

# ALIGNING ASSESSMENT TO GUIDE THE LEARNING OF ALL STUDENTS



## SIX REPORTS

### REPORT 1

THE WEB ALIGNMENT TOOL:  
DEVELOPMENT, REFINEMENT, AND DISSEMINATION  
NORMAN L. WEBB, MEREDITH ALT, ROB ELY, MARIA CORMIER, & BRIAN VESPERMAN

### REPORT 2

ALIGNMENT OF ALTERNATE ASSESSMENTS USING THE WEBB SYSTEM  
GERALD TINDAL

### REPORT 3

ASSESSING VERTICAL ALIGNMENT  
LAURESS WISE AND MEREDITH ALT

### REPORT 4

VERTICAL ALIGNMENT OF GRADE-LEVEL EXPECTATIONS FOR STUDENT ACHIEVEMENT:  
REPORT OF A PILOT STUDY  
LAURESS L. WISE, LIRU ZHANG, PHOEBE WINTER, LESLIE TAYLOR, AND D. E. (SUNNY) BECKER

### REPORT 5

ALIGNING ENGLISH LANGUAGE PROFICIENCY TESTS TO  
ENGLISH LANGUAGE LEARNING STANDARDS  
H. GARY COOK

### REPORT 6

DEVELOPING ALIGNED PERFORMANCE LEVEL DESCRIPTORS FOR THE  
ENGLISH LANGUAGE DEVELOPMENT ASSESSMENT K-2 INVENTORIES  
MARÍA H. MALAGÓN, MARJORIE B. ROSENBERG AND PHOEBE C. WINTER





## THE COUNCIL OF CHIEF STATE SCHOOL OFFICERS

**The Council of Chief State School Officers (CCSSO)** is a nonpartisan, nationwide, nonprofit organization of public officials who head departments of elementary and secondary education in the states, the District of Columbia, the Department of Defense Education Activity, and five U.S. extra-state jurisdictions. CCSSO provides leadership, advocacy, and technical assistance on major educational issues. The Council seeks member consensus on major educational issues and expresses their views to civic and professional organizations, federal agencies, Congress, and the public.

### Division of State Services and Technical Assistance

**The Council's Division of State Services and Technical Assistance** supports state education agencies in developing standards-based systems that enable all children to succeed. Initiatives of the division support improved methods for collecting, analyzing, and using information for decision making; development of assessment resources; creation of high-quality professional preparation and development programs; emphasis on instruction suited for diverse learners; and the removal of barriers to academic success.

### State Collaborative on Assessment and Student Standards

The State Collaborative on Assessment and Student Standards (SCASS) Project was created in 1991 to encourage and assist states in working collaboratively on assessment design and development for a variety of topics and subject areas. The Division of State Services and Technical Assistance of the Council of Chief State School Officers is the organizer, facilitator, and administrator of the projects. SCASS projects accomplish a wide variety of tasks identified by each of the groups including examining the needs and issues surrounding the area(s) of focus, determining the products and goals of the project, developing assessment materials and professional development materials on assessment, summarizing current research, analyzing best practice, examining technical issues, and/or providing guidance on federal legislation.

## COUNCIL OF CHIEF STATE SCHOOL OFFICERS

Valerie A. Woodruff (Delaware), President  
Elizabeth Burmaster (President-Elect), Wisconsin  
David P. Driscoll (Past President), Massachusetts  
G. Thomas Houlihan, Executive Director

Julia L. Lara, Deputy Executive Director  
Division of State Services and Technical Assistance

Don Long, Director, and Arthur Halbrook, Senior Associate  
State Collaborative on Assessment and Student Standards

ISBN: 1-933757-00-0

© 2006 by the Council of Chief State School Officers, Washington, DC  
*All rights reserved.*

Council of Chief State School Officers  
One Massachusetts Avenue, NW, Suite 700  
Washington, DC 20001-1431  
Phone (202) 336-7000  
Fax (202) 408-8072  
[www.ccsso.org](http://www.ccsso.org)

# ALIGNING ASSESSMENT TO GUIDE THE LEARNING OF ALL STUDENTS



## SIX REPORTS

### REPORT 1

THE WEB ALIGNMENT TOOL:  
DEVELOPMENT, REFINEMENT, AND DISSEMINATION  
NORMAN L. WEBB, MEREDITH ALT, ROB ELY, MARIA CORMIER, & BRIAN VESPERMAN

### REPORT 2

ALIGNMENT OF ALTERNATE ASSESSMENTS USING THE WEBB SYSTEM  
GERALD TINDAL

### REPORT 3

ASSESSING VERTICAL ALIGNMENT  
LAURESS WISE AND MEREDITH ALT

### REPORT 4

VERTICAL ALIGNMENT OF GRADE-LEVEL EXPECTATIONS FOR STUDENT ACHIEVEMENT:  
REPORT OF A PILOT STUDY  
LAURESS L. WISE, LIRU ZHANG, PHOEBE WINTER, LESLIE TAYLOR, AND D. E. (SUNNY) BECKER

### REPORT 5

ALIGNING ENGLISH LANGUAGE PROFICIENCY TESTS TO  
ENGLISH LANGUAGE LEARNING STANDARDS  
H. GARY COOK

### REPORT 6

DEVELOPING ALIGNED PERFORMANCE LEVEL DESCRIPTORS FOR THE  
ENGLISH LANGUAGE DEVELOPMENT ASSESSMENT K-2 INVENTORIES  
MARIA H. MALAGON, MARJORIE B. ROSENBERG AND PHOEBE C. WINTER



C O U N C I L   O F   C H I E F   S T A T E   S C H O O L   O F F I C E R S



## Acknowledgements

---

The Council would like to acknowledge the leadership of the State of Oklahoma, and especially Ronda Townsend, Executive Director, State Testing, Office of Accountability and Assessments, in guiding this research project to fruition. We also would like to acknowledge the contributions and commitment of the state members of the collaborative:

Alabama	Louisiana	North Carolina	Texas
California	Massachusetts	Oklahoma (lead state)	Wyoming
Delaware	Minnesota	Pennsylvania	West Virginia
Kansas	New Jersey	South Carolina	Wisconsin

In addition, the following members of the Technical Issues in Large-Scale Assessment (TILSA) SCASS project who were not part of the State Collaborative dedicated their time and insights to this project: Arizona, Georgia, Kentucky, Michigan, New Mexico, Ohio, and Rhode Island.

The Council also acknowledges the guidance and support from the U.S. Department of Education (ED), especially Sue Rigney, who served as Senior Project Officer. This research was supported by a grant from ED, No. S368A030011, to the Oklahoma Department of Education, with a subcontract to the Council of Chief State School Officers and the Wisconsin Center for Education Research, University of Wisconsin, for the period from April 14, 2003, through September 30, 2005, and a grant from the National Science Foundation, No. EHR 0233445, to the Wisconsin Center for Education Research, University of Wisconsin, for the period from October 1, 2002, to September 30, 2005.

Council staff member, Arthur Halbrook, Senior Associate and Coordinator for TILSA, was invaluable in his guidance of the project through TILSA and his editing work. Frank Philip, Director of Program Development and Operations for the Council, generously devoted his time and mastery in the final design and publication stages of this work. In addition, we wish to extend special acknowledgement to John Olson, former Director of Assessment at the Council, for his instrumental work in the initiation and planning of this project. Finally, we greatly appreciate the work of the authors and other researchers, who are also acknowledged in the individual reports, for this research and their commitment to improving the use of assessment in standards-based reform and student learning.



## Table of Contents

---

<b>Overview</b>	<b>vii</b>
<b>Report 1      The Web Alignment Tool: Development, Refinement, and Dissemination</b>	<b>1</b>
Norman L. Webb, Meredith Alt, Rob Ely, Maria Cormier, & Brian Vesperman	
<b>Report 2      Alignment of Alternate Assessments Using the Webb System</b>	<b>31</b>
Gerald Tindal	
<b>Report 3      Assessing Vertical Alignment</b>	<b>57</b>
Lauress Wise and Meredith Alt	
<b>Report 4      Vertical Alignment of Grade-Level Expectations for Student Achievement: Report of a Pilot Study</b>	<b>75</b>
Lauress L. Wise, Liru Zhang, Phoebe Winter, Leslie Taylor, and D. E. (Sunny) Becker	
<b>Report 5      Aligning English Language Proficiency Tests to English Language Learning Standards</b>	<b>131</b>
H. Gary Cook	
<b>Report 6      Developing Aligned Performance Level Descriptors for the English Language Development Assessment K-2 Inventories</b>	<b>155</b>
María H. Malagón, Marjorie B. Rosenberg and Phoebe C. Winter	





Alignment is now clearly recognized as a central driver of standards-based reform. Increasingly, over the last decade and especially since the passage of the *No Child Left Behind Act of 2001* (NCLB), educators, policy-makers, and researchers at the national, state and district level have become more focused on the importance and value of alignment in the development and implementation of comprehensive standards-based education systems that link content standards, curriculum, instruction, assessment, and professional development to guide student learning in achieving educational goals. This publication is a collection of new research reports that build upon this progress in the understanding and use of alignment systems in K-12 education. These reports together share the same primary purpose of seeking to empower state and district practitioners in improving the use of alignment through the development of an electronic, automated alignment analysis tool and how to extend this tool to alternate assessments and English language development assessments.

This research was made possible by a grant from the U.S. Department of Education (ED) for enhancing assessments (in response to the Request for Proposals OMB 1810–0576) in support of the goals of NCLB. Under this grant, the Oklahoma Department of Education lead this research project and partnered with the Council of Chief State School Officers (CCSSO) to organize and manage a collaborative of 16 states in carrying out this research project. The 16 states engaged in this collaborative were Alabama, California, Delaware, Kansas, Louisiana, Massachusetts, Minnesota, New Jersey, North Carolina, Oklahoma (lead state), Pennsylvania, South Carolina, Texas, Wyoming, West Virginia, and Wisconsin.

Building on prior work done by CCSSO on alignment, the collaborative used the alignment model developed by Norman Webb in 1997. In this model, alignment is defined as “the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do” (Webb, 1997, p. 4). As such, alignment is a quality of the relationship between expectations and assessments and not a specific attribute of either of these system components. Alignment describes the match between expectations and assessment that can be legitimately improved by changing either student expectations or assessments. Noting a central goal of the reauthorization of Title I of the elementary and Secondary Education Act of 1965 (ESEA) under the Goals 2000 Act in 1994, and then under NCLB, was “to use assessment to drive challenging instruction for students,” the first report by Webb articulates the unifying theme of these six research reports: to continue to provide the latest research and tools for supporting education systems in the use of assessment as centered on student learning.

The collaborative sought to automate the alignment process and extend its use by incorporating modifications into the process to make it valid for assessing special populations of students and expand the procedures used so they would be applicable with assessments at all grade levels. The collaborative focused on three major goals:

- |           |   |
|-----------|---|
| GOAL I:   | To produce an electronic, CD-based and web-based alignment analysis process that could be used by state and district staff to enter coding data, automatically producing data analysis and report creation.                             |
| GOAL II:  | To adapt the existing alignment analysis process by refining and modifying decision rules and other procedures to make the process more applicable to the full range of expectations and assessments of all students with disabilities. |
| GOAL III: | To expand the alignment procedures to be applicable for every-grade assessments and vertical scaling.   |

In addition, as the research proceeded, the work was extended to investigate how alignment can improve assessments for English Language Learners.

This project has supported making the automated alignment process system available on the Internet and CDs that can be readily distributed to states, thus increasing the use of the alignment tool in assessment development and verification. The automated Web Alignment Tool (WAT), with a training manual and associated support materials, is available on the Internet at [www.wcer.wisc.edu/wat](http://www.wcer.wisc.edu/wat) and is available at no charge to states and districts. The WAT enables users to conduct alignment evaluations and automatically calculates the results and produces tables needed

for evaluating the alignment of an assessment to standards. By late 2005, the WAT has already been used by over 17 states and by several countries. The rapid dissemination and widespread usage of the tool underscores its important role in determining the alignment between assessments and content expectations in order to ultimately improve and enhance student learning.

This research refined the automated process to include necessary definitions and considerations for applying the technique to assessment activities for students with disabilities and for making provisions for the varied contexts in which these would be administered. Along with automating the process, the collaborative generated the necessary decision-making rules and additional criteria needed in evaluating every-grade assessments and their alignment to content standards. This vertical alignment analysis sought to address both the different models used by states for every-grade assessments and the psychometric issues of vertical scaling. Finally, this research was extended to adapt the Webb alignment methodology to English language proficiency assessments and standards.

As a culminating activity of the automated alignment work of the project, the collaborative hosted two workshops to train state assessment, curriculum, and Special Education staffs in the use of the alignment process. The purpose of the dissemination workshops was to improve the capacity of state department of education staff members to manage and oversee quality control of assessment programs.

This publication is organized as a collection of six reports, each of which can stand alone but are collected here as a useful single reference work for practitioners and researchers.

**THE FIRST CHAPTER**, *The Web Alignment Tool: Development, Refinement, and Dissemination*, by Norman Webb, Meredith Alt, Maria Cormier, Rob Ely, and Brian Vesperman, details the capabilities and development of the Web Alignment Tool (WAT). It includes a description of the Webb alignment system and a study of the reliability of this system as field-tested using the WAT. It also provides a broad review of the research about alignment and describes how the Webb alignment model fits into this body of research.

**THE SECOND CHAPTER**, *Alignment of Alternate Assessments Using the Webb System* by Gerald Tindal, presents a method for determining the alignment between alternate standards and assessments based on the Webb system, including examples from field-testing.

**THE THIRD CHAPTER**, *Assessing Vertical Alignment* by Lauress L. Wise and Meredith Alt, is a concept paper outlining a method for determining the alignment between content standards across grades.

**THE FOURTH CHAPTER**, *Vertical Alignment of Grade-Level Expectations for Student Achievement: Report of a Pilot Study*, Lauress L. Wise, Liru Zhang, Phoebe Winter, Leslie Taylor, and D. E. (Sunny) Becker, presents the results of a pilot study in Delaware on the vertical alignment process. This report, which examines the relationship of content standards and expectations from grade to grade, finds that a review of the consistency of expectations across grades can provide validity evidence for content standards used for accountability under NCLB.

**THE FIFTH CHAPTER**, *Aligning English Language Proficiency Tests to English Language Learning Standards*, by H. Gary Cook, extends alignment research to an investigation of English language proficiency test (ELP) alignment. This report describes a process for ELP alignment, which matches an assessment's linguistic skills and acquisition levels to English language development standards.

**THE SIXTH CHAPTER**, *Developing Aligned Performance Level Descriptors for the English Language Development Assessment K-2 Inventories*, by María H. Malagón, Marjorie B. Rosenberg and Phoebe C. Winter, describes the process and results for developing aligned performance level descriptors for the assessment of English language development. The assessment instrument used in this study is the English Language Development Assessment (ELDA) K-2 inventories

# Alignment Report 1

---

## The WEB Alignment Tool: Development, Refinement, and Dissemination

Norman L. Webb, Meredith Alt, Rob Ely,  
Maria Cormier, Brian Vesperman

December 2005

## Acknowledgements

---

We would like to acknowledge the following people for their invaluable work on the Web Alignment Tool:

Norman Webb, Project Leader

Meredith Alt, Project Assistant

Maria Cormier, Project Assistant

Rob Ely, Project Assistant

Brian Vesperman, Computer Programmer

Lynn Lunde, Secretary

Margaret Powell, Editor

We would like to thank Liru Zhang, Brenda West, and Sarah McManus for their help and support in the alpha and beta testing of the WAT. We would also like to thank Arthur Halbrook, Laress Wise, Gerald Tindal, Phoebe Winter, Ronda Townsend, and all of the members of the Technical Issues on Large-Scale Assessment group for the invaluable expertise and advice they provided on the WAT and on alignment in general, as well as for their work in administering project meetings and conferences.

