A Summary Report

Findings of an Alignment Study in Delaware Student Testing Program Reading and Writing for Grades 2, 4, 6, 7, and 9

Prepared by

Rhonda Nowak
Loyola University New Orleans

Assessment and Analysis
Assessment and Accountability Branch
Delaware Department of Education

September, 2005

OFFICERS OF THE DELAWARE DEPARTMENT OF EDUCATION

Valerie A. Woodruff
Secretary of Education
Nancy J. Wilson, Ph.D.
Deputy Secretary of Education
Susan Keene Haberstroh
Executive Assistant

Robin R. Taylor, M.Ed.
Associate Secretary, Assessment and Accountability Branch
Lewis Atkinson, III, Ed.D.
Associate Secretary, Adult Education and Workforce Development
Martha Brooks, Ed.D
Associate Secretary, Curriculum and Instructional Improvement Branch
Dorcell Spence
Associate Secretary, Finance and Administrative Services

Wendy B. Roberts, Ph.D.
Director, Assessment and Analysis Work Group
Darlene J. Bolig, Ed.D.
Helen C. Dennis, M.A.
Jeffery Fleming, M.S.
Katia Foret, Ph.D.
James F. Hertzog, M.Ed.
Patsy G. Kersteter, Ph.D.
Joann F. Prewitt, M.A.
Liru Zhang, Ph.D.

***Support Staff*:**
Gail R. Truxon
Elaner M. Brown
Barbara F. O'Neal
Theresa L. Kreer
Kimberly K. Rodriguez

Document Control No. 95-01/05/11/02
This report is available on the World Wide Web at http://www.doe.k12.de.us/aab.

Table of Contents

List of Tables

A Summary Report

Findings of an Alignment Study in Delaware Student Testing Program
Reading and Writing for Grades 2, 4, 6, 7, and 9

Introduction

Alignment is an important attribute for educational systems. Although the value of aligning content standards and assessments has been known for some time (Carroll,1963; Cohen, 1987), alignment gained prominence in the early 1990s with the advent of improved standards in mathematics (NCTM, 1989) and widespread efforts at systemic reform (Smith & O'Day, 1991). Educators have increasingly recognized that if policy elements are not aligned, the system will be fragmented, will send mixed messages, and will be less effective (Consortium for Policy Research in Education, 1991; Newmann,1993). For example, the Systemic Initiatives program of the National Science Foundation (NSF) directed participating states, districts, and regions to set ambitious goals for student learning by developing coherent policy systems that focused, in part, on assessments aligned with those goals. The *Improving America's Schools Act* explained how assessments were to relate to standards: "…such assessments (high quality, yearly student assessments) shall … be aligned with the State's challenging content and student performance standards and provide coherent information about student attainment of such standards …" (U.S. Congress, 1994, p. 8). The U.S. Department of Education's explanation of the *Goals 2000: Educate Act* (1994) and the *Elementary and Secondary Education Act* (1994), which includes Title I, identified the alignment of curriculum, instruction, professional development, and assessments as key performance indicators for states, districts, and schools striving to meet challenging standards. More recently, the new *No Child Left Behind Act* continues the requirements from the former Title I legislation to the 2002-2003 school year through the 2004-2005 school year for language arts and mathematics and applies the requirements to science, beginning in 2007-2008 (Title I, Part A, Section IIII). As before, assessments are to be aligned with academic content and student achievement standards (Council of Chief State School Officers, 2002).

The *No Child Left Behind Act* requires that states report student progress toward meeting the state standards by grade for grades 3 through 8 even though the content standards developed by grade cluster. Early 2005, the Grade Level Expectations were developed based on the Content Standards for every grade from 2 through 11 in mathematics. These expectations will be used at the local level to develop curriculum and instruction. At the state level they will be used to develop or identify items for inclusion on the DSTP mathematics. Although the expectations for students by the end of a grade cluster have little changes, the goals and expectations by the end of each grade are specified.

The Delaware Student Testing Program (DSTP) is a statewide, mandated assessment program that is designed to measure student progress toward Delaware Content Standards. The Delaware Student Testing Program (DSTP) started in 1998 for grades 3, 5, 8, and 10. Beginning in spring 2002 the DSTP was expanded in reading for grades 2, 4, 6, 7, and 9; and writing for grades 4, 6, 7, and 9. The grades 2, 4, 6, 7, and 9 reading assessments consist of two parts: items developed by the contractor for Delaware that measure the Content Standards and items selected from the Stanford Achievement Test (SAT), abbreviated version of Reading comprehension. Items developed by contractor are reviewed and approved by Delaware educators before the field test. Three types of items: multiple-choice (MC), short answer (SA), and extended constructed-response items (ER), are used. The writing assessment includes a stand-alone writing prompt and a text-based writing prompt. Student performance on the DSTP reading (grades 2, 4, 6, 7, and 9) and writing (grades 4, 6, 7, and 9) will be reported in five performance levels -- Distinguished, Exceeds the Standard, Meets the Standard, Below the Standard, and Well Below the Standard, in 2006.

English Language Arts Grade Level Expectations - The original English/Language Arts Content Standards were adopted in 1995 and included Performance Indicators that described what students should know and be able to do by the end of a cluster of grades: K-3, 4-5, 6-8, and 9-10. In 2005, Grade Level Expectations (GLE's) were developed to provide clarity for classroom teachers as to their specific responsibilities in guiding students to meet and exceed the English/Language Arts Content Standards. These Grade Level Expectations will be used at the local level to develop curricula and assessments to monitor progress towards meeting and exceeding the standards. At the state level they will be used to develop or identify items for inclusion on the Delaware Student Testing Program (DSTP). End of cluster expectations have not changed; what has changed are the descriptors at each grade level that provide teachers guidance on how to achieve those standards. The document has since been reviewed by several groups of English/Language Arts teachers from across the state and by national consultants.

