator can be reviewed in the alignment table to guide selection of specific topic areas to review more closely. It is the review of the descriptive results at this fine grain level that the alignment analysis procedures are designed to facilitate, assisting one in targeting weak areas of alignment in order to consider strategies for improving alignment, or at least better understand the nature of the alignment relationship between the descriptions being reviewed. The overall process is graphically summarized in Figures 7 through 10 below. The table displayed in Figure 8 is an alignment table, from which all alignment indicators are drawn. The content viewers making up Appendices A-C all have interactive alignment tables that allow for the reporting of all the alignment indicators for all of the grade specific assessment forms analyzed for the study. Figure 8 highlights two rows in the table with low alignment indices; Number Sense and Geometric Concepts one is labeled 'A' and the other 'B', to indicate which results from Figure 8 are being detailed in Figures 9 & 10. The maps displayed in Figures 9 & 10 can similarly be accessed through the content viewers, which can be used to generate both coarse grain and fine grain content maps depicting the content of any assessment or standards target selected.

Figure 7

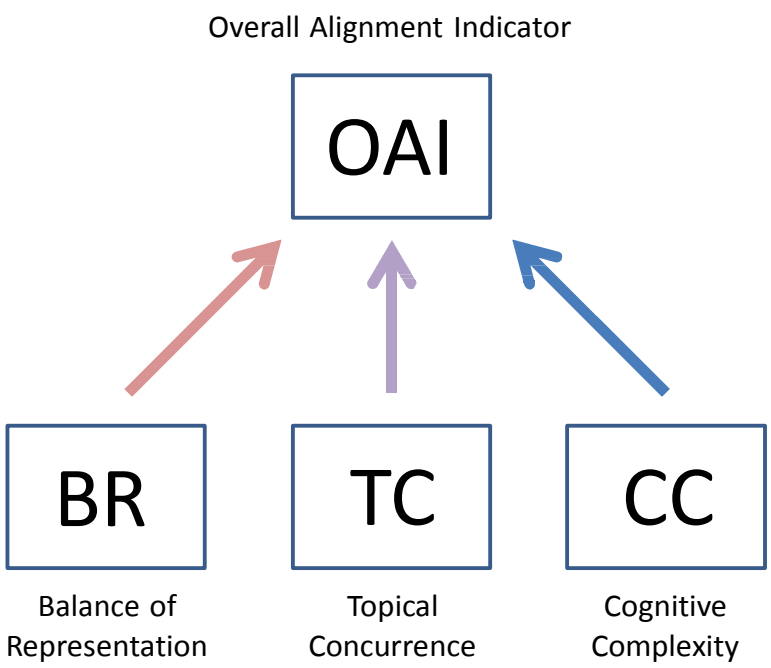


Figure 8

Alignment Analysis Summary Table

Coarse Grain	0.40	0.68	0.68	0.41
	OAI	BR	TC	CC
			(Topics)	(Cog. Dmnd.)
ALL GR.3 MATH TO: NCCSS Gr. 3	Alignment	Balance of Representation	Categorical Concurrence	Cognitive Complexity
Number Sense	0.15	-0.18	0.51	0.27
Operations	0.44	0.24	0.57	0.53
Measurement	0.32	-0.14	0.68	0.47
Consumer Applications	NA	0.00	NA	NA
Basic Algebra	0.40	0.02	0.65	0.46
Advanced Algebra	NA	0.00	NA	NA
Geometric Concepts	0.17	0.03	0.67	0.25
Advanced Geometry	NA	0.00	NA	NA
Data Displays	0.62	0.03	0.82	0.69
Statistics	NA	0.00	NA	NA
Probability	NA	0.00	NA	NA
Analysis	NA	0.00	NA	NA
Trigonometry	NA	0.00	NA	NA
Special Topics	0.00	0.00	NA	NA
Functions	NA	0.00	NA	NA
Instructional Tech	NA	0.00	NA	NA
Fine Grain	***0.30***	***0.68***	***0.56***	***0.40***