This work was supported by a grant from the U. S. Department of Education, No. S368A030011, to the Oklahoma Department of Education, with a subcontract to the Council of Chief State School Officers and the Wisconsin Center for Education Research, University of Wisconsin, for the period from April 14, 2003, through September 30, 2005, and a grant from the National Science Foundation, No. EHR 0233445, to the Wisconsin Center for Education Research, University of Wisconsin, for the period from October 1, 2002, to September 30, 2005.

## Executive Summary

---

The following research project originated in response to the Request for Proposals OMB 1810–0576 issued by the U.S. Department of Education (ED) in 2003 for enhancing assessments. As detailed in the proposal ED accepted, a state collaborative organized by the Council of Chief State School Officers and led by Oklahoma developed an electronic, automated system for determining the alignment between the content of state standards and assessments, including methods of determining content alignment across grades and with alternate standards and assessments.

This report details the development of this automated system, including how the system fulfills and exceeds the requirements set forth in the 2003 proposal. In addition to a description of the capabilities of the Web Alignment Tool (WAT) and a timeline of its development, the report incorporates a study of the reliability of Norman Webb's alignment system as employed by the WAT, as well as a review of the WAT within the broader context of alignment. The increasing attention devoted to alignment over the last five years, both by the government and by the educational research community, marks the development of the WAT as an important step in improving the assessments and content expectations that guide student learning.



## Table of Contents

<b>Executive Summary</b>	<b>3</b>
<b>Introduction</b>	<b>7</b>
<b>Accomplishments</b>	<b>8</b>
The Webb Alignment System .....	8
Entry of Content Standards and Assessments .....	9
Ease of Coding .....	9
Creating Studies that Accommodate more Complex Standards Structures .....	9
Other Alignment Criteria or Systems .....	9
Reports .....	10
Training .....	11
Timeline of Alignment Studies .....	11
<b>What We Learned</b>	<b>12</b>
Timeline: Development of the Web Alignment Tool .....	12
Reliability of Coding for the Webb Alignment System .....	14
<b>The Web Alignment Tool in the Broader Context of Alignment</b>	<b>19</b>
Background on Alignment .....	19
Other Interpretations of Alignment .....	21
Alignment Studies .....	21
Alignment Methodologies .....	22
Alignment Models .....	22
The Achieve Model .....	23
The Webb Model .....	24
Issues with Alignment Models .....	26
<b>Conclusion</b>	<b>28</b>
<b>References</b>	<b>29</b>





# The Web Alignment Tool: Development, Refinement, and Dissemination

---

## Introduction

---

This research project originated in response to the Request for Proposals OMB 1810–0576 issued by the U.S. Department of Education (ED) in 2003 for enhancing assessments. As detailed in the proposal ED accepted, a state collaborative organized by the Council of Chief State School Officers and led by Oklahoma developed an electronic, automated system for determining the alignment between the content of state standards and assessments, including methods of determining content alignment across grades and with alternate standards and assessments.

This final report details the development of this automated system, describes how the system developed fulfills and exceeds the requirements set forth in the 2003 proposal. This introduction summarizes these requirements for an automated system for studying alignment. The remainder of the report describes the fulfillment of these requirements, focusing primarily on development of the Web Alignment Tool.

The project goal, specifically addressed in this final report, was to produce an electronic, CD-based alignment analysis process that could be used by state and district staff to enter coding data, automatically producing data analysis and subsequent reports. As an important deliverable, the project produced a CD of the alignment analysis process that provided for incorporating adaptations for Special Education and expansion for vertical alignment. As a second important deliverable, the project produced a written report on field-testing the CD alignment analysis process, including measures of quality, such as agreement among reviewers. And, as a culminating activity, the project held a series of workshops for training state and district staff members to use the CD-based alignment analysis process in conjunction with traditional assessments, assessments of students with disabilities, and every-grade assessment.

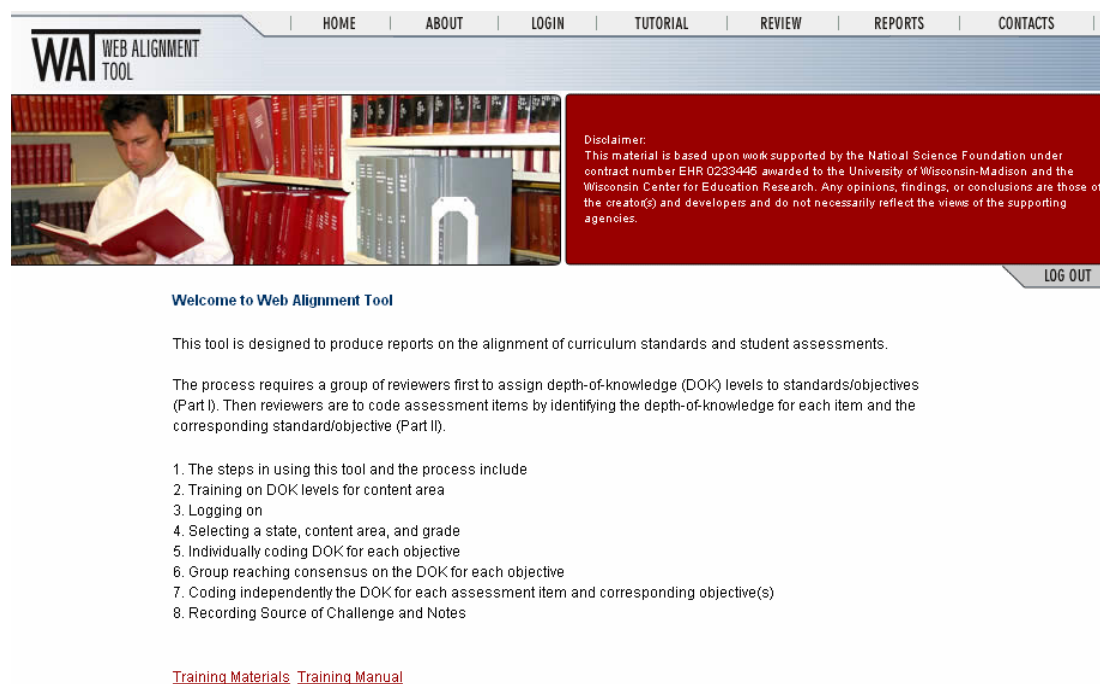
Although the original proposal called for a CD-based alignment system, it became evident in the early stages of the project implementation that an internet-based alignment system would also be useful and in many respects, more versatile. Therefore, the project directed its focus on the development of the Web Alignment Tool (WAT), and to fulfill the proposal requirements, it also made a CD version of the WAT. For sake of clarity, this report refers to the alignment tool as the WAT with the understanding that the CD alignment tool is essentially identical to the web-based version.

By late 2005, the WAT has already been used by over 17 states and by several countries. The rapid dissemination and widespread usage of the tool underscores its important role in determining the alignment between assessments and content expectations in order to ultimately improve and enhance student learning.

## Accomplishments

The primary goal of the project was to produce an electronic, CD-based alignment analysis process that would enable state and district staff members to enter coding data and automatically produce tables for data analysis and report creation. We have accomplished this through the development of the Web Alignment Tool (WAT), which can be found at [www.wcer.wisc.edu/wat](http://www.wcer.wisc.edu/wat) (see Figure 1). We have also developed a CD-ROM version of the WAT, with all of the capabilities of the Web version. Included with both versions is a 175-page Training Manual, which will be discussed later. Provided below is the home page for the Web Alignment Tool, which provides a gateway to the many valuable features of the electronic alignment process:

FIGURE 1. WAT HOME PAGE.



### The Webb Alignment System

Before detailing the capabilities of the Web Alignment Tool, an overview of the alignment system is an important first step. The WAT produces reports on alignment based on five criteria, as specified by Webb (1997):

**Categorical Concurrence** — This criterion measures the extent to which the same or consistent categories of content appear in the standards and the assessments. The criterion is met for a given standard if there are six or more assessment items targeting that standard.

**Depth-of-Knowledge Consistency** — This criterion measures the degree to which the knowledge elicited from students on the assessment is as complex within the content area as what students are expected to know and do as stated in the standards. The criterion is met if more than half of targeted objectives are hit by items of the appropriate complexity.

**Range-of-Knowledge Correspondence** — This criterion determines whether the span of knowledge expected of students on the basis of a standard corresponds to the span of knowledge that students need in order to correctly answer the corresponding assessment items/activities. The criterion is met if at least half of the assessment items have a DOK level as high as the DOK level of the objectives to which they were coded.

**Balance of Representation** — This criterion measures whether objectives that fall under a specific standard are given relatively equal emphasis on the assessment.

**Source of Challenge** — This criterion is met if the primary difficulty of the assessment items is significantly related to students' knowledge and skill in the content area represented in the standards.

In order to determine whether these alignment criteria are met, the WAT is used to create studies for each grade and content area to be examined. Each study consists of pairing content standards from one grade with an assessment. The WAT allows any number of reviewers to be in the group studying the alignment between the standards and the assessment. Once the standards and assessment are entered into the WAT and the study's reviewers are adequately trained in the Webb alignment system and the use of the WAT, the reviewers participate in a two-part alignment process. This process usually takes two to three hours per study (for one grade level).

In Part I, the reviewers first individually determine the depth-of-knowledge (DOK) values of each objective in the grade's standards, and then come together as a group to gain consensus. In Part II of the WAT, the reviewers individually code each assessment item, specifying the item's DOK value, up to three content objectives that the item targets, and any source-of-challenge issues with the item. Once these data are entered, the WAT immediately produces tables describing the degree to which the assessment satisfies the five alignment criteria. It also produces a number of additional tables that aid in the analysis of the alignment and how to improve it. The different WAT capabilities with respect to the complete alignment process are discussed below in detail.

### ***Entry of Content Standards and Assessments***

In order to produce studies using the WAT, the appropriate content standards must be entered into the system. The WAT allows for standards to be entered with either two or three levels of specificity. To support the use of common terminology, the alignment process relies on a three-level system referred to as Standards, Goals, and Objectives, where objectives are the most specific content expectations and standards are the most general. A two-level system omits the goal level. The data are entered into the system at the Objectives level, and the alignment criteria are judged at the Standards level. For ease of entry, the WAT allows for standards to be uploaded from an Excel spreadsheet. The assessment information is even easier to enter into the WAT. The item numbers for all items on the assessment are entered (no item content is entered). When an item number is entered, each item may be specified according to how many points it is worth and whether it is a field-test item. Once a set of standards or an assessment has been entered into the system, it may be edited or used for other studies. It is also possible to download an entered set of standards from the WAT onto a computer, which can save time when using the CD alignment tool.

### ***Ease of Coding***

Once the assessments and standards have been entered and studies have been created, reviewers may begin coding. The WAT allows reviewers to code objectives individually and for these individual responses to be printed all together for use during the consensus process. In Part II of each study, when reviewers are individually coding items, the WAT allows reviewers to easily edit their entries for other items and for the group leader to monitor their progress. The WAT also creates paper coding sheets for standards and items that may be printed as back-up copies, or to make coding easier for some reviewers.

### ***Creating Studies that Accommodate more Complex Standards Structures***

Some standards, such as those for English Language Learning, have multiple dimensions (e.g., an objective may be designated by its content and its proficiency level). Although one can judge the alignment satisfaction of each content standard, it may be desirable to judge the alignment satisfaction of each proficiency level as well. The WAT allows for studies to be created and reported on according to such multiple dimensions.

### ***Other Alignment Criteria or Systems***

Although the WAT is intended to implement the Webb alignment system, it is also flexible with respect to alignment criteria. For example, the cutoff values for a particular study may be changed to reflect a different theory, or a particular state's needs. Cognitive levels other than the DOK levels may also be used, with any number of different levels.

## Reports

The WAT produces the following reports and tables for each study. The first four tables display the numerical information for the relevant alignment criteria, showing “YES,” “NO,” or “WEAK” for each standard according to the degree of alignment between it and the given assessment. By checking a box, these tables can address the different weights of the assessment items so the items count toward the alignment to these different degrees. Additionally, for the following four tables, any number of particular reviewers, items, or objectives may be omitted from the alignment results:

- Categorical Concurrence
- Depth-of-Knowledge Consistency
- Range-of-Knowledge Correspondence and Balance of Representation
- Summary of Attainment of Alignment Criteria—Summarizes information from the previous three tables
- Source-of-Challenge Comments
- Notes by Reviewer
- DOK Values and Intraclass Correlation—Lists the DOK values assigned by all of the reviewers to each item, and produces two statistics measuring reliability across the reviewers: intraclass correlation and pairwise comparison
- Standard Coverage—Shows all assessment items that were coded to each given objective (extremely useful for interpreting alignment results)
- Item Coverage—Shows all objectives that were coded as being targeted by each given assessment item
- Uncodeable Items—Lists items that any of the reviewers felt did not target any of the objectives
- Item-Agreement Report—Similar to the Standard Coverage report, but color-codes entries according to the degree of agreement among reviewers
- High-Low Report—Similar to the Item-Agreement Report, but color-codes each entry according to how much higher or lower the average DOK level for the item is than the consensus DOK for the matching content objective
- Graphs—A graphical representation of the data in the first four alignment tables
- Summary of Inputs, Individual Analysis, and All Reviewer Individual Analysis—Three tables showing various alignment information for individual reviewers
- Group Summary—Shows raw statistics for the entire group in the alignment study
- All Reviewer Data—Shows all quantitative data entered by each reviewer, including pairwise comparison statistics for the reliability of standard coding and objective coding across the reviewers
- Group Consensus—Lists the standards and the group consensus DOK values for each
- Cross Study Report—Shows information for multiple studies that all employ the same set of standards. This table is handy for a state with multiple assessment forms for each grade; it also helps show which of a number of proposed assessments should be adopted by the state.
- Advanced Reporting and Get Formatted Reports—Includes many of the above reports in one page
- Get Word Formatted Reports—Includes many of the above tables, creating them in a Microsoft Word document that can be easily printed and/or incorporated into written reports
- Group Summary By Goal—Describes alignment criterion satisfaction for each goal rather than each standard (only applies with a set of standards that include an intermediate goal level)
- Objective Mean—Describes alignment criterion satisfaction for each objective rather than each standard
- All Reviewer Data Check—Insures that any subsequent changes to the standards in the system do not affect the given alignment study

## Training

In the Webb model, reviewers are trained at the Alignment Analysis Institute in the use of the WAT and in the DOK-level definitions for their appropriate content areas. This process usually takes about two hours. Reviewers may individually learn about the DOK levels and the alignment process using the Tutorial link on the WAT. This link also allows any user to download the Training Manual, a much more comprehensive resource for training. This manual is included as an Appendix to this report. It includes detailed information about all aspects of the WAT, including a step-by-step guide to registering, creating groups, entering standards, assessments, studies, coding data, and producing and interpreting reports. The manual also includes materials for training reviewers, including DOK-level definitions and sample items and objectives, as well as a comprehensive checklist for organizing and coordinating an alignment study, copies of all necessary handouts, and technical computer specifications, instructions, and installation instructions for the CD alignment tool.