The Alignment Study was conducted for the Delaware Student Testing Program (DSTP, 2003 Form) for English Language Arts in Grades 2, 4, 6, 7, and 9. A two-day alignment analysis institute took place on April 4-5, 2005 at the Delaware Technical and Community College in Dover. The team of reviewers consisted of teachers with English Language Arts expertise from Delaware, as well as English Language Arts experts from Louisiana and Wisconsin (Appendix A). The mix provided some balance between external reviewers without extensive prior knowledge of Delaware Content Standards and the DSTP, and teachers who were more knowledgeable about the Delaware Standards and state assessment. An external consultant facilitated the group of Delaware teacher reviewers for Grades 2, 4 and 6; a second external consultant facilitated the group of teacher reviewers for Grades 7 and 9. The groups consisted of six teachers and the group facilitator for Grades 2, 4, 6 and seven teachers and the group facilitator for Grades 7 and 9 (one teacher did not return on the second day for this group).

Training was provided the morning of April 4 for all reviewers to understand Webb's Alignment Model and Criteria and how to use the Web Alignment Tool (WAT). Reviewers were trained to identify *Depth of Knowledge* (DOK) levels for state reading performance indicators/grade level expectations and reading-related DSTP test items. Additional whole-group training took place the afternoon of April 5 for state writing performance indicators/grade level expectations and writing-related DSTP test items. During the training, all participants reviewed definitions and criteria for the four DOK levels and examples of performance indicators and test items at each level. Following the whole-group training, the large group of reviewers divided into Grades 2, 4, 6 and Grades 7, 9. Each sub-group briefly discussed grade-specific alignment procedures. The alignment study process for each sub-group included four steps:

- Step One - Reviewers individually read the state reading performance indicators/grade level expectations for one grade and coded DOK levels using the WAT. When all individual entries had been made on the computers, a coding report was used to facilitate the reviewers in reaching consensus on appropriate DOK levels for each reading performance indicator/grade level expectations for that grade level. A similar coding and consensus process was subsequently used for all designated grade levels within each sub-group. In completing this step of the alignment process, reviewers gained increasing familiarity with the Delaware Content Standards and the four DOK levels.

- Step Two – After the performance indicators/grade level expectations had been coded, the team of reviewers turned their attention to the reading-related DSTP test items, which consisted of reading passages with multiple-choice questions and constructed responses. Some group discussion took place regarding aligning specific test items to one appropriate primary performance indicators and up to two secondary performance indicators. The reviewers then independently coded a few test items for one grade level and informally discussed their assigned DOK levels and corresponding performance indicators. Reviewers were allowed to code only one DOK level to each assessment item, even if the item corresponded to more than one performance indicator.

- Step Three – This step involved reviewers independently coding test items and identifying appropriate primary and secondary performance indicators. Reading-related test items for one grade level were coded on April 4 and reading test items for the other grade levels were coded on April 5. After all of the reading DSTP items had been coded, group discussion about the alignment process allowed reviewers to calibrate their coding. However, agreement was not the purpose for the process. Even though the average of reviewers' coding was used, the variation among reviewers was considered legitimate because the true DOK level for an item might fall somewhere between the assigned values. Moreover, the variation could signify the robustness of an item that can legitimately correspond to more than one objective or a lack of clarity in how the performance indicators/grade level expectations were written. Following the writing

training on April 5, a similar process was used to individually code writing-related DSTP test items, followed by group discussion.

- Step Four – On the afternoon of April 5, debriefing sessions took place within each sub-group, in which reviewers were encouraged to offer their opinions on the quality of the Delaware performance indicators/grade level expectations and DSTP test items. Reviewers could also identify any concerns they had regarding the alignment process itself.

Alignment Criteria

The alignment of an assessment to the corresponding expectations for student learning is an essential attribute for an effective standards-based educational system. Alignment is the degree to which assessments and expectations are in agreement. Alignment can be legitimately improved by making appropriate changes of expectations and/or assessments. The quality of this relationship or the degree of alignment between an assessment and the expectations can be determined using multiple criteria described in detail in a National Institute of Science Education (NISE) research monograph, *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education* (Webb, 1997).

The objective of the current analysis was to use Webb's four Depth of Knowledge levels (1997) in order to determine alignment between the Delaware Student Testing Program (DSTP) and the Delaware English Language Arts (reading and writing) Performance Indicators/Grade Level Expectations for Grades 2, 4, 5, 7, and 9.

Categorical Concurrence

An important aspect of alignment is whether both standards and assessments address the same content categories. The *Categorical Concurrence* provides a very general indication of alignment if both documents incorporate the same content. The criterion of *Categorical Concurrence* is met if the same or consistent categories of content appear in both standards and assessment. This criterion is judged by the number of test items for each standard. It is assumed that at least six items should be included in an assessment per standard to achieve an acceptable level. The minimum six items per standard were determined based on the procedure developed by Subkoviak (1988). For example, assume that the reliability of one item is .10; six items would produce an estimated agreement coefficient of .63 when the cut-score is near the mean. This indicates that about 63% of the examinees would be consistently classified as master or non-master if two equivalent test administrations are employed without considering other factors, such as the chance probability, measurement error. However, if sub-score is reported for each standard, six items are far from adequate to achieve reliable and stable results.

Depth of Knowledge Consistency

The alignment does not only mean the agreement of content categories covered by both standards and the corresponding assessment, but also the complexity of knowledge required by each. *Depth of Knowledge* indicates the consistency of what is elicited from students in the assessment and cognitive demands that students are expected to know and can do in the standards. The criterion of *Depth of Knowledge (DOK)* is met if 50% of the items (or score points) are at or above the knowledge level specified in the corresponding standard. The minimum requirement for DOK is based on the assumption that the cut-score is set at a minimum of 50% for each standard so that students are expected to successfully answer some items at or above the DOK levels. If 40% to less than 50% of items are judged at or above the DOK levels specified in the standards, the criterion of *Depth of Knowledge* is weakly met (Table 1).