A

B

Figure 9

Fine Grain Content Descriptions

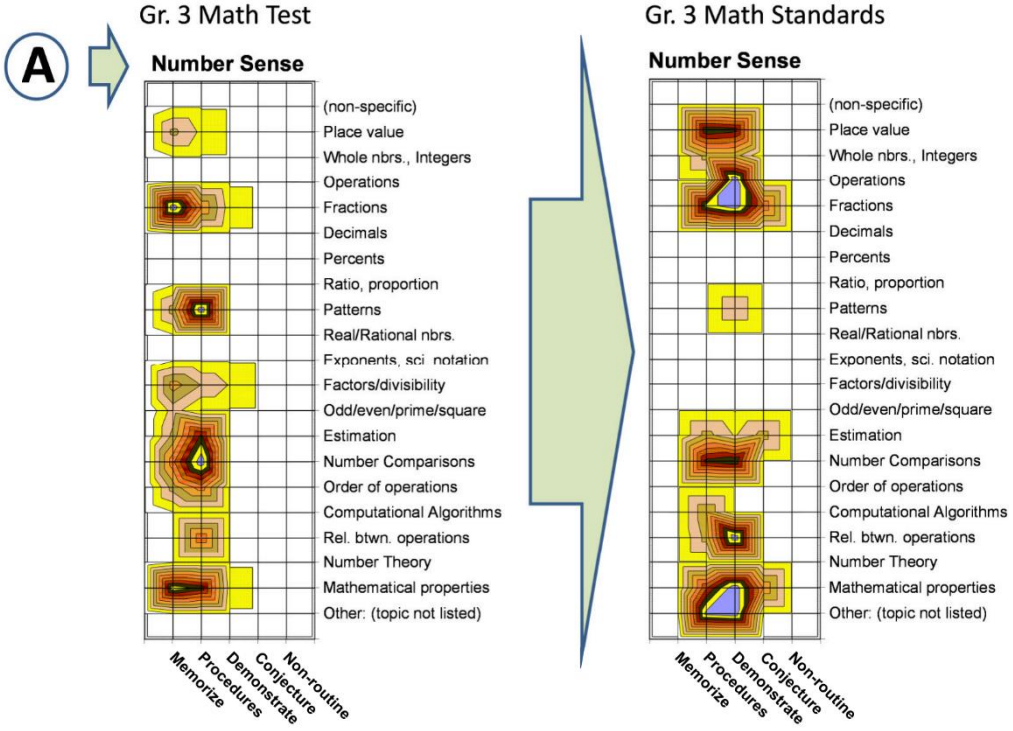
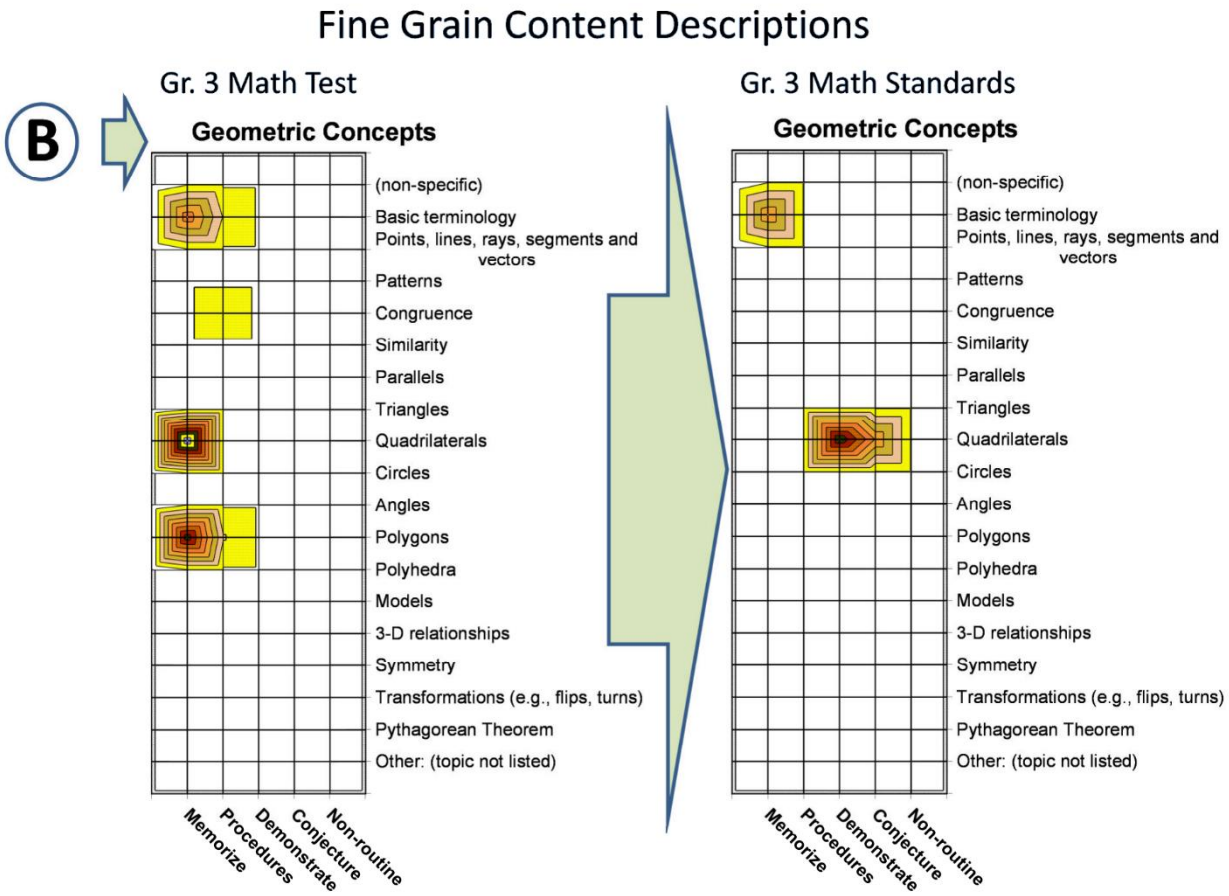


Figure 10



Using Appendices D-F

Appendices D-F house subject-specific 'content viewers'. These are macro-enable Excel-based files that provides an interactive user interface to generate alignment tables, content maps and charts with marginal measures (all the data sources needed to conduct a fine-grain analysis for any given subject and grade level, or even grade level form.

The files must be run on PC-platforms with macros-enabled in order for the interactive features to work. Each viewer contains an introduction and set of instructions to get the user started. That with the examples of Appendix B and familiarity with the descriptive displays explained in Appendix A provides the necessary background information to examine the full data set used for the study in-depth.