## Timeline of Alignment Studies

The following table shows all of the alignment studies that have made use of the WAT. Seventeen states (and one separate nation) have used the WAT to date. An asterisk indicates studies participated in by Norman Webb and/or his WAT staff at the Wisconsin Center for Education Research.

State or Nation	Month of Study	Grades and Subjects Studied
*State A	March 2003	Grades 3-10—mathematics, reading, and science
*State B	August 2003	Grades 3, 5, 8, 10—language arts, Grades 3, 8—mathematics, Grades 4, 6, 8, 11—science
*State C	August 2003	Grade 10—English/language arts, science
*State D	August 2003	Grade 8, 10—mathematics, Grade 10—science
*State B	September 2003	Grades 5, 10—mathematics
State Q	September 2003	Grades 3-8, 11—mathematics, reading (for testing company)
State K	November 2003	Grades 1, 4, 8, 10—alternate assessments in mathematics and language
*State E	January 2004	Grades 3, 5, 7, 9—mathematics
State B	March 2004	Grades 3, 5, 8, 12—social science
*State F	April 2004	Grades 3, 5, 8, 10—mathematics, English/language arts
*State G	May 2004	Grades 2-8, 10—mathematics, reading, and language arts, multiple forms and practice forms
*State G	June 2004	Biology and U.S. History
*State E	October 2004	Grades 4, 6, 8—mathematics, language arts
*State H	December 2004	High school—mathematics, science, language arts, six different assessments for each subject area
*State B	March 2005	Grades 2-9—mathematics, ELA
State I	February and May 2005	Grades 2, 5, 8, 12—ELL
State J	April 2005	Grades 3-10—mathematics, reading, writing, and science
*State K	May 2005	Grades 2, 5, 8, 12—ELL
State L	May 2005	Grade 7—mathematics
State M	May 2005	Grades 2-9—mathematics, reading (for testing company)
*Nation I	June 2005	Grades 1-12—mathematics, science, English
State N	July 2005	Grade 9—mathematics, English (SAT and ACT)
Nation I	August 2005	Grades 1-12—Arabic
State O	September 2005	Grades 3-8—English/language arts, mathematics, science, including NAEP items
State P	October 2005	Grades 3, 8, 10—reading, Grades 5, 7, 9—mathematics

## What We Learned

This section of the report contains two parts. The first is a timeline of events which influenced the development of the Web Alignment Tool. The second section, a study of the reliability of the WAT and the Webb alignment system, describes Deliverables 2 and 5.

### *Timeline: Development of the Web Alignment Tool*

The Web Alignment Tool (WAT) is a product of years of careful testing and refining. The table of alignment studies on the previous page shows that the workers on this project have been directly involved in studies for 11 states and one separate nation. The experience accumulated by conducting these studies and addressing the various issues raised by them has made it possible to develop the WAT as it is today. The following is an account of the issues that arose over the course of the WAT's development and the changes made to the WAT and to the Webb alignment system in order to address these issues.

Date	Event	Description of Event
Dec 2002	Grant begins	
March 2003	Alignment Study—State A	Grades 3-10—Mathematics, reading, and science. Coding was done with pen-and paper, but the data computation and analysis used the fledgling WAT. TILSA member from several states participated and gave feedback on the alignment process.
August 2003	Alignment Study—State B	Grades 3, 5, 8, 10—Language arts; Grades 3, 8—Mathematics; Grades 4, 6, 8, 11—Science. Grade 10—English/language arts and science for State C. Grade 8 and 10—Mathematics and Grade 10—Science for State D.  This study served as the alpha test for the WAT, and was entirely performed using the automated system.  Questions were raised about how standards should be entered into the WAT. State B had multiple versions of the standards—one for instructional guidance, one for assessment specifications, etc. Which should be used in the alignment study? The state also had more than three levels of specificity. At which level should the data be coded, and at which level should the information be reported out?  These questions arising through the course of the alignment study underscored how important it is to make the appropriate decisions about standards at the beginning. Our experience and advice about this went into Part IV of the Training Manual: the content expectations to use should be the ones that provide most of the instructional guidance in the state. The two or three levels of specificity selected should be specific enough to allow for DOK coding (that is, they describe sets of activities).
Sept 2003	Alignment Study—State B	Grades 5, 10—Mathematics. This served as the beta test for the WAT, fully automated and with more reports for ease of analysis.
Jan 2004	Alignment Study—State E	Grades 3, 5, 7, 9—Mathematics. This served as the beta test for the CD alignment tool.
Feb 2004	TILSA Mtg—San Diego	This meeting, the first of several, provided a chance for members of the three working groups to report on their progress and to get feedback from the TILSA members.  At the conference, states raised the concern that since some assessment items counted for multiple points toward student scores, there should also be a way for these items to count for multiple points toward the alignment results.  At the conference, states also recommended a way to change the cutoff values for the alignment criteria, depending on the features of the particular state's standards and assessments.  Now when an assessment is entered into the WAT, it is

		<p>possible to enter item weights for each item. The reports may (optionally) weight the items accordingly toward the alignment results.</p> <p>When reports are produced, the cutoff values for the alignment criteria may be adjusted. The Training Manual cautions states to provide a rationale for any such changes.</p>
April 2004	Alignment Study—State F	<p>Grades 3, 5, 8, 10—Mathematics, English/language arts. Each grade level report now includes roughly 6 tables in the written portion and 11 tables in the Appendix.</p> <p>After the study was completed, it came to our attention that some of the assessment items were field test items and thus should not count for alignment purposes.</p> <p>To account for field test items or other items that should be omitted from the alignment analysis, when an assessment is entered it is now possible to designate certain items as “missing.”</p>
May 2004	Alignment Study—State G	<p>Grades 2–8, 10—Mathematics, reading, and language arts. Three assessment forms for each grade and subject area, plus several additional practice forms for the high school grades and writing prompts for grades 4 and 7. Within the following month, we conducted studies in high school Biology and U.S. History, where the training and consensus processes were done using conference calls.</p> <p>Because of all the forms and grades, this alignment institute consisted of 73 different studies. The reports therefore required the production and formatting of a thousand tables—a daunting task to do by hand.</p> <p>In one group, we noticed that the internal reviewers coded item DOK values in a way consistently different from that of the external reviewers. This issue indicated problems in the consensus process, training, and ultimately in reporting.</p> <p>The Get Word Formatted Reports option was developed, saving many hours of work. When a group leader selects this report, a Microsoft Word file is immediately created with 11 of the most important tables for the study formatted and correctly labeled.</p> <p>We did three things to address this issue:</p> <ol style="list-style-type: none"> <li>1) We added to the Training Manual a section of advice for Group Leaders on ways to monitor and involve all of the reviewers.</li> <li>2) We clarified the DOK definitions and improved the training materials, adding specifically selected sample items to the training, many of which have DOK values on the borderline.</li> <li>3) We added Advanced Reporting, which allows the temporary removal of one or more outlying reviewers from the alignment analysis. This option also allows the omission of any number of standards or objectives from the alignment analysis.</li> </ol>
June 2004	TILSA Mtg—Boston	This conference directly preceded the national Large-Scale Assessment conference in Boston.
Oct 2004	Alignment Study—State E	Grades 4, 6, 8—Mathematics, language arts.
Dec 2004	Alignment Study—State H	<p>High school science, mathematics, and reading/language arts—six assessment forms for each subject, including MEAP, SAT, ACT, PSAT, PLAN, and WorkKeys. The purpose of this alignment study was explicitly to determine the alignment costs and benefits of replacing the Michigan-designed assessment (MEAP) with a college-entrance examination, and how to supplement such an examination with further items.</p> <p>Because of the purpose of this alignment study, we needed some way to easily compare the alignment and content coverage of different assessments with the same set of standards.</p> <p>We developed the Cross Study Report, which allows a group leader to select any number of assessments measuring the same set of standards. The report shows the differences in how these assessments target the standards. This report has also been convenient for states with multiple assessment forms per grade.</p>

Feb 2005	Deliverable 5: Dissemination Conference I—Tempe, AZ	<p>The first of the dissemination conferences, in which representatives from the western states were invited to learn about the WAT and were given drafts of the Training Manuals. The conference had two sessions. Session A involved a complete sample alignment study and Session B involved learning how to interpret the alignment reports and how to organize an alignment institute. About 35 participants.</p> <p>The primary difficulty in this dissemination conference was due to faulty and fickle wireless Internet connections at the site.</p> <p>In Session A, we realized that the process by which group leaders enter standards was needlessly complicated, confusing, and time-consuming.</p> <p>To save WAT users from similar problems at alignment institutes, in the Training Manual we recommend hardware Internet connections rather than wireless ones.</p> <p>In Session B, it was pointed out that determining how to improve Depth-of-Knowledge Consistency was very complicated using the current computer alignment reports.</p> <p>We changed the standard entry process in two ways. First, the nested structure of the standards, goals, and objectives can be imposed by increasing or decreasing the indent on the entries, rather than by having to reroute hierarchy trees. Second, standards can be uploaded from a spreadsheet, rather than having to be typed by hand.</p> <p>We created the High-Low Report, which enables an analyst, using only one table rather than three, to see which items have DOK values higher or lower than the objectives to which they were coded.</p>
March 2005	Alignment Study for ELL standards —State H	<p>This was the first alignment study in which the WAT was used for English Language Learner (ELL) standards. These standards were more complex because they had both a content structure and a proficiency level structure.</p> <p>The ELL standards had two dimensions of structure, but the WAT could only handle one. There had to be some way to analyze the alignment according to both structures, rather than having to enter all of the data into the system twice.</p> <p>We created the Additional Standard Structure option, which allows additional standard structures to be easily superimposed on a particular set of standards. Likewise, the alignment data can be displayed for any and all of these overlaying structures.</p>
June 2005	Alignment Study —Nation I	<p>Grades 1-12—Mathematics, science, and English. This was the first alignment study we performed for another country.</p>
July 2005	Deliverable 5: Dissemination Conference II—Boston	<p>The second of the dissemination conferences, with the same structure as the first. Open to representatives of the eastern states. About 30 participants</p>
September 2005	Grant ends	

## Reliability of Coding for the Webb Alignment System

### I. Introduction

The purpose of this analysis is, in part, to assess the reliability of the Webb alignment system. Our analysis uses data collected in 34 alignment studies during seven different alignment institutes held from 2003 to 2005. We chose these particular alignment studies for analysis because they span the development of the Web Alignment Tool (WAT), because they represent all four content areas allowed for by the WAT (mathematics, language arts, science, and social science), and because they represent a variety of types of standards, assessments, and studies.



This study is structured as follows: First, we describe the reliability statistics and criteria that are used. Then these statistics are displayed for each selected alignment study in Table 1. These results are analyzed and discussed in light of the characteristics of each selected alignment study. We conclude with a brief description of how the reliability of coding affects the reliability of the alignment results based on Webb's four alignment criteria.

## II. Measures of Reliability of the Coding

When using Webb's alignment system to determine the alignment between a set of content objectives, participants must enter three kinds of data: Depth-of-knowledge (DOK) levels of each content objective, DOK levels of each assessment item, and the objective(s) targeted by each assessment item. In order to assess the reliability of this alignment system, we must therefore assess the reliability of these three types of data.

The DOK levels of the content objectives are determined by a process of consensus among the study participants (reviewers). Assessing the reliability of this kind of data, which would require an experimental analysis, is beyond the scope of the study. The reliability in assigning DOK levels to the content objectives could be determined by randomly assigning reviewers to groups who code the same standards and then comparing the results. Resources were not available to conduct such a study. Rather than having multiple groups agree on the DOK levels assigned to the content objectives, the process depends more on reviewers within a group consistently applying the DOK levels to the objectives and to the assessment items. The results are reported on the comparison between the DOK levels of the assessment items with the DOK levels of the content objectives.

The other two kinds of data were analyzed for reliability using standard reliability statistics. Since DOK levels can be analyzed as categorical data, Levels 1 to 4, these data are analyzed using intraclass correlation. However, since the intraclass correlation cannot be used with low variance, the reliability of these data is also calculated using pairwise agreement among reviewers. Pairwise agreement is a more stringent comparison, but can produce more accurate results on consistency particularly when there is high agreement among reviewers and minimal change of values across items. We have adjusted the traditional method for calculating pairwise comparison results to allow for reviewers to code from 0 to 3 objectives to each assessment item. A detailed explanation of our pairwise comparison statistic is shown below.

### a. Intraclass Correlation for DOK-Level Coding

The intraclass correlation is calculated according to the method of Shrout and Fleiss (1979):

$$ICC = \frac{\sigma^2(i)}{\sigma^2(i) + \sigma^2(r)}$$

Here,  $\sigma^2(i)$  is the variance in the data between the assessment items, and  $\sigma^2(r)$  is the variance in the data

between the reviewers. In other words, the statistic measures the percent of variance in the data due to the differences between the items rather than the differences between the reviewers. An intraclass correlation value of, say, 0.7, means that 70% of the variance in the data can be explained by differences between the items, while the other 30% is due to differences between the reviewers. Intraclass correlation is considered adequate for values greater than 0.7 and good for values greater than 0.8. Of course, in cases where there is very low variance between the items, the intraclass correlation statistic can be meaningless or misleading. In such instances, it is preferable to use pairwise agreement as a reliability measure instead.

### b. Pairwise Agreement for DOK-Level Coding

The pairwise agreement for a particular assessment item is calculated as follows: For each possible pair of reviewers, determine whether the two reviewers assigned the item to the same DOK level or not. Divide the number of agreeing pairs of reviewers by the total number of possible pair comparisons considering each pair of reviewers. The quotient of the pairs of agreement over the total possible pairs is computed over all of the assessment items.

### c. Pairwise Agreement for Objective, Goal, and Standard Coding

The WAT allows a reviewer to code each assessment item as measuring up to three different content objectives. Reviewers may instead mark the item "uncodeable." For any given assessment item, the pairwise agreement for objectives is calculated as follows:

First, choose a pair of reviewers. Find the reviewer who coded the greater number of objectives to this item, and call this number  $n$ . Now take the number of entries the two reviewers agree on and divide this by  $n$ . This gives the agreement between the two reviewers. Perform this calculation for all possible pairs of reviewers, and take the sum of the agreements. Then divide this sum by the total number of pairs of reviewers. This is the pairwise agreement value for the given assessment item. An example is provided in Figure 2. This calculation method of the pairwise correlation was used for reporting the results in this report. However, subsequently the computation method has been changed to dividing the total number of agreements (numerators) by the total number of comparisons (denominators). This method of calculation better represents when values when there is either perfect agreement or no agreement.

The pairwise agreement for objectives is averaged over all of the assessment items to give the pairwise agreement for objectives statistic for the alignment study as a whole.

This method of calculating pairwise agreement assumes that two reviewers should not be considered in full agreement on an item unless they code it identically. The reason for this is that if a reviewer codes, say, two objectives to a given item, the WAT will count this item twice for alignment purposes (for this particular reviewer), once for each coded objective. For this reason, reviewers are trained to code multiple objectives to an item only if the item fully targets each objective.