Four levels of *Depth of Knowledge* were defined for different content areas (for details, please see Appendix B). Although, the definition of each level varies some by content area, in general Level 1 represents *Recall* or *Reproduction of Information*, Level 2 represents *Uses of Skills and Concepts*, Level 3 represents *Strategic Thinking*, and Level 4 represents *Extended Thinking*. In applying each of these *Depth of Knowledge* levels, reviewers should think about what a typical student at the grade level being assessed would be expected to know or to perform. Thus, the four levels vary some by grade of a given content area.

Range of Knowledge Correspondence

The breadth of content knowledge should be comparable on both standards and assessments. The *Range of Knowledge* criterion is used to judge whether the span of knowledge expected in a standard is the same or corresponds to the span of knowledge covered by the test items. This criterion considers the number of test items or activities to the corresponding objectives of each standard. If 50% of the objectives have at least one related test item, the *Range of Knowledge Correspondence* is acceptable, which is based on the assumption that an assessment should measure over half of the content domains specified in a standard and each standard is given equal weight. The minimum requirement of 50% for this criterion and the balance distribution of test items for each will increase the likelihood for students to demonstrate what they know and can do. States may choose a more rigorous requirement for this criterion. However, the *Range of Knowledge Correspondence* is more difficult to attain with large number of standards and objectives. If 50% or more of the objectives for a standard have one corresponding assessment item, the *Range of Knowledge* criterion is met. If 40% to 49% of the objectives have at least one test item, the criterion is weakly met (Table 1).

Balance of Representation

In addition to comparable depth and breadth of content knowledge, alignment also requires an even distribution of content domains in both standards and assessments. The *Balance of Representation* criterion considers the number of items corresponding to objectives (number of objectives within a standard hit), but does not take into consideration the distribution of test items/activities (hits) across these objectives. The index computed using the formula below is used to judge the *Balance of Representation* criterion indicating the degree of balance of test items among objectives of a standard.

Balance of Representation Index $= 1 - (\sum \left| 1/Ok=1 - I\,k\,/H \right|)/2$

Where

O = Total number of objectives hit for the standard
I k= Number of items corresponding to the objective
K = Objectives
H = Total number of items hit for the standard

The index value of 1.0 signifies perfect balance, in which the corresponding items (or hits) are equally distributed among the objectives and each objective is measured with the same number of items; the index value approaching 0 signifies a large number of corresponding items (or hits) measure one or two objectives. If most of the items related to one objective and only one item related to each of the remaining objectives (with a unimodal distribution), an index value less than .50 is expected. A bimodal distribution of test items has an index value about .55 to .60. If all objectives are measured with at least two items, the index value is .70 or higher. The acceptable level on this criterion is .70; the index values greater than .60 and less than .70 is considered weakly met (Table 1).

Source of Challenge Criterion

The *Source of Challenge* criterion is used to identify items on which the major cognitive demand is inadvertently mismatched with the targeted skill, concept, or application. Cultural bias or specialized knowledge could be the source of problems. Such item characteristics may result in bias or unfairness of assessment. The comments from reviewers will only be studied by the Test Development Committees due to the security of test items.

Table 1. Alignment Levels for the Four Criteria

Alignment Level	Categorical Concurrence	Depth of Knowledge	Range of Knowledge	Balance of Representation
Acceptable	6 items per standard	50%	50%	.70
Weak		40% - 49%	40% - 49%	.60 - .69
Unacceptable	less than 6 items per standard	less than 40%	less than 40%	less than .60

Results of the Alignment Analysis

The current study took item format into account and used the maximum score point as weight (1 for multiple-choice; 2 for short answer; 4 for extended constructed-response; and 5 for writing prompt) for alignment analysis. The results are reported for English Language Arts Grades 2, 4, 6, 7, and 9. To examine the consistency of assigning the DOK level to test items from reviewers, the intraclass correlation (Shrout and Fleiss, 1979) was applied**.** The correlations are ranged from .70 to .88 in this study. Corresponding content standards can be found in Appendix C and detailed information from analysis is included in Appendix D.

Delaware English Language Arts Standards

In Delaware, there are four standards in English Language Arts (Appendix C); however, Standard 3, which addresses expectations for listening and oral skills, was not included in this alignment analysis. Performance Indicators represent sub-categories of expectations for students and Grade Level Expectations represent even more specific ELA expectations for students at specific grade levels. For all grades analyzed for reading (Grade 2, 4, 6, 7, and 9), judgments regarding Depth-of-Knowledge Levels were made at the level of Performance Indicators; however, Grade Level Expectations corresponding to the Performance Indicators were also considered in order to assign DOK levels to the Performance Indicators. (See Appendix C for a more detailed explanation of the Grade Level Expectations.)

Table 2 is a summary of reviewers' coding on the *Depth of Knowledge* (DOK) levels for Performance Indicators for English Language Arts Content Standards 1, 2, and 4 for Grades 2, 4, 6, 7, and 9. (Content Standard 1 was not analyzed for Grade 2.) Results in Table 2 suggest few performance indicators at DOK Level 1 (recall or reproduction) in Grade 2 and none at Grades 2, 4, 6, 7, and 9. Percentages of performance indicators at DOK Level 2 (Use of skills and concepts) relatively high at Grade 2, decreased sharply

from Grade 2 (61%) to Grade 4 (35%), and then continued to decrease at higher grade levels (26% for Grade 6, 21% for Grade 7, and 18% for Grade 9). Although it can be very reasonable for the proportion of objectives by DOK level to remain the same across grades, an increasing proportion of objectives at higher DOK levels in the later grades do indicate a shift towards more reasoning and less emphasis on skills and concepts for these grades.