The pairwise agreement for goals and the pairwise agreement for standards are calculated similarly. These calculations are also represented in Figure 2. Notice that the number of goals or standards coded by a reviewer to an item may be less than the number of objectives this reviewer coded. For example, in Figure 1, Reviewer A coded three different objectives, but only two different goals (1.1 and 2.1). Therefore, in a pairwise comparison of Reviewer A and Reviewer B at the goal level, the two reviewers agreed on one out of two coded goals, even though reviewers only agreed on one out of three coded objectives.

**Figure 2. Example of a pairwise agreement calculation for one assessment item.**

Reviewers								
A			B	C			D	
1.1.b	1.1.c	2.1.d	1.1.c	1.1.b	1.1.c	2.2.a	1.1.b	2.2.a

Reviewer Pair	Agreement *		
	of Objective	of Goal	of Standard
(A, B)	1/3	1/2	1/2
(A, C)	2/3	1/2	1
(A, D)	1/3	1/2	1
(B, C)	1/3	1/2	1/2
(B, D)	0	1/2	1/2
(C, D)	2/3	1	1
Sum of agreements	2.333	3.5	4.5
Pairwise Agreement	0.389	0.583	0.75

Note: This method was used to compute the pairwise agreement in the analyses reported here. However, subsequently the computation method for computing pairwise agreement for the WAT was changed by computing the sum of the numerators and dividing by the sum of the denominators.

The pairwise agreement statistic almost always returns a lower value than the intraclass correlation statistic. Usually, a pairwise comparison result of 0.7 or higher is considered to reflect good agreement, 0.6 or higher reflects adequate agreement (four out of five reviewers agreeing), and 0.5 or less reflects poor agreement. See Figure 3 for some examples. It is also noteworthy that the method used above for dealing with multiple reviewer responses tends to lower the pairwise comparison statistic, so these guidelines are relatively strict.

**Figure 2. Examples of pairwise agreement results.**

Reviewers' Responses	Pairwise Comparison
(6 reviewers)	
a a b b c c	0.2
a a a b b b	0.4
a b b b b b	0.66
(9 reviewers)	
a a a b b b c c c	0.25
a a b b b b b b b	0.61
a b b b b b b b b	0.75

### III. Studies Chosen for Analysis

We analyzed 34 alignment studies from seven different alignment institutes. We chose these studies to represent a broad range with respect to the following factors:

- **Time**—we include the first study after the beginning of the TILSA working grant (State A, May, 2003), as well as a recent study conducted using the Webb system (State B, March, 2005). Some of the studies used the Web Alignment Tool (WAT) in different stages of its development. Additionally, training materials were developed to some extent over this two-year period.
- **Content area**—we include studies for social science, mathematics, reading, language arts, and science.
- **Grade level**—from grade 2 to grade 12
- **Group size**—from three reviewers to nine reviewers
- **Types of standards**—from a set of standards with 4 total objectives (State E Grade 4 Language Arts) to a set with 77 objectives (State H Mathematics).
- **Variance**—for instance, the variance of the DOK coding was so small for the State H Science PLAN study that it resulted in a nonsensical intraclass correlation coefficient (see Table 1).

### IV. Reliability Results

Table 1 shows reliability measures for the two types of coding involved in the Webb alignment system: DOK coding and Objective coding.

Table 1

Reliability Statistics for Several Studies (2003–2005)

State	Grade level	Reviewers	Objectives/Standards Analyses		Item DOK Analyses	
			Objective Pairwise Comp.	Standard Pairwise Comp.	DOK Intraclass Correlation	DOK Pairwise Comp.
State A Mar. 2003	Grade 3 Math (II)	8	0.76	0.94	0.93	0.76
	Grade 8 Math (II)	8	0.77	0.94	0.94	0.72
	Grade 3 Reading (II)	8	0.55	0.93	0.84	0.55
	Grade 8 Reading (II)	8	0.60	0.94	0.79	0.58
	Grade 3 Science (II)	8	0.62	0.78	0.83	0.66
	Grade 6 Science (II)	8	0.44	0.70	0.87	0.68
State B Aug. 2003	Grade 10 Science (I)	4	0.46	0.84	0.74	0.62
	Grade 8 Language Arts	5	0.39	0.82	0.77	0.47
	Grade 10 Language Arts	8	0.43	0.91	0.88	0.55
	Grade 3 Math	6	0.43	0.72	0.86	0.59
March 2004	Grade 8 Math	6	0.34	0.62	0.87	0.60
	Grade 4 Social Studies	7	0.62	0.91	0.93	0.56
March 2005	Grade 8 Social Studies	8	0.45	0.80	0.91	0.56
	Grade 2 ELA	7	0.49	0.91	0.86	0.68
Apr. 2005	Grade 9 ELA	6	0.35	0.84	0.84	0.49
	Grade 4 Math	7	0.33	0.62	0.87	0.63
State G May 2004	Grade 9 Math	6	0.40	0.67	0.78	0.66
	Grade 6 Math Form B	5	0.58	0.76	0.75	0.48
	Grade 6 Math Form C	5	0.64	0.88	0.86	0.62
	HS Math Practice Form 1	3	0.47	0.69	0.60	0.67
State E Oct. 2004	HS Math Practice Form 2	3	0.48	0.64	0.69	0.61
	Grade 4 Math	8	0.80	0.91	0.82	0.69
	Grade 8 Math	8	0.70	0.89	0.83	0.61
	Grade 4 ELA	8	0.47	0.47	0.80	0.49
State H Dec. 2004	Grade 8 ELA	8	0.66	0.66	0.78	0.61
	HS ELA HEAP	9	0.40	0.48	0.88	0.66
	HS ELA ACT	9	0.48	0.60	0.83	0.58
	HS ELA PLAN	9	0.59	0.69	0.84	0.56
	HS Science HEAP	9	0.57	0.77	0.81	0.53
	HS Science ACT	9	0.68	0.86	0.30 *	0.76
	HS Science PLAN	7	0.46	0.68	-0.086 *	0.81
	HS Math HEAP	9	0.36	0.74	0.84	0.61
	HS Math ACT	9	0.30	0.71	0.83	0.68
	HS Math PLAN	9	0.30	0.75	0.85	0.66

\* Low variance among DOK levels assigned to items by reviewers makes the Intraclass Correlation invalid. In such cases, the pairwise comparison is a better measure of agreement.

The intraclass correlation results for DOK-level coding are all greater than the desired value of 0.7, except for four studies. Two of these results, the State G high school mathematics practice examinations, were due to the small number of reviewers (3). The other two, the State H Science ACT and PLAN assessments, were due to a very small variance in the DOK entries. For such situations, the intraclass correlation statistic is not meaningful, and the reliability can better be measured by pairwise comparison. For these two studies, the pairwise comparison results are quite high.

The pairwise comparison results for standard coding are all greater than the desired value of 0.6 except for two studies. The State E English/Language Arts (ELA) study for grade 4 is an unusual case, due to State E's standards. The Language Arts framework includes only four standards with no goals or objectives beneath them. This is why the pairwise comparison is the same for objectives and for standards for the State E ELA studies. The lower pairwise comparisons are because reviewers had a difficult time distinguishing between two of the standards. The other study with a low pairwise comparison for standards is the State H HEAP English/Language Arts study. This relatively low result is due to the unusually large number of standards (12) in the Language Arts framework and the inability of reviewers to distinguish among some of the standards. The pairwise comparison results for objective coding are much lower. Half of the studies showed relatively poor agreement. As expected, half of the studies with the lowest agreement (below 0.37) used the State H high school mathematics standards, which contained the greatest number of objectives of all the analyzed studies (77).

Reliability appears to have changed very little over time. For instance, we can compare the results from the two State B institutes that were nearly two years apart. Although both institutes had similar numbers of reviewers, the reliability measures are almost identical. As expected, use of the WAT did not change the coding reliability, but made the process faster and more efficient. Reliability is more a function of training than the tool or the process.

There is one improvement in reliability over time, which can be seen by comparing the 2003 State B mathematics studies with the 2004 State G mathematics studies. Even with fewer reviewers, the State G studies have equal DOK-level coding reliability and higher objective coding reliability than the State B studies. This might be attributed to developments in the training materials over this time. On the other hand, the State B mathematics standards contain a greater number of objectives than the State G mathematics standards.

Objective and standard coding reliability do not appear to change with grade level, although DOK item coding reliability tends to slightly decrease at higher grade levels. This may be because assessments for lower grade levels often contain items that are easier to identify at Level 1.

As expected, content area does not appear to be highly related to objective and standard coding reliability, but is slightly related to DOK-level coding reliability. This is perhaps due to the greater precision of DOK-level definitions in some content areas. The intraclass correlation for mathematics studies is generally a bit higher than for English/Language Arts and science, but lower than for social studies.

Finally, we note that group size is the largest factor involved in a change in reliability. The reliabilities increase as the number of reviewers increases, with only a few exceptions. For the best reliability results, this system should be employed with groups of eight or more reviewers, and will more likely produce lower reliabilities with three or fewer reviewers.

## **V. Reliability of the Webb Alignment Criteria**

Webb's four quantitative alignment criteria are Categorical Concurrence, Depth-of-Knowledge Consistency, Range-of-Knowledge Correspondence, and Balance of Representation.

For a given alignment study, the Categorical Concurrence criterion is met if each standard is targeted by at least six assessment items. For this criterion to be reliable, the coding must simply be reliable at the standard level. The standard pairwise comparison statistic is below a minimum level (0.5) for only four (11%) of the studies analyzed, so we consider the Categorical Concurrence statistic to be generally reliable because it is computed at the standard level.

For a given study, the Depth-of-Knowledge Consistency criterion can be roughly understood as being met for a balanced standard if at least half of the assessment items have a DOK level as high as the DOK level of the objectives to which they were coded. For this criterion to be reliable, the DOK item coding and the objective coding must be reliable. The DOK coding reliability is adequate for 94% of the studies we analyzed. The

objective coding is less reliable—less than half of the studies had an objective pairwise comparison statistic greater than 0.5.

In a study with an objective comparison value of 0.5, we estimate that a balanced standard could have a Depth-of-Knowledge Consistency value error of up to 0.13, just from variance in the coding of objectives. Furthermore, the smaller DOK coding variance may add up to 0.05 to this error. However, since objectives within the same standard generally have similar DOK values, and since the standard pairwise comparison statistic is reasonable for most studies, we expect the actual maximum error to be significantly smaller. Nonetheless, it is reasonable to examine with extra care standards within 0.1 of the Depth-of-Knowledge Consistency cutoff value.

The Range-of-Knowledge Correspondence criterion is met for a standard in a given study if 50% of its objectives are targeted by at least one item. This criterion's reliability clearly depends on the reliability of the objective coding. In a study with an objective pairwise comparison value of 0.5, we calculate that a standard may have a Range-of-Knowledge Correspondence value error of up to 0.1. However, this is extremely unlikely, especially given the reasonably high standard pairwise comparison statistics for most studies. This error decreases even more for standards where objectives are targeted by multiple items.

The Balance-of-Representation criterion is met for a standard in a given study if the items targeting that standard are reasonably evenly distributed across the objectives under that standard. This criterion also depends on the reliability of the objective coding and the standard coding. This is also probably the criterion influenced the most by reviewer fatigue; a fatigued reviewer may look less closely to find the best matching objective for an item, choosing instead a more familiar and often-coded objective.

Reported alignment results are derived from taking an average of the results from the individual reviewers and reported at the standard level. With eight reviewers, the preferred number, our experience has shown that we get reasonably high reliabilities in coding the depth-of-knowledge levels of items and the standard measured by the item. When reviewers did vary in their judgments, the averages lessened the error that might result from any one reviewer's finding. Standard deviations are reported, which give one indication of the variance among reviewers. Some variation among reviewers can be the result of an alignment issue. In some states, the curricula standards can overlap. When two reviewers code one item to two different objectives or even standards, this can indicate the lack of precision in how the standards are written. This means that in any alignment study the lack of agreement among reviewers needs to be analyzed to determine if the cause is an alignment issue or training.

---

## The Web Alignment Tool (WAT) in the Broader Context of Alignment

---

In recent years, states have shown a heightened interest in methods of determining the alignment between curriculum and assessment. Now more frequently looked at in terms of the relationship between state content standards and state assessments, alignment has become a subject of greater concern as it relates to standards-based reform and accountability measures.

### **Background on Alignment**

Bhola, Impara, and Buckendahl (2003) define alignment as “the degree of agreement between a state's content standards for a specific subject area and the assessment(s) used to measure student achievement of these standards” (p. 21). Other definitions also stress alignment as a relationship between standards and assessments, with Webb (1997) defining alignment as “the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do” (p. 4).

Concern with alignment in recent years stems in part from federal, as well as state, legislation, originating in the 1990s and elaborated in the No Child Left Behind Act. This Act requires aligned assessments and standards.

Still, the concept “has existed in one way or another almost since the beginning of the use of formal tests to aid in decision making” (Bhola, Impara, & Buckendahl, 2003, p. 22). Earlier attention to alignment was evident in validity concerns with assessment, which have been heightened in an era of high-stakes testing.

Alignment can be seen to relate to at least three different aspects of test validity: content validity, construct validity, and consequential validity (Ananda, 2003a). Alignment is first related to content validity, the degree to which content on assessments reflects the targeted content domain. Strong alignment between content on assessments and standards is frequently used as evidence of assessment content validity, given that the purpose of the assessment is to measure the content knowledge as specified by the standards. Alignment is also closely related to construct validity, the extent to which a test accurately measures the concepts or abilities (i.e., constructs) it is intended to measure. Demonstrating alignment of standards and assessments across grade levels to reflect presumed changes in a developmental construct offers supporting evidence of construct validity (Ananda, 2003a). Alignment is also related to consequential validity, the degree to which an intervention such as an assessment system achieves its intended purposes. Both the content validity of alignment and the consequential validity “are ultimately concerned with the social consequences of testing, including the desired outcome of improved student learning” (Ananda, 2003a).

Prior to the 1990s, alignment in terms of validity concerns was primarily the domain of education and psychology researchers (Cronbach & Meehl, 1955; Messick, 1975). The quantity and significance of education research on validity grew, in part, because the introduction of the Title 1 program in 1965 (and changes and reauthorizations since then) dramatically increased the number of tests that states and districts receiving federal education funding were required to administer (National Research Council, 1999). However, the federal requirement of Title 1 that schools test students using nationally normed tests meant that student performance could be compared against a nationally representative norming group, but the tests were not analyzed for their alignment to an objective standard (National Research Council, 1999). In the 1990s, amid concerns that the normed tests did not provide sufficient information about students’ knowledge and skills and in the context of reformers’ focus on tests since the 1980s, the federal government introduced changes to Title 1 (Elementary and Secondary School Act, 1965). Changes introduced in 1994 through the Goals 2000: Improving America’s Schools Act mandated that states create standards and tests that could be compared to the state standards (National Research Council, 1999). Of the changes brought about through legislation, some of the most far-reaching were those in the assessment arena (National Research Council, 1999).

As part of the new requirements, Goals 2000 for the first time introduced the concept of alignment into legislation, noting that “. . . 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). A central goal of the reauthorization of Title 1 was to use assessment to drive challenging instruction for all students, which could only be achieved if the assessments were appropriately aligned with the standards (National Research Council, 1999). The U.S. Department of Education’s explanation of the Goals 2000 Act further highlighted alignment of curriculum, instruction, professional development, and assessments as a key performance indicator for states, districts, and schools striving to meet challenging During the 1990s, legislators also made alignment of assessments for students with disabilities a new issue by noting that the assessment guidelines developed in 1994 were applicable to all students (U. S. Congress, 1994). In 1997, Final Regulations for the Amendments to the Individuals with Disabilities Education Act specifically addressed alternate assessments, stating that these assessments “need to be aligned with the general curriculum standards set for all students” (U. S. Department of Education, 1997).