Table 2. Summary Coding of Depth of Knowledge for Objectives

Grade	Number of Objectives	DOK Level	Objectives by DOK Level	
			Number	*%*
2	21	*1*	1	4
		2	13	61
		3	7	33
		4	0	0
4	27	*1*	0	0
		2	14	35
		3	25	62
		4	1	2
6	28	*1*	0	0
		2	11	26
		3	26	63
		4	4	9
7	28	*1*	0	0
		2	9	21
		3	22	53
		4	10	24
9	30	*1*	0	0
		2	8	18
		3	19	44
		4	16	37

Percentages of DOK Level 3 increased sharply from Grade 2 (33%) to Grade 3 (62%), increased slightly for Grade 6 (63%), and then decreased for Grade 7 (53%) and Grade 9 (44%). The decrease in percentages of performance indicators at DOK Level 3 for higher grade levels may indicate performance indicators reach a maximum of complexity in Grade 6 and then taper off in terms of complexity for students in Grade 7 and 9. Another

explanation might be the Grade Level Expectations that were used to determine DOK levels for performance indicators at Grades 7 and 9 resulted in lower DOK assignments. Percentages of performance indicators at DOK Level 4 increased across grade levels (0% for Grade 2, 2% for Grade 4, 9% for Grade 6, 24% for Grade 7, and 37% for Grade 9. These data suggest ELA performance indicators increasingly require students to engage in extended thinking at higher grade levels.

The alignment analysis in English Language Arts raised some issues. Although DOK levels were determined at the level of ELA Reading Performance Indicators, reviewers also considered a sub-category of Grade Level Expectations for reading in order to make DOK level assignments. Reviewers suggested it was sometimes difficult to assign appropriate DOK levels to performance indicators because the corresponding Grade-Level Expectations represented such a wide range of difficulty levels. A second issue related to the alignment analysis process for the writing standard. Although there is only one performance indicator for Standard 1 (Use written and oral English), 14 objectives were coded for DOK levels in writing. Reviewers also considered specific Grade-Level Expectations for writing in order to determine a DOK level for each of the 14 objectives. While this process allowed reviewers to code more items for writing, the redundancy of the writing objectives may have led to some ambiguous results. Furthermore, even though the Delaware Student Testing Program is designed to measure student reading comprehension and writing skills separately, the current analysis was conducted on the basis of standards rather than how the test was constructed.

Table 3 presents the summary results of the alignment analysis in English Language Arts by grade using the four criteria. 'Yes' indicates meeting the acceptable level to a corresponding alignment criterion, 'Weak' indicates a weakly meet, and 'No' indicates unacceptable.

- Grade 2 (Appendix D Tables 1.1 – 1.4): Standard 1 (Use written and oral English) was not assessed for the four criteria. For Standard 2 (Construct, examine, and extend the meaning of literary, informative, and technical texts), reviewers' judgments indicate an acceptable level for Categorical Concurrence, Depth of Knowledge Consistency, and Range of Knowledge Consistency. However, the test showed only a weak distribution of test items/activities across performance indicators (Balance Index) for Standard 2 (.65). For Standard 4 (Use literary knowledge), reviewers rated only Depth of Knowledge Consistency between test items and performance indicators at an acceptable level. Categorical Concurrence (5), Range of Knowledge Consistency (14%), and Balance of Representation (.54) were rated unacceptable for this standard. These data suggest improvement in alignment is needed, particularly for Standard 4.

- Grade 4 (Appendix D Tables 2.1 - 2.4): For Standard 1, alignment results indicated an acceptable number of test items; however, there was an unacceptable level (31%) of DOK matches for this standard. In addition, Range of Knowledge

Consistency was considered weak (42%). The distribution of test-items/activities across performance indicators was judged by reviewers as unacceptable (.56). For Standard 2, Categorical Concurrence, Depth of Knowledge Consistency, and Range of Knowledge Consistency were acceptable. The distribution of test items/activities across performance indicators was considered to be unacceptable (.57) for Standard 2. For Standard 4 an acceptable number of test items and a consistent match of appropriate DOK levels for test items were found. Reviewers found Range of Knowledge Consistency between test items and standards to be weak (42%); they also judged Balance of Representation for this standard as weak (.64). These data suggest improvement in alignment is needed for Standards 1, 2, and 4.

Table 3. Summary of Alignment Results for English Language Arts

Grade	Standard	Categorical Concurrence	Depth of Knowledge	Range of Knowledge	Balance of Representation
Grade 2	1		*Not assessed*		
	2	Yes	Yes	Yes	*Weak*
	4	No	Yes	No	No
Grade 4	1	Yes	No	*Weak*	No
	2	Yes	Yes	Yes	No
	4	Yes	Yes	*Weak*	*Weak*
Grade 6	1	Yes	Yes	Yes	*Weak*
	2	Yes	Yes	Yes	*Weak*
	4	Yes	Yes	Yes	Yes
Grade 7	1	Yes	Yes	Yes	Yes
	2	Yes	Yes	Yes	*Weak*
	4	Yes	Yes	*Weak*	Yes
Grade 9	1	Yes	Yes	Yes	Yes
	2	Yes	*Weak*	Yes	*Weak*
	4	Yes	Yes	Yes	Yes

- Grade 6: the results (Appendix D Tables 3.1 - 3.4): For Standard 1, reviewers judged Categorical Concurrence, Depth of Knowledge Consistency, and Range-of-Knowledge Consistency to be at acceptable levels. Balance of Representation was judged as weak

(.69). Similar judgments were made regarding Standard 2 – Categorical Concurrence, Depth of Knowledge Consistency, and Range-of-knowledge Consistency were acceptable, and Balance of Representation was weak (.60). For Standard 4, all four criteria were judged by the reviewers as acceptable. These data suggest alignment is acceptable overall; some improvement in Balance of Representation could be made for Standard 2 and Standard 4.

- Grade 7 (Appendix D Tables 4.1 – 4.4): For Standard 1, results of the analysis suggest acceptable levels for Categorical Concurrence, Depth of Knowledge Consistency, Range of Knowledge Consistency, and Balance of Representation. For Standard 2, the first three criteria were considered acceptable; however, reviewers found Balance of Representation to be weak (.64).) For Standard 4, Categorical Concurrence, Depth of Knowledge Consistency, and Balance of Representation were considered to be acceptable; however, for this standard, reviewers found Range of Knowledge Consistency to be weak (48%). These data suggest alignment is acceptable overall; some improvement could be made in Range of Knowledge Consistency for Standard 4 and Balance of Representation for Standard 2.

- Grade 9 (Appendix D Tables 5.1 – 5.4): For Standard 1, reviewers considered all four criteria to be met at acceptable levels. For Standard 2, Categorical Concurrence and Range of Knowledge Consistency were judged as acceptable; however, Depth of Knowledge was considered weak (41%), as was Balance of Representation (.65). For Standard 4, reviewers considered all four criteria to be met at acceptable levels. These data suggest alignment is acceptable overall; some improvement could be made in Depth of Knowledge Consistency for Standard 2 and Balance of Representation for Standard 2.

Recommendations

The following recommendations for improving the degree of alignment in English Language Arts are based upon the results of the current analysis using Webb's model and evaluation criteria (Table 2 and Table 3).

1. For Grade 2, it is recommended that alignment procedures be conducted for Standard 1 (Use oral and written English).

2. For Grade 2, improvement in alignment between the DSTP and Standard 4 (Use literary knowledge) is recommended. Increasing the number of test items that correspond to performance indicators for this standard would improve Categorical Concurrence with this standard. If these test items are designed to assess the appropriate DOK levels and cover a broad range of performance indicators, then they will also help to improve Range of Knowledge Consistency and Balance of Representation between the DSTP and Standard 4.

3. For Grade 4, improvement in alignment between the DSTP and Standards 1, 2, and 4 is recommended, particularly for Range of Knowledge Consistency and Balance of Representation criteria. Adding test items that assess appropriate DOK levels and cover a broad range of performance indicators will improve alignment for these criteria. Adding writing prompts in Grade 4 which require higher DOK levels will improve Depth of Knowledge Consistency between the test and Standard 1 (Use oral and written language).

4. For Grades 6, 7, and 9, the results indicate a sufficient or approaching sufficient alignment between the DSTP and the English Language Arts Performance Indicators for Content Standards 1, 2, and 4. Four of the six (66%) weaknesses rated by reviewers at these grade levels occurred for lack of Balance of Representation (2 weaknesses identified for Standard 1; 1 weakness identified for Standard 2; and 1 weakness identified for Standard 4). Adding more test items that address a broader range of performance indicators will support an even distribution of test items/activities across the ELA Content Standards.

5. For Grades 7 and 9, the results indicate that some improvement is needed related to consistency in Depth of Knowledge levels between the test and ELA performance indicators. Adding test items that address higher DOK levels will improve consistency between the DSTP and performance indicators for Standard 1 (Use written and oral English) and Standard 2 (Construct, examine, and extend the meaning of literary, informative, and technical texts).

6. For Grades 7 and 9, the results indicate that improvement is needed in developing performance indicators that are more representative of higher DOK levels. Fewer performance indicators at DOK Level 1 should be represented at higher grade levels.

7. Some consideration might be given to improving alignment between performance indicators for Standard 1, 2, and 4 and the corresponding grade level expectations. Another possible alternative to the coding process used for this study might involve coding DOK levels for specific grade level expectations.

8. An alternate alignment process is recommended for the writing standard. An alternative to the process used for this study would include assigning DOK levels to specific grade level expectations for Standard 1, rather than the one performance indicator or 14 objectives. Also, consideration might be given to using student work (anchor papers) more extensively for alignment purposes than was possible during this study. The writing rubrics might be revised to include more specific criteria that would be more useful in determining DOK levels.

References

Carroll, J. B. (1963). A model for school learning. *Teachers College Record, 64*, 723-733

Cohen, S. A. (1987). Instructional alignment: Searching for a magic bullet. *Educational Researcher, 16*(8), 16-20.

Consortium for Policy Research in Education. (1991). *Putting the pieces together: Systemic school reform* (CPRE Policy Briefs). New Brunswick, NJ: Rutgers, The State *University of New Jersey, Eagleton Institute of Politics.*

Council of Chief State School Officers. (2002). *No Child Left Behind Act. A description*
Newmann, F. M. (1993). Beyond common sense in educational restructuring: The issues of content and linkage. *Educational Researcher*, *22*(2), 4-13, 22.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Subkoviak, M. J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement, 25*(1), 47-55.

U.S. Congress, House of Representatives. (1994, September 28). *Improving America's Schools Act*. Conference Report to accompany H. R. 6 Report 103-761. Washington, DC: U.S. Government Printing Office.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (NISE Research Monograph No. 6). Madison: University of Wisconsin–Madison, National Institute for Science Education. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessment in four states* (NISE Research Monograph No.18). Madison: University of Wisconsin–Madison, National Institute for Science Education. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states – A study of the State Collaborative on Assessment & Student Standards (SCASS) Technical Issues in Large-Scale Assessment (TILSA).* A Publication of Chief State School Officer.

Webb, N. L. and Delaware Department of Education (2004). *A special summary report – Findings of an alignment study in DSTP English Language Arts and mathematics for Grades 3, 5, 8, and 10; and Science for Grades 4, 6, 8, and 11.* A Publication of State of Delaware.