The education legislation of the 1990s led many states to adopt content standards; however, few states used formal processes to develop assessments that were aligned with the standards (Case, Jorgensen, & Zucker, 2004). In comparison to the earlier legislation, the No Child Left Behind Act of 2001 brought about both increased mandatory testing provisions and additional attention to alignment. Specifically, state eligibility for federal funds under Title 1 of NCLB requires adoption of content standards in reading, mathematics, and science, along with annual testing programs in reading and mathematics for students in grades 3 through 8 and once in high school. The act also requires, by 2007–8, assessment in science at least once in grades 3 through 5, once in grades 6 through 9, and once in grades 10 through 12 (No Child Left Behind Act of 2001). The stipulations of NCLB have led almost all states to develop content standards in reading and mathematics and to employ a form of statewide assessment (Bhola, Impara, & Buckendahl, 2003). The assessments measure state progress toward proficiency in each subject area, with repeated failure by schools or districts to meet the requirements for Adequate Yearly Progress (AYP) yielding greater consequences than did the 1990s legislation.

The No Child Left Behind legislation requires that assessments be aligned with each state’s content standards for each subject and grade level (Case, Jorgensen, & Zucker, 2004). The assessments may be designed by the

state, or they may be pre-existing tests that have been augmented in order to be aligned with state standards (Ananda, 2003a). For states that select to use a norm-referenced test (NRT), NCLB requires that each state perform an independent alignment study relative to the state's content standards to identify omitted content. States are then required to augment the assessment based on the findings (Ananda, 2003a).

The Individuals with Disabilities Education Improvement Act of 2004 (IDEA), which is closely paired with the No Child Left Behind Act, also carries alignment implications (No Child Left Behind Act of 2001). The Act asserts that states must provide alternate assessments that "are aligned with the State's challenging academic content standards and challenging student academic achievement standards" (H. R. 1350: IDEA, 2004). Up to 1% of the "proficient" scores that count toward annual yearly progress (AYP) may be drawn from students who are assessed using an alternate assessment based on alternate standards. These alternate standards may be different standards entirely, or they may be off-grade-level standards. Whatever form the alternate standards take, IDEA states that assessments for students with disabilities should be aligned with those alternate standards. In cases where assessment is tied to alternate standards, IDEA also requires the development of benchmarks or objectives gauging students' goals based on their Individualized Education Program (IEP) (IDEA, 2004). Since alternate assessments are often in non-standardized formats like portfolios and work samples, methods for determining alignment may include the development of reasonable sampling procedures and parsing the sampled assessments into discrete "items." For further information on issues and procedures regarding the alignment of alternate standards and assessments, see Tindal (2005).

### **Other Interpretations of Alignment**

The growing attention to and research on the alignment between state standards and state assessments has also given rise to new questions about other types of alignment in education systems, or the varied ways in which instruction and assessment should be aligned. For instance, Porter (2004) identifies horizontal alignment as the degree to which state tests align with state standards. But, he also notes questions that arise with vertical alignment, such as whether the district test is aligned to the state test, or whether the district standards are aligned to the state standards (Porter, 2004).

The term "vertical alignment" is also used to describe the appropriate progression of content in the standards from one year to the next. Wise (2005) has developed a set of criteria for describing the relationship between content in one grade's standards and the content in the previous grade's standards, including the proportion of the content that is broadened from the previous grade, deepened from the previous grade, or is simply material that is new to the grade level. Alignment of performance standards is also an issue with this kind of vertical alignment; Lewis and Haug (2003) point out that performance expectations should increase coherently across grades and scores in order for the measuring of proficiency at a specific grade level to relate appropriately to scores at the grades above and below that grade.

Beyond assessment and standards, alignment is also related to curriculum. Improving student learning relies on a coherent curriculum, that "exemplifies both broadly and concretely the intentions of the standards, and the content and skills to be taught and learned" (Baker, 2004, p. 15). As they have come to recognize the importance of alignment between curriculum and standards, many states now offer teachers better training and elaboration of the types of content and the skills that reflect the standards. Some alignment studies (Wise, Harris, Sipes, Hoffman, & Ford, 2000) also address the alignment of the curriculum being taught in the classroom with assessment items. By asking panel members questions about the percentage of their school's students (at specific grades) that had been provided sufficient instruction to answer the assessment items, the studies recognize a distinction between the intended and enacted curriculum.

Alignment can also be seen in terms of student identities and the manner in which students align their actions with the activities or communities in which they want to participate (Nasir, 2002). For instance, as students imagine their own connection to a broader community (of scholars, of athletes, of other chosen fields), they align their actions by emulating or adopting the practices and requirements of those communities (Nasir, 2002). From numerous potential relationships, research on alignment now seeks to identify the critical places in which alignment must operate in order to improve student performance (Baker, 2004).

### **Alignment Studies**

In an era of standards-based reform and accountability, states and districts seek alignment studies of their standards and assessments to serve a range of purposes (Ananda, 2003b). States may seek to identify areas of vulnerability that signal possible problems for their alignment, such as content gaps on their assessment. More detailed alignment studies may be sought to provide information about whether or not to restructure an existing assessment system, or about which changes to implement. States may also seek information about



how their standards and assessments compare to those in other systems. They may desire to use alignment analyses for future item-development activities in cases where specific standards are underrepresented on an assessment. Finally, they may seek evidence of content validity from an independent, expert source (Ananda, 2003b).

### **Alignment Methodologies**

The models of alignment in use today typically incorporate one of several alignment methodologies. Webb (1997) and Case, Jorgensen, and Zucker (2004) classify three traditional methodologies for analyzing the alignment between standards and assessments: sequential development, expert review, and document analyses.

#### **Sequential Development**

With sequential development, standards are developed first and then used by test developers to shape the structure and content of the assessment (Case, Jorgensen, & Zucker, 2004). The standards outline what assessment items should address, and this order of development allows the standards and assessments to be open to the public for review. Test developers can indicate which items correspond to each standard, and public review can ensure that all items are appropriately aligned (Case, Jorgensen, & Zucker, 2004).

#### **Expert Review**

With expert review, specialists analyze the relationship between standards and content, generally after both have already been developed. This frequently involves item-by-item review by people who have been trained to assess the alignment (Case, Jorgensen, & Zucker, 2004).

#### **Document Analyses**

In document analyses, standards and assessments are analyzed using a system for encoding their structure and content. After each set of documents is encoded, their alignment can be systematically compared (Case, Jorgensen, & Zucker, 2004).

### **Alignment Models**

With increased attention to alignment, models of alignment have become increasingly sophisticated (Case, Jorgensen, & Zucker, 2004). Additional guidelines provided under NCLB added new considerations to older ways of approaching alignment, stating that assessments must “measure the depth and breadth of the state academic content standards for a given grade level,” with attention to qualities such as comprehensiveness and clarity for users (U. S. Department of Education, 2003, p. 12).

Alignment models in use today range from low complexity to high complexity, and include at least four different models. The operational models include the Survey of the Enacted Curriculum, the Achieve model, the Council for Basic Education (CBE) model, and the Webb model (now the Web Alignment Tool) (CCSSO, 2002). These models will be outlined in further detail in the following discussion of low-, moderate-, and high-complexity models of alignment. The categorization of the models is based on Bhola, Impara, and Buckendahl's (2003) discussion of the specific models.

#### **Low Complexity**

A low-complexity alignment model forms the basis for research to provide evidence of the content validity of test scores (Impara, 2001). This model defines alignment as “the extent to which the items on a test match relevant content standards (or test specifications)” (Bhola, Impara, & Buckendahl, 2003, p. 22).

With a low-complexity model, content specialists, such as subject-area teachers and higher education faculty, examine each assessment item and determine the degree to which an item matches a content standard, using a scale that may range from “no match at all” to “matches exactly” (Bhola, Impara, & Buckendahl, 2003). While this model forms a basis for other models, its limited scope (simple content match) makes it inadequate for developing assessment systems that might inform instruction (Impara, 2001).

#### **Moderate Complexity**

More complex models have been developed from the simple content match. Bhola, Impara, and Buckendahl (2003) refer to one of the simplest of the moderate models, in which content panelists examine the match



between content standards and assessment items in terms of both content match and cognitive complexity match.

Another model of greater complexity, which has been fieldtested in 11 states and 4 large urban districts (Council of Chief State School Officers, 2002) is the Survey of the Enacted Curriculum (SEC) model. The SEC model analyzes the alignment of standards, instruction, and assessments via use of a common content matrix that allows comparison across schools, districts, or states. As part of the alignment analysis, standards, assessments, instruction, or curriculum materials are categorized, using a common framework of content topics, according to cognitive demand (Council of Chief State School Officers, 2002). The model distinguishes between the intended curriculum, which is frequently identified in state content standards; the enacted curriculum, which is actually taught by teachers in their classrooms; and the assessed curriculum, or the content that is ultimately tested.

To measure the intended curriculum, a team of researchers codes each part of the standards documents with regard to content and cognitive demand. Data can then be analyzed using topographical maps that indicate what material is emphasized in the standards and what is not. Porter (2004) notes that measuring the enacted curriculum is more challenging than measuring content standards, since it requires teachers to either track the content they teach daily, or over a longer period of time. Teachers' self-assessment is translated into the topics-by-cognitive-demand matrix, and data are analyzed and displayed visually. The assessed curriculum is measured, using a common content language, by a team of experts that performs content analyses of assessments using the common language. Like the intended and enacted curriculum, the assessed curriculum is analyzed and displayed on a topographical map.

SEC's content maps and graphs portray visually the similarities and differences in content from standards to assessments to instruction (Council of Chief State School Officers, 2002). The maps allow for visual comparison of the relationship between the different components of instruction and assessment. With SEC, alignment between standards and assessment is perfect "if the proportions for the assessment match cell by cell the proportions for the standards" on the matrix (Porter, 2004, p. 13). Written interpretations of the content charts are also provided and statistics of alignment for each grade and subject are computed (Council of Chief State School Officers, 2002).

Another model of moderate complexity, the Council for Basic Education (CBE) alignment process, outlines the degree of match in content and performance level between standards and assessments. The model determines alignment based on four criteria: content, content balance, rigor, and item-response type (Council of Chief State School Officers, 2002). Content is understood in ways similar to other models; content balance considers the number and distribution of items used to assess the standards. Rigor refers to higher-order thinking skills that are similar to the cognitive complexity dimension of some models (Bhola, Impara, & Buckendahl, 2003). Finally, item-response type determines whether the item requires a supply-versus-select response type and the appropriateness of the response type in measuring the standard. In the CBE model, reviewers work in pairs to apply an evaluation rubric to judge alignment (Council of Chief State School Officers, 2002).

### High Complexity

Scholars disagree on the complexity of the model described by LaMarca, Redfield, Winter, Bailey, and Despriet (2000) that has multiple dimensions, with Impara categorizing it as a moderate- complexity model, while Bhola, Impara, and Buckendahl (2003) categorize it as a high- complexity model. LaMarca et al. describe a model for aligning state standards and assessment systems that is built around the requirements of Title 1 (LaMarca et al., 2000). Their model emphasizes five dimensions in its consideration of how well assessments measure content standards: content match, depth match, emphasis, performance match, and accessibility. Content match and depth match are similar to the content and cognitive complexity dimensions outlined in the moderately complex models (Bhola, Impara, & Buckendahl, 2003). The emphasis dimension analyzes whether an assessment sufficiently covers the target content domain in relation to the emphasis outlined in the standards. Performance match indicates whether assessment items reflect the types of student performance articulated in the standards and the methods through which teachers taught the content of the standards. Accessibility analyzes the degree to which the assessment includes items with a range of difficulty that enables students at varying levels of proficiency to reveal what they know on the assessment.

Aside from this model, the Achieve and Webb models are the two alignment models generally considered to have high complexity.

### ***The Achieve Model***

Achieve, Inc. has developed an alignment model that provides an in-depth analysis of the alignment of assessments to state standards (Council of Chief State School Officers, 2002), or the relationship between a state's standards and those of other states or nations (Case, Jorgensen, & Zucker, 2004) on the basis of four criteria. The Achieve alignment criteria are: construction of the test blueprint, content centrality, performance centrality, and challenge (Achieve, 2003).

Under Achieve's confirmation or construction of test blueprint criteria, reviewers use either a blueprint provided by the state, district, or test company, or one created by Achieve, to ensure that each assessment item corresponds to at least one state standard or objective (Achieve, 2003). The content centrality criterion examines the quality of the match between the content of the assessment item and the content of the standard or objective, using one of four categories, from "not aligned" to "clearly aligned." The performance centrality criterion examines the match between what students are asked to do (i.e., to "analyze" or to "identify"), using the four categories for the content centrality criterion.

Achieve's last criterion, the challenge criterion, examines several different aspects of alignment. The challenge criterion is applied to both individual items and a set of items that comprise a strand such as "Measurement." Reviewers determine whether items have appropriate or inappropriate sources of challenge and code the level of cognitive demand of the items. Achieve (2003) borrows from Webb's (2001) Levels for Determining Depth of Knowledge, and depth-of-knowledge levels are a central component of Webb's criteria for determining alignment. Achieve's challenge criterion also addresses the set of items, with reviewers comparing the overall demand of a set of items with the overall level of challenge of the standard. Reviewers consider whether the knowledge and skills of the standard receive the same emphasis on the assessment and whether the emphasis is appropriate, using both qualitative and quantitative measures (Achieve, 2003).

As with the Webb model, the Achieve model uses a team of teachers, curriculum specialists, and content experts to do the analysis (Achieve, 2003). In contrast to Webb's model, where reviewers participate in a consensus process to determine the levels of cognitive demand of state standards and then analyze the assessment items independently, Achieve reviewers conduct content analyses as a group (Rothman, Slattery, Vranek, & Resnick, 2002).

### ***The Webb Model***

Another operational alignment model, and one upon which the LaMarca model draws heavily, the Webb model outlines five general categories and criteria for alignment under those categories (Webb, 1997). Webb's broader categories of content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability act as a means to analyze agreement among expectations and assessments. The categories emphasize the relationship between multiple variables, listed "in an order first to consider content, then students, then instruction, and finally application to a system" (p. 14). Webb's alignment studies of various states (1999) have further refined that alignment process and the four criteria under the category (content focus) that receives the most attention when conducting an alignment study. Those four criteria—categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation—are considered in greater detail below.

#### **Categorical Concurrence**

Categorical concurrence indicates that the same or consistent categories of content appear in both standards and assessments. This criterion is judged by determining whether an assessment includes items measuring content from each state standard. If, for instance, a standard deals with reading comprehension, an assessment with a sufficient number of items dealing with reading comprehension would satisfy the categorical concurrence criterion. The number of items Webb determined would satisfy the criterion, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. Webb's decision is based on a procedure developed by Subkoviak (1988), and assumes that the six items when considered together produce a score having adequate reliability.