Appendix A

List of Reviewers

List of Reviewers

Name	District	School
Barbara Coe		Retired teacher
Shay Eli	Cape Henlopen	Milton Elementary
Linda Hand	Christina	Maclary Elementary
Debbie O'Brien	Lake Forest	Central Elementary
Ann Lewis	Laurel	Dunbar Elementary
Sharon Biss	Red Clay	Stanton Middle School
Sharon Bryant-Horsey	New Castle County Vo-Tech	District Office
Debbie Buffington	Indian River	District Office
Pat Clements	New Castle County Vo-Tech	District Office
Kathy Gilbert	Brandywine	Springer Middle School
Becky Sharp	Polytech	Polytech High School
Kathy Casey	Brandywine	Maple Lane Elementary
Rhonda Nowak*	Loyola University New Orleans	Consultant
Cynthia Jacobson*	Retired from Holmen School District in Wisconsin	Consultant
Brian Vesperman	University of Wisconsin - Madison	Consultant

* Group Leader

Appendix B

Levels of Depth of Knowledge

Subject	Depth of Knowledge			
	Level 1	**Level 2**	**Level 3**	**Level 4**
English Language Arts	Requires students to recall, observe, question or represent facts or simple skills or abilities. Requires only surface understanding of text often verbatim recall. Examples: ◆ Support ideas by reference to details in text ◆ Use dictionary to find meaning ◆ Identify figurative language in passage ◆ Identify correct spelling or meaning of words	Requires processing beyond recall and observation. Requires both comprehension and subsequent processing of text. Involves ordering, classifying text as well as identifying patterns, relationships and main points. Examples: ◆ Use context to identify unfamiliar words ◆ Predict logical outcome ◆ Identify and summarize main points ◆ Apply knowledge of conventions of standard American English ◆ Compose accurate summaries	Requires students to go beyond text. Requires students to explain, generalize and connect ideas. Involves inference, prediction, elaboration and summary. Requires students to support positions using prior knowledge and to manipulate themes across passages. Examples: ◆ Determine effect of author's purpose on text elements ◆ Summarize information from multiple sources ◆ Critically analyze literature ◆ Compose focused, organized, coherent, purposeful prose	Requires extended higher order processing. Typically requires extended time to complete task, but time spent not on repetitive tasks. Involves taking information from one text/passage and applying this information to a new task. May require generating hypotheses and performing complex analyses and connections among texts. Examples: ◆ Analyze and synthesize information from multiple sources ◆ Examine and explain alternative perspectives across sources ◆ Describe and illustrate common themes across a variety of texts ◆ Create compositions that synthesize, analyze, and evaluate

Depth-of-Knowledge for English Language Arts – Reading

Level 1 (Recall) requires students to receive or recite facts or to use simple skills or abilities. Oral reading that does not include analysis of the text, as well as basic comprehension of a text, is included. Items require only a shallow understanding of the text presented and often consist of verbatim recall from text, or simple understanding of a single word or phrase. Some examples that represent, but do not constitute all of, Level 1 performance are:

- Support ideas by reference to details in the text.
- Use a dictionary to find the meanings of words.
- Identify figurative language in a reading passage.

Level 2 (Skills and Concepts) includes the engagement of some mental processing beyond recalling or reproducing a response; it requires both comprehension and subsequent processing of text or portions of text. Intersentence analysis of inference is required. Some important concepts are covered, but not in a complex way. Standards and items at this level may include words such as summarize, interpret, infer, classify, organize, collect, display, compare, and determine whether fact or opinion. Literal main ideas are stressed. A Level 2 assessment item may require students to apply skills and concepts that are covered in Level 1. Some examples that represent, but do not constitute all of, Level 2 performance are:

- Use context cues to identify the meaning of unfamiliar words.
- Predict a logical outcome based on information in a reading selection.
- Identify and summarize the major events in a narrative.

Level 3 (Strategic Thinking) requires deep knowledge that becomes a greater focus at Level 3. Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be encouraged to explain, generalize, or connect ideas. Standards and items at Level 3 involve reasoning and planning. Students must be able to support their thinking. Items may involve abstract theme identification, inference across an entire passage, or students' application of prior knowledge. Items may also involve more superficial connections between texts. Some examples that represent, but do not constitute all of, Level 3 performance are:

- Determine the author's purpose and describe how it affects the interpretation of a reading selection.
- Summarize information from multiple sources to address a specific topic.
- Analyze and describe the characteristics of various types of literature.

Level 4 (Extended Thinking) requires higher-order thinking and deep knowledge. The standard or a test item at this level will probably be an extended activity, with extended time provided for completing it. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require the application of significant conceptual understanding and higher-order thinking. Students take information from at least one passage of a text and are asked to apply this information to a new task. They may also be asked to develop hypotheses and perform complex analyses of the connections among texts. Some examples that represent, but do not constitute all of, Level 4 performance are:

- Analyze and synthesize information from multiple sources.
- Examine and explain alternative perspectives across a variety of sources.
- Describe and illustrate how common themes are found across texts from different cultures.

Depth-of-Knowledge for English Language Arts – Writing

Level 1 (Recall and Reproduction) requires the student to write or recite simple facts. The focus of this writing or recitation is not on complex synthesis or analysis but on basic ideas. The students are asked to list ideas or words, as in a brainstorming activity prior to written composition; are engaged in a simple spelling or vocabulary assessment; or are asked to write simple sentences. Students are expected to write and speak using the conventions of Standard English. This includes using appropriate grammar, punctuation, capitalization, and spelling. Some examples that represent, but do not constitute all of, Level 1 performance are:

- Use punctuation marks correctly.
- Identify Standard English grammatical structures and refer to resources for correction.

Level 2 (Skills and Concepts) requires some mental processing. At this level, students are engaged in first-draft writing, or brief extemporaneous speaking for a limited number of purposes and audiences. Students are expected to begin connecting ideas, using a simple organizational structure. For example, students may be engaged in note taking, outlining, or simple summaries. Text may be limited to one paragraph. Students demonstrate a basic understanding and appropriate use of such reference materials as a dictionary, thesaurus, or web site. Some examples that represent, but do not constitute all of, Level 2 performance are:

- Construct compound sentences.
- Use simple organizational strategies to structure written work.
- Write summaries that contain the main idea of the reading selection and pertinent details.

Level 3 (Strategic Thinking) requires some higher-level mental processing. Students are engaged in developing compositions that include multiple paragraphs. These compositions may include complex sentence structure and may demonstrate some synthesis and analysis. Students show awareness of their audience and purpose through focus, organization, and the use of appropriate compositional elements. The use of appropriate compositional elements may include addressing chronological order in a narrative, or including supporting facts and details in an informational report. At this stage, students are engaged in editing and revising to improve the quality of the composition. Some examples that represent, but do not constitute all of, Level 3 performance are:

- Support ideas with details and examples.
- Use voice appropriate to purpose and audience.
- Edit writing to produce a logical progression of ideas.

Level 4 (Extended Thinking) requires higher-level thinking. The standard at this level is a multi-paragraph composition that demonstrates the ability to synthesize and analyze complex ideas or themes. There is evidence of a deep awareness of purpose and audience. For example, informational papers include hypotheses and supporting evidence. Students are expected to create compositions that demonstrate a distinct voice and that stimulate the reader

or listener to consider new perspectives on the ideas and themes addressed an example that represents, but does not constitute all of, Level 4 performance is:

- Write an analysis of two selections, identifying the common theme and generating a purpose that is appropriate for both.

Appendix C

Content Standards for English Language Arts

Standard 1. Student will use written and oral English appropriate for various purposes and audiences.

Standard 2. Student will construct, examine, and extend the meaning of literary, informative, and technical texts through listening, reading, and viewing.

Standard 3. Student will access, organize, and evaluate information gained by listening, reading, and viewing.

Standard 4. Student will use literary knowledge accessed through print and visual media to connect self to society and culture.

Appendix D

Summary of Results of Alignment Analysis

Table 1.1. Summary of Categorical Concurrence for Grade 2

Standard	Coded Obj. Mean	Level of Performance Indicators Level	N. of Obj.	% of Obj	Hits Mean	SD	Categorical Concurrence
		1	1	6			
2	16.29	*2*	12	75	37.43	4.62	Yes
		3	3	18			
		4	0	0			
		1	0	0			
4	37.14	*2*	1	20	1.86	2.37	No
		3	4	80			
		4	0	0			
		1	1	4			
Total	21.29	*2*	13	61	39.29	4.13	
		3	7	33			
		4	0	0			

Table 1.2. Summary of Depth of Knowledge Consistency for Grade 2

Standard	Coded Obj. Mean	Hits Mean	SD	% of Items at DOK Level Under Mean	SD	At Mean	SD	Above Mean	SD	DOK Consistency
2	16.29	37.43	4.62	30	35	62	38	9	25	Yes
4	5.00	1.86	2.37	20	40	73	39	7	13	Yes
Total	21.29	39.29	4.13	29	36	63	38	9	24	

Table 1.3. Summary of Range of Knowledge for Grade 2

Standard	Coded Obj.	Hits		N. of Hits		% of Hits		Range of
	Mean	Mean	SD	Mean	SD	Mean	SD	Knowledge
2	16.29	37.43	4.62	9.00	0.53	55	4	Yes
4	5.00	1.86	2.37	0.71	0.69	14	14	No
Total	21.29	39.29	4.13	4.86	3.76	35	16	

Table 1.4. Summary of Balance of Representation for Grade 2

Standard	Coded Obj.	% of Hits		Index		Balance of
	Mean	Mean	SD	Mean	SD	Representation
2	16.29	95	6	0.65	0.07	Weak
4	5.00	5	6	0.54	0.42	No
Total	21.29	50	42	0.59	0.16	

Table 2.1. Summary of Categorical Concurrence for Grade 4

Standard	Coded Obj. Mean	Level of Performance Indicators Level	N. of Obj.	% of Obj.	Hits Mean	SD	Categorical Concurrence
		1	0	0			
1	14.00	*2*	2	14	9.17	5.36	Yes
		3	11	78			
		4	1	7			
		1	0	0			
2	18.33	*2*	10	55	68.33	8.73	Yes
		3	8	44			
		4	0	0			
		1	0	0			
4	8.00	*2*	2	25	14.50	4.15	Yes
		3	6	75			
		4	0	0			
		1	0	0			
Total	40.33	*2*	14	35	92.00	13.76	
		3	25	62			
		4	1	2			

Table 2.2. Summary of Depth of Knowledge Consistency for Grade 4

Standard	Coded Obj. Mean	Hits Mean	SD	% of Items at DOK Level Under Mean	SD	At Mean	SD	Above Mean	SD	DOK Consistency
1	14.00	9.17	5.36	65	44	31	40	4	14	No
2	18.33	68.33	8.73	32	41	61	40	7	19	Yes
4	8.00	14.50	4.15	45	45	52	43	4	9	Yes
Total	40.33	92.0	13.76	43	45	51	43	5	17	

Table 2.3. Summary of Range of Knowledge for Grade 4

Standard	Coded Obj. Mean	Hits Mean	Hits SD	N. of Hits Mean	N. of Hits SD	% of Hits Mean	% of Hits SD	Range of Knowledge
1	14.0	9.17	5.36	5.83	2.95	42	21	Weak
2	18.33	68.33	8.73	11.83	1.67	64	8	Yes
4	8.00	14.50	4.15	3.33	0.94	42	12	Weak
Total	40.33	92	13.76	7.0	3.90	49	14	

Table 2.4. Summary of Balance of Representation for Grade 4

Standard	Coded Obj. Mean	% of Hits Mean	% of Hits SD	Index Mean	Index SD	Balance of Representation
1	14.0	10	5	0.56	0.28	No
2	18.33	75	8	0.57	0.05	No
4	8.00	16	3	0.64	0.08	Weak
Total	40.33	33	29	0.59	0.13	

Table 3.1. Summary of Categorical Concurrence for Grade 6

Standard	Coded Obj. Mean	Level of Performance Indicators Level	N. of Obj.	% of Obj.	Hits Mean	SD	Categorical Concurrence
		1	0	0			
1	11.67	*2*	1	7	11.20	3.51	Yes
		3	10	71			
		4	3	21			
		1	0	0			
2	17.00	*2*	9	45	69.20	8.77	Yes
		3	11	55			
		4	0	0			
		1	0	0			
4	5.83	*2*	1	14	8.40	3.44	Yes
		3	5	71			
		4	1	14			
		1	0	0			
Total	34.50	*2*	11	26	74.00	17.89	
		3	26	63			
		4	4	9			