#### **Depth-of-Knowledge Consistency**

Depth-of-knowledge consistency indicates a minimum level of consistency between the cognitive demands of state standards and the cognitive demands of the assessment items. For consistency to exist between the assessment and the standard, as judged in Webb's analysis, at least 50% of the items corresponding to an objective must be at or above the level of knowledge of the objective. If a standard has between 40% and 50% of items at or above the depth-of-knowledge levels of the objectives, then the criterion is "weakly" met.

In comparing the cognitive demands of the standards and assessment items, Webb (1999) defines four levels of cognitive complexity: Level 1 (Recall), Level 2 (Skills and Concepts), Level 3 (Strategic Thinking), and Level 4 (Extended Thinking). As an example, we include the depth-of-knowledge level definitions for mathematics.

**Level 1 (Recall)** includes the recall of information such as a fact, definition, term, or a simple procedure, as well as performing a simple algorithm or applying a formula. That is, in mathematics, a one-step, well-defined, and straight algorithmic procedure should be included at this lowest level. Other key words that signify Level 1 include “identify,” “recall,” “recognize,” “use,” and “measure.” Verbs such as “describe” and “explain” could be classified at different levels, depending on what is to be described and explained.

**Level 2 (Skill/Concept)** includes the engagement of some mental processing beyond an habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps. Keywords that generally distinguish a Level 2 item include “classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data.” These actions imply more than one step. For example, to compare data requires first identifying characteristics of objects or phenomena and then grouping or ordering the objects. Some action verbs, such as “explain,” “describe,” or “interpret,” could be classified at different levels depending on the object of the action. For example, interpreting information from a simple graph, or reading information from the graph, also are at Level 2. Interpreting information from a complex graph that requires some decisions on what features of the graph need to be considered and how information from the graph can be aggregated is at Level 3. Level 2 activities are not limited only to number skills, but may involve visualization skills and probability skills. Other Level 2 activities include noticing or describing non-trivial patterns, explaining the purpose and use of experimental procedures; carrying out experimental procedures; making observations and collecting data; classifying, organizing, and comparing data; and organizing and displaying data in tables, graphs, and charts.

**Level 3 (Strategic Thinking)** requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is at Level 3. Activities that require students to make conjectures are also at this level. The cognitive demands at Level 3 are complex and abstract. The complexity does not result from the fact that there are multiple answers, a possibility for both Levels 1 and 2, but because the task requires more demanding reasoning. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be at Level 3.

Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and deciding which concepts to apply in order to solve a complex problem.

**Level 4 (Extended Thinking)** requires complex reasoning, planning, developing, and thinking, most likely over an extended period of time. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified at Level 2. However, if the student is to conduct a river study that requires taking into consideration a number of variables, this would be a Level 4. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas within the content area, or among content areas—and have to select one approach among many alternatives on how the problem should be solved, in order to be at this highest level. Level 4 activities include designing and conducting experiments and projects; developing and proving conjectures, making connections between a finding and related concepts and phenomena; combining and synthesizing ideas into new concepts; and critiquing experimental designs.

### Range-of-Knowledge Correspondence

Webb’s range-of-knowledge correspondence criterion is used to judge whether a comparable span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer the assessment items/activities (Webb, 1999). The criterion for correspondence between span of knowledge for a standard and an assessment considers the number of objectives within the standard with one related assessment item/activity. Fifty percent of the objectives for a standard have to have at least one related assessment item in order for the alignment on this criterion to be judged acceptable. Webb bases this level on the assumption that students’ knowledge should be tested on content from over half of the domain of knowledge for a standard (Webb, 1999). This assumes that each objective for a standard should be given equal weight. As with the other criteria, a state may alter the acceptable level on this criterion—for instance, making the necessary level more rigorous by requiring an

assessment to include items related to a greater number of the objectives. Range-of-knowledge correspondence is more difficult to attain if the content expectations are partitioned among a greater number of standards and a large number of objectives.

### **Balance of Representation**

The final criterion from Webb's content category of alignment is balance of representation. The range-of-knowledge criterion only considers the number of objectives that had corresponding assessment items; it does not take into consideration how the hits (or assessment items/activities) are distributed among the objectives. The balance-of-representation criterion is used to indicate the degree to which one objective is given more emphasis on the assessment than another. Webb (1999) identifies an index he uses to judge the distribution of assessment items that have at least one hit. The index is computed by considering the difference in the proportion of objectives and the proportion of hits assigned to the objective. An index value of 1 signifies perfect balance and is obtained if the hits (corresponding items) related to a standard are equally distributed among the objectives for the given standard. Index values that approach 0 signify that a large proportion of the hits are on only one or two of all of the objectives hit. Webb suggests that index values of .7 or higher indicate that items/activities are distributed among all of the objectives at least to some degree (e.g., every objective has at least two items) and uses this value as the acceptable level on this criterion.

### **Source of Challenge**

The source-of-challenge criterion is based on a similar idea in the Achieve alignment model. Source-of-challenge issues are identified for any assessment item where the major cognitive demand of the item is something other than the intended or targeted concept of skill. In other words, students may answer the item correctly without knowing the intended content, or they may answer the item incorrectly while knowing the intended content. This criterion is different from depth-of-knowledge consistency. A source-of-challenge item does not assess the appropriate content at an inappropriate level of cognitive complexity, but rather assesses irrelevant and unintentional content. One purpose of this criterion is to identify issues of cultural bias on an assessment.

### **Webb Model Process**

In the Webb model, groups of reviewers are formed to examine and code both the standards and assessments according to a common framework. Each group is comprised of 5 to 12 contents experts in one subject and grade range. Some of these experts are internal to the state or district performing the study, in order to provide insight about how the standards are used and interpreted. Some of the experts are external to the state or district, to give external validation and an outside perspective on student knowledge.

After learning about the alignment process, the group leaders train the reviewers on the depth-of-knowledge (DOK) level definitions for their appropriate content area. Then for each grade level being studied, the reviewers participate in a two-step alignment process. The first step is to achieve group consensus for the DOK values for each content objective in the appropriate grade-level standards. This process may take one to two hours for each grade in each content area. The second step is to individually code each item on the given assessment. Each reviewer in the group independently assigns each item a DOK value and selects up to three objectives in the standards that the item targets (usually one).

Reviewers enter these data into the Web Alignment Tool (WAT) ([www.wcer.wisc.edu/wat](http://www.wcer.wisc.edu/wat)), which then produces reports and tables with general alignment information, as well as specific information about how the alignment meets the criteria described above.

### ***Issues with Alignment Models***

The quality of the results of an alignment process will depend on both the model used and the extent to which important problems are identified and either solved, or clearly acknowledged to be limitations (Impara, 2001). Impara (2001) asserts that alignment studies using any of the models may have problems associated with the comprehensiveness of the alignment criteria, or with the methodology associated with using the criteria in an alignment study.

One issue with alignment criteria comprehensiveness emerges when the finest level of detail in a set of standards is still broader than the level of detail represented by the individual assessment items. In most alignment systems, an item may be considered to "target" an objective (goal, standard, benchmark, indicator, etc.), even if it only addresses part of the content in the objective (Impara, 2001). This kind of issue is addressed by the alignment criterion called structure of knowledge, discussed by Webb (1997). For an alignment system like Webb's, this kind of issue should be addressed when determining the form in which the

standards will be entered into the WAT to be coded. Poor matches between items and objectives should also be specifically noted by reviewers. The Achieve model addresses this issue by allowing different degrees of alignment between objectives and items. However, even in the Achieve model, multiple items, each of which only slightly targets a given objective, may in total completely target the objective, or may in total target only part of the objective.

Another issue with the comprehensiveness of the alignment criteria comes from the different cut scores and classifications of the assessments. For instance, Webb's Depth-of-Knowledge Consistency criterion is set at 50% because 50% is a common passing score for an assessment. With this as a passing score, a student passing an aligned assessment must answer at least one item at the appropriate DOK level. However, with this rationale, an assessment with a proficiency cut score of 30% or 70% should use different cutoff values for the Depth-of-Knowledge Consistency criterion as well. The Webb model addresses this kind of issue by simply allowing alignment analysts to change the cutoff values for the alignment criteria as long as they can justify this, based on the specific factors of the assessments and standards.

Methodology issues with alignment studies include reviewer training and assessment selection. Reviewer training is clearly an issue: if reviewers are not appropriately or adequately trained in the use of the given alignment model, their coding procedures will have limited validity. Assessment selection can also be troublesome, because assessments may change considerably from one year to the next or from one form to another even if they are based on the same blueprints.

A recent paper (Herman et al., 2005) addresses some of these methodology issues with data from a case study. An alignment study based on the Webb model was performed with 20 reviewers between a set of mathematics standards and an assessment in California. According to several statistical measures, the group of 20 adequately trained reviewers had good agreement for item coding with respect to DOK levels, targeted objectives, targeted standards, and item centrality (a measure of the degree to which the item matches the targeted content). On the other hand, various different subgroups of six reviewers displayed much lower reliability, especially with respect to targeted objectives. This finding is somewhat reinforced by the reliability study found earlier in this report. The evidence suggests that while it may be convenient in an alignment study to have reviewer groups of fewer than seven persons, this may compromise the reliability of the coding.

## Conclusion

---

Over the past several years, the Webb alignment system has shown itself to be one of the most important and innovative models for understanding the relationship between content standards and assessments. Before the development of the Web Alignment Tool, alignment analyses using high complexity models like the Webb model were time-consuming and tedious to perform. Now as a result of this grant, the WAT has made such alignment studies so quick and easy that over the past three years it has been employed with the standards and assessments of over 17 states and one separate nation.

This paper has summarized the WAT's development and refinement, described its capabilities, analyzed its reliability, and discussed its relationship to other alignment systems. This account raises further questions and leads to other avenues for research and development. Research into vertical alignment and the alignment of alternate assessments, conducted by Lauress Wise and Gerald Tindal respectively, has also been funded by the same grant. Although these models make use of the WAT, what are the limitations of using the WAT in these contexts? To what degree can the WAT and the Webb model answer questions about the alignment of assessments with performance standards (as opposed to content standards)? How can the WAT be expanded to address the alignment of curricula with standards and assessments? The growing recognition of the importance of alignment and the usefulness of the WAT indicate that these questions will be explored in the near future.

## References

- Achieve, Inc. (2003). *Measuring up: A report on education standards and assessments for Montgomery County*. Washington, DC: Author.
- Ananda, S. (2003a). *Rethinking issues of alignment under No Child Left Behind*. San Francisco: WestEd.
- Ananda, S. (2003b). *Achieving alignment*. *Leadership*, 33(1), 18-21.
- Baker, E. (2004). *Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform*. (CSE Report 645.) Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Bhola, D., Impara, J., & Buckendahl, C. (2003). *Aligning tests with states' content standards: Methods and issues*. *Educational Measurement: Issues and Practice*, 22(3), 21-29.
- Case, B., Jorgensen, M., & Zucker, S. (2004). *Alignment in educational assessment*. San Antonio: Harcourt Assessment.
- Council of Chief State School Officers. (2002). *Models for alignment analysis and assistance to states*. Paper prepared by the Council of Chief State School Officers. Washington, DC: Author.
- Cronbach, L. J., & Meehl, P. E. (1955). *Construct validity in psychological tests*. *Psychological Bulletin*, 52, 281-302.
- Elementary and Secondary School Act. Public Law 89-10 (April 11, 1965).
- Herman, J., Webb, N., & Zuniga, S. (2005). *Measurement issues in the alignment of standards and assessments: A case study*. (CSE Report 653.) Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- IDEA — Individuals with Disabilities Education Act, 20 U.S.C. § 1400 et seq., as amended by the Individuals with Disabilities Education Act Amendments of 1997, Pub.L. No. 105-17, 111 Stat. 37 (1997). <http://thomas.loc.gov/cgi-bin/query/C?c108:/temp/~c108Km0L17>
- Impara, J. (2001, April). *Alignment: One element of an assessment's instructional unity*. Paper presented at the 2001 annual meeting of the National Council on Measurement in Education, Seattle, WA.
- LaMarca, P., Redfield, D., Winter, P., Bailey, A., & Despriet, L. (2000). *State standards and state assessment systems: A guide to alignment*. Washington, DC: Council of Chief State School Officers.
- Lewis, D. M. & Haug, C.A. (2003). *Aligning policy and methodology to achieve consistent across-grade performance standards*. Monterey, CA: CTB/McGraw-Hill.
- Messick, S. (1975). *The standard problem: meaning and values in measurement and evaluation*. *American Psychologist*, 30, 955-66.
- Nasir, N. (2002). *Identity, goals, and learning: Mathematics in cultural practice*. *Mathematical thinking and learning*, 4(2 & 3), 213-247.
- National Research Council. (1999). *Testing, teaching, and learning: A guide for states and school districts*. Washington, DC: National Academies Press.
- No Child Left Behind Act of 2001. Public Law 107-110, 115 Stat. 1425 (2002).
- Porter, A. (2004). *Curriculum assessment*. In J. Green, G. Camilli, & P. Elmore (Eds.), *Complimentary methods for research in education*. Washington, DC: American Education Research Association.
- Rothman, R., Slattery, J., Vranek, J., & Resnick, L. (2002). *Benchmarking and alignment: CSE Technical Report 566*. Los Angeles: University of California, National Center for Research on Evaluation.

- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.
- Subkoviak, M. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery tests. *Journal of Educational Measurement*, 25(1), 47-55.
- Tindal, G. (2005). Alignment of alternate assessments using the Webb system. Eugene, OR: Council of Chief State School Officers, Technical Issues in Large Scale Assessment (TILSA).
- U. S. Congress. (1994). Goals 2000: Educate American Act. Washington, DC: Author.
- U. S. Department of Education. (1997). Final regulations for the amendments to the Individuals with Disabilities Education Act, P.L. 105-17. [http://www.cec.sped.org/law\\_res/doc/law/downloads/Fullregs97.doc](http://www.cec.sped.org/law_res/doc/law/downloads/Fullregs97.doc)
- 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.
- Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states. (Research Monograph No. 18.) Washington, DC: Council of Chief State School Officers.
- Webb, N. L. (2001). Levels for determining depth of knowledge. Council of Chief State School Officers, TILSA Alignment Study, Version 2.0, May 21–24, 2001.
- Wise, L., Harris, C., Sipes, D., Hoffman, G., & Ford, R. (2000). California High School Exit Examination (CAHSEE). Alexandria, VA: Human Resources Research Organization (HumRRO).
- Wise, L. (2005). Assessing vertical alignment. Alexandria, VA: Human Resources Research Organization (HumRRO).



## Alignment Report 2

---

### Alignment of Alternate Assessments Using the Webb System

Gerald Tindal  
University of Oregon

September 2005



## Table of Contents

<b>Introduction</b>	<b>35</b>
<b>Section I The Alignment Process: Definitions, Models, and Caveats</b>	<b>35</b>
Definitions of Alignment.....	35
Alignment Review Categories.....	36
Models of Alignment.....	37
Caveats to Consider in Alignment of Alternate Assessments.....	39
<b>Section II Alignment in the Webb System</b>	<b>42</b>
Steps in Conducting Alignment.....	44
<b>Section III Alignment Alternate Assessments Using Performance Tasks</b>	<b>45</b>
Alignment of Alternate Assessment with Performance Tasks .....	45
Alignment of Alternate Assessment with Portfolios.....	49
<b>Summary of State Two's Alignment</b>	<b>54</b>
<b>References</b>	<b>55</b>



# Alignment of Alternate Assessments Using the Webb System

---

## Introduction

---

The purpose of this paper is to delineate a procedure for alignment of assessments with standards in special education, using the Webb (2002) model as a base. The paper is divided into three sections that include (a) a summary of critical literature and an overview of the alignment process, highlighting some unique issues in special education; (b) the actual process used in applying the model to alternate assessments, including the specific steps to complete in conducting an alignment study; and (c) an illustration of the alignment process and outcomes using performance and portfolio assessments. The assumption is made that one alignment system is best for both general and special education and the process is applied in the same manner as Webb uses it so that states are not using different systems (one for general education and a different one for special education). However, some adjustments are needed, which are described below.