Table 3.2. Summary of Depth of Knowledge Consistency for Grade 6

Standard	Coded Obj. Mean	Hits Mean	SD	% of Items at DOK Level: Under Mean	SD	At Mean	SD	Above Mean	SD	DOK Consistency
1	11.67	11.20	3.51	34	39	60	39	6	20	Yes
2	17.00	69.20	8.77	26	35	65	38	9	23	Yes
4	5.83	8.4	3.44	32	44	59	47	10	29	Yes
Total	34.50	74.0	17.89	29	38	63	40	8	24	

Table 3.3. Summary of Range of Knowledge for Grade 6

Standard	Coded P.I. Mean	Hits		N. of Hits		% of Hits		Range of Knowledge
		Mean	SD	Mean	SD	Mean	SD	
1	11.67	11.20	3.51	7.0	2.18	50	16	Yes
2	17.0	69.20	8.77	14.20	0.75	70	4	Yes
4	5.83	8.4	3.44	4.20	1.60	60	23	Yes
Total	34.50	74	17.89	8.47	4.42	60	15	

Table 3.4. Summary of Balance of Representation for Grade 6

Standard	Coded P.I. Mean	% of Hits		Index		Balance of Representation
		Mean	SD	Mean	SD	
1	11.67	12	4	0.69	0.18	Weak
2	17.0	78	7	0.60	0.05	Weak
4	5.83	10	4	0.73	0.06	Yes
Total	34.50	33	32	0.67	0.12	

Table 4.1. Summary of Categorical Concurrence for Grade 7

Standard	Coded Obj. Mean	Level of Performance Indicators Level	N. of Obj.	% of Obj.	Hits Mean	SD	Categorical Concurrence
		1	14	100			
1	10.88	*2*	0	0	16.50	1.12	Yes
		3	0	0			
		4	0	0			
		1	0	0			
2	15.00	*2*	5	25	67.50	9.03	Yes
		3	11	55			
		4	4	20			
		1	0	0			
4	5.25	*2*	1	14	12.00	3.87	Yes
		3	4	57			
		4	2	28			
		1	14	35			
Total	31.12	*2*	6	14	72.00	26.04	
		3	15	37			
		4	6	14			

Table 4.2. Summary of Depth of Knowledge Consistency for Grade 7

Standard	Coded Obj. Mean	Hits Mean	SD	% of Items at DOK Level: Under Mean	SD	At Mean	SD	Above Mean	SD	DOK Consistency
1	10.88	16.50	1.12	0	0	3	17	97	17	Yes
2	15.00	67.50	9.03	32	38	56	38	12	25	Yes
4	5.25	12.00	3.87	47	41	47	41	6	14	Yes
Total	31.12	72.0	26.04	22	35	35	41	43	47	

Table 4.3. Summary of Range of Knowledge for Grade 7

Standard	Coded Obj. Mean	Hits Mean	Hits SD	N. of Hits Mean	N. of Hits SD	% of Hits Mean	% of Hits SD	Range of Knowledge
1	10.88	16.50	1.12	11.0	1.15	76	7	Yes
2	15.0	67.50	9.03	14.83	1.07	74	5	Yes
4	5.25	12.0	3.87	3.33	1.37	48	20	Weak
Total	31.12	72	26.04	9.72	4.93	66	18	

Table 4.4. Summary of Balance of Representation for Grade 7

Standard	Coded Obj. Mean	% of Hits Mean	% of Hits SD	Index Mean	Index SD	Balance of Representation
1	10.88	17	2	0.84	0.03	Yes
2	15.0	70	3	0.64	0.05	Weak
4	5.25	12	3	0.76	0.08	Yes
Total	31.12	33	26	0.75	0.10	

Table 5.1. Summary of Categorical Concurrence for Grade 9

Standard	Coded Obj. Mean	Level of Performance Indicators Level	N. of Obj.	% of Obj.	Hits Mean	SD	Categorical Concurrence
		1	14	100			
1	14.0	*2*	0	0	15.50	0.76	Yes
		3	0	0			
		4	0	0			
		1	0	0			
2	21.0	*2*	4	19	97.17	17.02	Yes
		3	9	42			
		4	8	38			
		1	0	0			
4	8.0	*2*	1	12	17.33	11.66	Yes
		3	4	50			
		4	3	37			
		1	0	0			
Total	43.0	*2*	8	18	130.0	26.46	
		3	19	44			
		4	16	37			

Table 5.2. Summary of Depth of Knowledge Consistency for Grade 9

Standard	Coded Obj. Mean	Hits Mean	SD	% of Items at DOK Level: Under Mean	SD	At Mean	SD	Above Mean	SD	DOK Consistency
1	14.0	15.50	0.76	0	0	0	0	100	0	Yes
2	21.0	97.17	17.02	50	41	41	35	9	19	Weak
4	8.0	17.33	11.66	29	38	53	37	18	31	Yes
Total	43.0	130.0	26.46	30	40	29	36	41	46	

Table 5.3. Summary of Range of Knowledge for Grade 9

Standard	Coded Obj. Mean	Hits Mean	Hits SD	N. of Hits Mean	N. of Hits SD	% of Hits Mean	% of Hits SD	Range of Knowledge
1	14.0	15.50	0.76	11.0	0	79	0	Yes
2	21.0	97.17	17.02	16.50	2.75	79	13	Yes
4	8.0	17.33	11.66	4.83	2.19	60	27	Yes
Total	43.0	130.0	26.46	10.78	5.18	73	20	

Table 5.4. Summary of Balance of Representation for Grade 9

Standard	Coded Obj. Mean	% of Hits Mean	% of Hits SD	Index Mean	Index SD	Balance of Representation
1	14.0	12	2	0.83	0.01	Yes
2	21.0	75	5	0.65	0.04	Weak
4	8.0	12	7	0.74	0.14	Yes
Total	43.0	33	30	0.74	0.11	