## Section I: The Alignment Process: Definitions, Models, and Caveats

---

Before adopting a specific model of alignment, it is important to consider the variables involved in the process and how they are considered in relation to the literature on and models of alignment. Even with this contextual understanding, however, the model cannot be blindly applied in special education without considering some very unique issues that arise with the use of alternate assessments.

### **Definitions of Alignment**

Alignment at its simplest level is close in definition to the term of overlap that had been present in the early work on content validity. Similar to content validity, alignment begs the question of “to what?” or “with what?” However, unlike the earlier research on content validity, alignment is standards-based, systemic in nature, and prospective in development. These three features have the potential for (a) bridging content validity to construct validity and (b) focusing our attention on inferences as the bedrock for validity. These features connect curriculum and instruction with standards and assessment. However, it appears that most writers make the assumption that standards, once enacted, have a reach into the classroom that provides parity and equality of

what is taught and how it is taught. If this is more or less true, then our accountability systems can be trusted; if this is not true, then confusion exists between the outcome and the inference.

Alignment to **standards** is the starting point. For example, consider the following three definitions.

Alignment refers to the degree of match between test content and the subject area content identified through state academic standards. Given the breadth and depth of typical state standards, it is highly unlikely that a single test can achieve a desirable degree of match. This fact provides part of the rationale for using multiple accountability measures and also points to the need to study the degree of match or alignment both at the test level and at the system level. Although some degree of match should be provided by each individual test, complementary multiple measures can provide the necessary degree of coverage for systems alignment. This is the greater accountability issue. (La Marca, 2001, p. 1)

LaMarca suggests that sound standards and assessment development activities are needed to create alignment.

More recently, Bhola, Impara, & Buckendahl (2003) stated that

“alignment can be defined as the degree of agreement between a state’s content standards for a specific subject area and the assessment(s) used to measure student achievement of these standards” (p. 21).

Finally, Webb (1997) defines alignment as

“the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do” (p. 4).

As intimated by Webb in the quotation above, a definitional feature of alignment (which was not present with content validity work from the 1980s) is a systemic focus. Notice the language in the second point below from the definition provided by La Marca, Redfield, Winter, Bailey, and Despriet (2000):

- 1) “Alignment is a match between two or more things. Webster’s New World College Dictionary defines align as ‘to bring into a straight line; to bring parts or components into a proper coordination; to bring into agreement, close cooperation.’ In an aligned system of standards and assessments, all components are coordinated so that the system works toward a single goal: educating students to reach high academic standards” (Hansche, 1998, p. 21).
- 2) Alignment refers to how well all elements in a system work together to guide instruction and student learning (Webb, 1997).
- 3) Alignment directly affects the degree to which valid and meaningful inferences about student learning can be made from assessment data (Long & Benson, 1998).

Finally, alignment is more prospective than retrospective. Research on content validity indicates that most tests provided broad surveys of several curricula that were analyzed post hoc for overlap, but these tests were not designed from the outset with specific planning around any particular curriculum. La Marca (2001) provides the best description of this feature with his suggestion that sound standards and assessment development activities create alignment when the following three conditions are present:

- (a) As standards are developed, assessment design needs to be considered (determining what and how to measure achievement);
- (b) items and tasks should be designed to measure specific objectives as outlined by academic standards; and
- (c) a post hoc review of alignment should be conducted following assessment development.

### **Alignment Review Categories**

La Marca, Redfield, Winter, Bailey, and Despriet (2000) articulate the following categories to consider in reviewing the alignment of assessments with standards.

**Content Match** When evaluating content match between standards and assessment, one should consider whether assessments are designed to match the content standards, whether all items and tasks

are related to the content standards, and whether the assessment fully covers the content standards.

**Depth Match** When evaluating depth match between standards and assessment, one should take into consideration whether both the assessment as a whole and items/tasks are at a level of difficulty matching that prescribed by content standards, and whether item/task specifications both indicate the depth at which knowledge should be measured and elicit responses reflecting the depth of knowledge it measures.

**Emphasis** Assessment items and tasks en masse should measure knowledge and skills representative of those in the content standards in order for the assessment and content standards to have emphasis match.

**Performance Match** When evaluating performance match, one should consider whether the assessment blueprint specifies how the entire range of performance descriptors will be measured by the assessment, whether item specifications are referenced to the levels of knowledge and skills in the performance descriptors, whether the assessment as a whole covers knowledge and skills at each defined performance level and each aspect of the performance descriptors is covered by one or more items/tasks.

**Accessibility** Accommodations and modifications should be available for students with disabilities or English language learners, groups of selected-response items should cover a variety of ways of expressing knowledge and skills related to the content standard(s), and the assessment should be free of irrelevant factors that are likely to interfere with students' opportunity to demonstrate knowledge/skills in order to have accessibility.

**Reporting** To evaluate reporting, one should take into consideration whether score reports clearly illustrate levels of student proficiency on content standards, whether reports contain information that can be used to make valid inferences and decisions and information about the standard error of measure regarding reported scores, and whether the reported information can be applied for the intended purpose(s) of the assessment

### **Models of Alignment**

In a review by Bhola, Impara, and Buckendahl (2003), two alignment systems are classified as moderately complex: (a) Surveys of Enacted Curriculum (SEC) (Council of Chief State School Officers, 2002) in which content and cognitive demand are considered, and (b) Council for Basic Education (CBE), which includes content, content balance, rigor, and item response type (Council of Chief State School Officers, 2002). Likewise, two alignment systems reflect a highly complex alignment model: Achieve (Resnick, Rothman, Slattery, & Vranek, 2003–2004) and Webb's system (2002).

#### **The SEC Model**

Alignment with Surveys of Enacted Curriculum (Council of Chief State School Officers, 2002) uses a two-dimensional (content topics and cognitive demand) content matrix; it yields alignment analyses of standards, assessments, and instruction and allows comparison across schools, districts, or states. Four reviewers code each assessment item or benchmark, which is systematically categorized according to content topics by cognitive demand (divided into five subject-appropriate categories) into the matrix. Likewise, surveys are conducted with teachers who use the same matrix to report on instruction or curriculum materials. Statistics of alignment for each grade and subject are computed, and content maps and graphs visually portray differences and similarities in content from instruction to standards to assessments.

#### **The CBE Model**

The Council for Basic Education alignment system (Council of Chief State School Officers, 2002) is designed to conduct alignment work, or train and supervise a team of educators to conduct the alignment. The process identifies test items or framework specifications that match benchmarks and records the degree of match in content and performance level. Reviewers work in pairs to apply an evaluation rubric and exemplars to determine degree of match, and subsequently make decisions. States receive a CBE report following the alignment process, which advises states as to the content areas and grade levels that need to be augmented for alignment with content standards, as well as steps that need to be taken for the test to be in compliance with ESEA/NCLB requirements.

### The Achieve Model

This alignment system (Resnick, Rothman, Slattery, & Vranek, 2003–2004) was developed in 2001 and is based on five criteria:

- 1) accuracy of the test blueprint (every item corresponds to at least one standard)
- 2) content centrality (the number of standards being assessed)
- 3) challenge (degree of cognitive complexity that considers both the source and level)
- 4) balance (weight of items and emphasis of standards)
- 5) range (representativeness of item sampling from the content domain of the standards)

The Achieve model provides a qualitative and quantitative analysis on the degree of alignment between assessment items and standards using a panel of content experts' judgments on five criteria:

- **content centrality**, a comparison of the content required by an assessment item to that of the related standard
- **performance centrality**, a comparison of the performance required by an assessment item to that of the related standard
- **challenge**, rated on a 1–4 scale of level of demand
- **balance**, the extent to which the content delineated in the standards receives the same emphasis on the related item set
- **range**, the proportion of objectives explicating a standard that are assessed by at least one item.

Achieve provides quantitative data on the test blueprint (important because blueprints are the basis for score reports), content and performance centrality, source of challenge, level of demand, and written commentary on overall patterns (including level of challenge and balance for each standard and for the test in entirety), as well as standards benchmarking (in which a state's standards are compared to exemplary state and international standards), augmentation analysis (in which a state's standards are compared to "off-the-shelf," norm-referenced tests, and suggestions are provided as to how a state can improve alignment and better conform to No Child Left Behind [NCLB]), professional development (to build a state's capacity to conduct its own alignment studies), and policy audits (to determine the effectiveness of a state's reform efforts).

### The Webb Model

This broader view of alignment focuses not only on content (which addresses categorical concurrence, range of knowledge, depth of knowledge, and balance of representation), but also on articulation across the grades, equity, pedagogical implications, and systems applicability. Webb's model (1997, 2002) combines qualitative expert judgments and quantified coding and analysis, yielding a set of statistics for each standard and grade on the degree of alignment between state content standards and state assessments. Trained reviewers individually identify the content standard objectives that match each assessment item, and the depth of knowledge required by each objective/benchmark of the content standards being analyzed; the reviewers use four levels: (a) recall, (b) skill/concept, (c) strategic thinking, and (d) extended thinking. Next, reviewers determine the objective/benchmark represented by each item or task on the state assessment and rate the depth of knowledge required for successful completion. Reviewers' ratings are entered into a spreadsheet and analyzed across reviewers, producing statistics on the following four categories (Wisconsin Center for Educational Research, 1999; Webb, 2002).

- 1) **Categorical Concurrence**, or the degree to which standards and assessments address the same content categories. This criterion is met if both documents display the same or consistent categories of content.
- 2) **Depth-of-Knowledge Consistency**, or the degree to which the depth or complexity of knowledge required by the standards and assessments are in agreement. If the assessment is as demanding conceptually as the expectations standards set for the students, this criterion is met. Depth-of-knowledge is judged at four levels: (a) recall of fact, information, or procedure; (b) skill in using information, conceptual knowledge, or procedures of two or more steps; (c) strategic thinking, reasoning, developing a plan or sequence of steps, complexity, more than one possible answer,



requiring less than 10 minutes to do; and, (d) extended thinking, requiring an investigation, time to think and process multiple conditions of the problem or task, and requiring more than 10 minutes to do non-routine manipulations.

- 3) **Range-of-Knowledge Correspondence**, or the degree to which the span of knowledge a standard expects of students matches that required to correctly answer the assessment items or activity.
- 4) **Balance of Representation**, or the extent to which assessment items are evenly dispersed across learning objectives within a standard.

In summary, the literature on alignment presents a considerable advancement over previous work on content validity. The components are far more sophisticated and the implications more extensive. No longer is the process carried out only in the context of research, but as part of large-scale assessments with the outcomes used to guide systems development. No longer is overlap sufficient; rather more complex and complete dimensions are considered, which may be a natural extension of a standards-based system with high stakes decisions. Therefore, in considering alignment in understanding multiple dimensions of an assessment system, it also is important to consider all components of it—not just the test used in general education but also that used in special education.

### **Caveats to Consider in Alignment of Alternate Assessments**

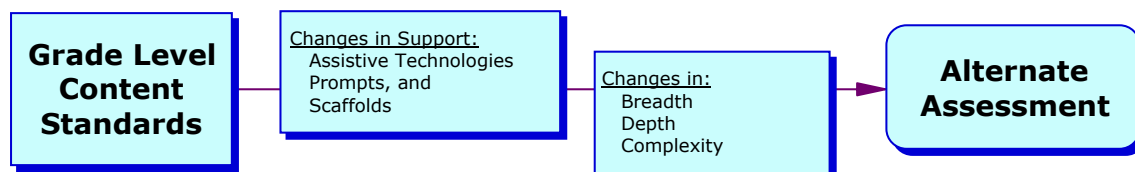
Three caveats need to be considered prior to beginning the alignment of alternate assessments with grade-level content standards. **First**, a number of different alternate assessments are possible, with the alternate assessment judged against alternate achievement standards as well as when judged against modified achievement standards. **Second**, considerable diversity exists in the approaches used with the alternate assessments that employ not only selected responses (SR), but also constructed responses (CR) and often both, and thereby require careful consideration for defining an “item.” **Third**, because some alternate assessments employ observations and collections of evidence over time (portfolios), it is not possible to align anything until the assessment is completed, thereby requiring careful sampling of students.

#### **Different types of alternate assessments**

The Executive Summary of *Including Students with Disabilities in Large-Scale Assessments* (Technical Work Group, 2005) provides a summary of the incredible pace of change in the laws and regulations that have been enacted. In this process, changes have been made in the participation options with five possibilities now available: (a) the general education assessment, (b) the general education assessment with accommodations, (c) an alternate assessment judged against grade-level achievement standards, (d) an alternate assessment judged against modified achievement standards, and (e) an alternate assessment judged against alternate achievement standards. Although the purpose of this document is to address the alignment of assessments with grade-level content standards, it is not possible to conduct such an assessment without considering these various options. Given the examples in the current document, alignment is confined to the alternate assessment judged against grade-level content standards for two reasons: (a) the notice of proposed rule making is not yet available [at the time of submitting this report in December 2005] to clarify the requirement of an alternate assessment judged against modified achievement standards, and (b) no examples from states are currently available for illustrating the process.

Suffice it to say at this point, however, that the distinctions between these options rest on a validity argument in which claims are made about proficiency and then evidence is provided to support the claim. The Technical Work Group (2005) has defined three levels of inference that can be made in understanding proficiency of achievement on grade-level content standards: Are the expectations (a) comparable to inferences made from the general education assessment; (b) constrained because of changes made in support or breadth, depth, and complexity; or (c) stipulated to only the format and application actually used? These options are relevant for the alignment process in that the assessment items need to take into account the changes made (either support or breadth, depth, and/or complexity). In essence, though the alignment (between standards and tests) remains the same (as used by Webb with the general education test), the judgments need to be moderated or mediated relative to the changes made in the process of developing the alternate assessments. The critical question then may be less about the alignment per se than the changes made and their impact on alignment. Some changes may facilitate alignment, while others detract from it.

Figure 1. Moderating changes to be considered in the alignment process



The following definitions of these changes are used in the white paper (Technical Work Group, 2005).

**Assistive technologies** – Any encoding or decoding device (electronic, digital, manual, etc.) that allows information to be presented or responses to be made while interacting with test directions, items, or tasks.

**Prompts** – Any form of verbal, non-verbal, or physical cue to structure, pace, or signal a response to be made by the student. Examples include verbalisms like, “continue,” “next,” “now what,” or reminders of each step; physical guidance is an example of a prompt.

**Scaffolds** – Any type of structural assistance introduced to organize information or guide responses embedded in the presentation of the item or task. Examples include the addition of highlights, underlines, outlines, crib sheets, or other information to “essentialize” the task or response.

**Breadth** – The number of standards and objectives being addressed in the assessment.

**Depth** – The type of knowledge form implied in the standard and response demand, or the type of intellectual operation required to respond. Examples include various knowledge forms (concepts, principles, procedures) and intellectual operations needed to solve the problem or answer the question (e.g., make predictions, provide explanations, give illustrations, consider reasons and use criteria to make judgments).

**Complexity** – The confluence of breadth and depth with the additional component of requisite skills implied in either the standards or the assessments (e.g., more complex standards and assessments address requisite skills that are more advanced). Examples include simple problems with requisite skills considered in a task analysis that are vertically aligned with the problem.

Because these changes are substantive (and in essence change the kind of inference that can be made about the grade-level content standard), the alignment process needs to consider them concurrently while reviewing items and standards. All of these changes can be considered as a source of challenge either on the basis of Webb’s notion of construct irrelevant variance (2002), or in the context of difficulty as noted by Resnick et al. (2003–2004).

### Definition of an item

In Webb’s model, alignment is very specific, with no sweeping judgments of subtests or tests being made; rather, items are individually evaluated. Alternate assessments, however, are very diverse (much more so than on the traditional large-scale assessment using multiple-choice items), with a number of different approaches currently available (Quenemoen, Thompson, & Thurlow, 2003). Four different approaches to assessment are depicted in Chapter 3 of *Including Students with Disabilities in Large-Scale Testing*:

- teacher judgments (based on reflections and observations using a rating scale or checklist)
- portfolios
- performance tasks
- performance collections

Because of such diversity in approach, the definition of an item needs to be carefully operationalized. In this paper, the term **behavioral events** is a term used as a proxy for an item. Behavioral events have four attributes: (a) they reflect routines (with a beginning/middle/end), (b) they are captured in one session (e.g., sitting/setting), (c) they comprise a skill or multiple skills, and (d) they contain one or more items that may incorporate observations of students in different settings (and therefore use rating scales and checklists), collections of work samples in portfolios, performance tasks, and collections of performance.

This term is introduced so that the alignment process does not become too granular on the one hand or too vague on the other. If every response in an assessment system is task-analyzed and placed into the alignment process, the process would be too cumbersome. However, without some specificity applied to the various

behaviors comprising the assessment, alignment outcomes would not be very informing. In this adaptation for special education, the number/type of behavioral events should be considered in terms of fair representation of the domain from the grade-level content standard. With brief tasks, more events can be counted; with fewer (but longer) tasks, the standard may still be represented but fewer events can be counted. The critical effect is the reliability of this judgment. In general education testing, the number of items influences the level of reliability; with a greater number of items, it is easier to attain higher levels of reliability. In the examples below, an assessment system is described in terms of the critical dimensions requiring alignment in reading, using each of the three assessment formats.

**An example of an observation-based rating:** The student is observed while being read a story and then asked questions about the story. While the teacher is reading, target words (high frequency) are written on the board, and the student is asked to read them and tell what they mean. The observation schedule targets the independence of reading and description of word meaning. Raters are given a scale of assistance used by the student in saying the word and another for repeating the meaning of the word. Three behavioral events are present: (a) a picture of the student reading, (b) notes from the teacher's questions about the story, and (c) the word list with correct and incorrect words noted. In this observation, the tasks may be aligned with standards addressing reading and vocabulary.

**An example of a portfolio entry:** The student completes a mind map of images after having a story read, printing some important words in a graph, and relating them to each other and the images. This work sample is supplemented by a drawing the student has created about the story and is scored for the connection to the content of the story (not neatness or artistic skill). Both samples are scored for the relatedness of the words on the mind map to the story and the representations of the pictures to the words. Three behavioral events are present: (a) a mind map, (b) a picture of the student describing the mind map, and (c) the illustration. These critical tasks may be aligned with standards for interpretive comprehension.

**An example of a performance task or collection:** The student is asked to read from a list of words and then from a passage and asked to relate the gist of the story from the passage. The student's performance is scored for number of words (on the list) read correctly and number of words read per minute in the passage; comprehension is measured by counting the number of story-specific words used in the retelling. Partial credit is given for responses that are essentially but not completely correct (e.g., sounding out syllables or blends). Three behavioral events are present with (a) the word list, (b) the passage, and (c) the retell. The tasks may be aligned with the standards of fluency and story meaning as part of literal comprehension.

### Sampling plans for student selection

Performance assessments are most similar to general education in their use of structured (standardized) tasks, while observations, portfolios, and collections of performances may not have explicit (pre-specified or standardized) tasks defined a priori. Therefore, the alignment process may have to be conducted after the assessment is completed to reflect various student behaviors. If this is the case, a sampling plan (of students) is required to ensure that a representative range of tasks is selected. To the degree that the alternate assessment is highly specified (e.g., has prescribed directions about what to document), it may begin to look much like a performance assessment in which target tasks are required for administration in controlled ways. However, if such prescriptive direction is not provided (in the sampling of standards and manner of documentation), then the assessment cannot be aligned until all student performance is appropriately documented. Therefore, it is highly desirable to specify in advance the components of the portfolio (and possibly observations) with descriptors that are aligned with the standards. These descriptors should consider what tasks to include and how to develop and format them, as well as how to administer and score them. Without such specificity, the actual sample of student work needs to be collected first and then analyzed for these components.

However, if such specificity is lacking, then a sampling plan strategy from survey research is needed to continue the alignment process for portfolios (and possibly observations). The sampling plan is best initiated by articulating the important variables to document the students who are participating. Some of the most obvious variables may be the geographic regions of the state, the size of the school (district or program), the disability of the student, and the type of program (e.g., regional or residential). The actual number of measures at each grade level to be aligned may need to be considered in sampling students. If students' entries are idiosyncratic in their focus on standards, the number of students to be sampled may need to be larger. Finally, generalizability may need to be considered in sampling students: How much variation is present in the sample? If considerable variation exists in the assessments, more students need to be sampled. If the variation among student assessments is slight, fewer students need to be sampled. When the data are stable, the alignment variables can be aggregated over students to obtain a composite picture of the system. This kind of matrix sample allows a report to be generated at the systems level, but not at the individual student level.

## Section II: Alignment in the Webb System

With the four areas addressed in applying the Webb model to special education, the actual alignment process can now begin. In this model, five areas are addressed: (a) categorical concurrence, (b) depth of knowledge, (c) range of knowledge, (d) balance of representation, and (e) source of challenge.

**Categorical Concurrence** The focus is on whether the standards and the assessments address the same content categories. In Webb's system, a requirement is made that the assessment needs to contain at least six items from a standard, the rationale being related to the need for making reliable judgments for ascertaining mastery. In considering the categorical concurrence of behavioral events, however, it may be as important to consider the amount of behavior reflected in each event (in which case the sheer number of events is somewhat less relevant than the type of behavior). The goal is to achieve a fair count (percentage) of standards that have stable representation in the assessment at the level in which scores are reported. In whatever manner behavioral events are defined within any assessment format, the key issue is whether they address a standard.

**Depth of Knowledge (DOK)** The focus in depth of knowledge is on the cognitive demand required by the standard and the assessment with four (DOK) levels possible. If an assessment item fits more than one objective, it needs to be coded as primary or secondary. According to Webb, at least 50% of the target items need to be at or above the level required by the objective. Table 1 presents the four levels of the Webb system and the matching levels of an adaptation of that system into an alternate assessment. The key modification is the use of language that is more sensitive to the content of alternate assessments.

Table 1

Depth-of-Knowledge (DOK) Levels for Webb Items and for Alternate Assessment Tasks

Level	Webb Description	Alternate Assessment Description
1	Recall and reproduction: Recall and recognition of a fact, information, or procedure	A 'behavior event' with 1:1 correspondence completed in single context.
2	Skill and concept: Use information or conceptual knowledge with 2 more steps	A 'behavioral event' with more than 1:1 correspondence in more than one context with correct or incorrect responses.
3	Strategic thinking: Requires reasoning, developing a plan or a sequence of steps, some complexity, more than one possible answer (non-routine problem-solving)	A multiple step 'behavioral event' executed in more than one context with more than 1:1 correspondence and with partial correct scoring of responses.
4	Extended thinking: Requires an investigation, time to think and process multiple conditions of the problem (e.g. completing a project, including how to design and execute it.	A multiple step 'behavioral event' executed as an approach (of many) to completing a task that occurs in multiple settings.

Many alternate assessments are imbued with "independence of functioning" scales or "generalization," constructs that are embedded in the assessment. To the degree that the assessment targets multiple environments for a task to be completed or a behavioral event to be displayed, the depth of knowledge may be influenced. For example, a low DOK level may be reflected if the behavior is to be engaged using a paper/pencil task in the classroom (name a street sign); however, if this same task is assessed in the community, in which the student must behave accordingly (stop at a red light and in the presence of cars), the DOK level may be quite high. Therefore, depth of knowledge is evaluated in the context of the environment in which the student is assessed and the tasks that are presented (or behavioral events that are documented). Furthermore, in targeting the performance of the student (whether using Webb's system or this adaptation of it), alignment is not related to the correctness of performance, but the manner in which behavior is scored. Given the same range of distracters (as noted above), correct and incorrect (dichotomous format) responses represent lower DOK levels than responses that can reflect partial correctness.

### Range of Knowledge

How well represented is each standard in the alternate assessment? This category refers to the breadth or span of knowledge required by the assessment and matched to standards with at least 50 percent of the objectives needing at least one related item (as required by Webb, 2002). Before proceeding to this analysis, the

methodology of assessment needs to be considered; a separate analysis is needed for each methodology (performances or portfolios-observations). Range of Knowledge can be calculated from extant data (the number of objectives with items divided by the total number of objectives in the standard) and is reflected as a percentage. This index is averaged across standards.

### Balance of Representation

In Webb's (2002) system, this index reflects the degree to which more than one objective on an assessment is given more or less emphasis. A standard hit is a standard with a corresponding item (or behavioral event). This index is defined as the proportion of objectives MINUS the proportion of hits assigned to objectives. To adapt this dimension to alternate assessments, the essential definition is considered as degree of emphasis. With many alternate assessments, however, the number of items is typically low (e.g., only a few portfolio entries or observations are present, and they are applied to several standards). Balance of representation in the alignment system, therefore, reflects the proportion of targeted skills with multiple representations on the standards. This multiplicity usually comes about through different items obtained as unique tasks, performances, work samples, or observations.

In Table 2, an example of balance is depicted for a standard with five objectives and 17 instances in which the objectives were met. In the first column are standard objectives and in the second column the number of matches between an item on the alternate assessment and these objectives. Calculation of balance is in the last two columns.

Table 2

Example for Balance of Representation (Calculated only on Objectives with Hits)

Standard or Objective	# Items "hit" per objective	$1/(O) - I(k)/(H)$	ABS
1.	3 hits	$1/5 - 3/17 =$	0.02
2.	4 hits	$1/5 - 4/17 =$	$ - .04  = 0.04$
3.	3 hits	$1/5 - 3/17 =$	0.02
4.	4 hits	$1/5 - 4/17 =$	$ - .04  = 0.04$
5.	3 hits	$1/5 - 3/17 =$	0.02
	O = 5 H = 17		$\Sigma = 0.14$ $0.14/2 = 0.07$ $1 - 0.07 = 0.93$
			Balance Index = 0.93

$O$  = Total number of objectives hit for the standard  
 $I(k)$  = Number of items hit corresponding to objective (k)  
 $H$  = Total number of items hit for the standard  
 $Balance\ Index = 1 - (\Sigma [1/(O) - I(k)/(H)]) / 2$

### Source of challenge

The focus is on construct-irrelevant variance other than the targeted skill, concept, or applications, causing the student to answer incorrectly for the wrong reasons. Source is considered as construct-relevant variance to identify false positives—students get right answer for wrong reason versus false negatives—students get the wrong answer but have the relevant skill or knowledge. In analyzing alternate assessments, the source of challenge cannot be ignored and should be considered in the alignment process, particularly with respect to the design of the assessments. It is particularly important in thinking about changes in support or breadth, depth, and complexity as mediators in the development of appropriate alternate assessments.

In summary, before beginning the process, grade-level content standards need to be identified. Then, the alignment process is completed in the following steps. First, the type of alternate assessment needs to be documented. Second, a behavioral event needs to be defined. Note: For all formats of assessment (performance, portfolio or observation), "items" refer to behavioral events that result in a product that could include a photo entered into a portfolio, a worksheet completed by the student, or a set of items from a performance task. Third, depending on the format, a decision can then be made about sampling students (if the alternate assessment has no descriptors of behavioral events). After completing these three steps, the alignment process can begin using Webb's (2002) five variables: (a) categorical concurrence, (b) depth of knowledge, (c) range of knowledge, (d) balance of representation, and (e) source of challenge. In Table 3, these variables are listed with a brief summary, as applied to each assessment approach.

Table 3

## Cross-Tabulation of Webb Alignment Variables with Assessment Approach

Dimension to Evaluate	Observation Rating or Checklist	Portfolio Work Sample or Behavioral Event	Performance Assessment Rating or Count
<b>Assessment sampling plan</b> (test blueprint for the assessment).	The sampling plan is the setting in which observations are to take place and the types of students who are being observed.	The sampling plan is the type of documents present in the portfolio and the types of students who are being selected for portfolio review.	With performance assessments, the sampling plan may be implicit in the tasks or explicit in the manner in which items are constructed.
<b>Categorical Concurrence</b> (a) Are the behaviors in the assessment access or target skills? (b) Do the target skills match standards and objectives?	For each standard, note the prevalence of objectives with observations of target skills in the environment.	For each standard, note the prevalence of objectives with work samples as target skills.	For each standard, note the prevalence of objectives with tasks as target skills as displayed for each task.
<b>Depth of Knowledge</b> First, determine the depth of knowledge for the standard or objective. Second, analyze the depth of knowledge for the alternate assessments.	Rate the standard and the notes or pictures on a 4-point scale of DOK given the environment in which behavior is observed.	Rate the standard and the work sample products on a 4-point scale of DOK using only the evidence in the portfolio.	Rate the standard and the behavioral samples on a 4-point scale of DOK.
<b>Range of Knowledge</b> Ascertain matched standard objectives and one target skill.	Proportion of standard objectives with targeted skills in the notes and pictures.	Proportion of standard objectives with targeted work samples.	Proportion of standard objectives with targeted performance tasks.
<b>Balance of Representation</b> Ascertain matched standard objectives with at least one target skill.	Proportion of standard objectives with multiple (and varied) behavioral events providing stable inferences from the observations.	Proportion of standard objectives with multiple (and varied) behavioral events providing stable inferences from the judgments.	Proportion of standard objectives with multiple task scores providing stable inferences from the (sub)totals.

**Steps in Conducting Alignment**

To complete the alignment process, the following steps need to be followed, regardless of the type of assessment being used (performance, portfolio, or observation).

**Step 1**

List the appropriate standards and objectives on a spreadsheet in the first column.

**Step 2**

Note the format of assessment and develop a (student) sampling plan.

- For performance assessments, fill subsequent columns with task labels. The cell that defines the intersection of each standard and task can be filled in with the number of items for that task. The total number of cells with numbers equals the number of tasks (to be considered hits in Step 5 below). See examples of matrices in reading and mathematics/science for states later in this paper.
- For portfolio or observation alternate assessments, mark frequency counts of behavioral events appearing in the folder. These counts should appear in subsequent columns for each standard and objective. A behavioral event is any entry that was completed within a sitting/setting. For example, a picture of the student performing a task qualifies as a behavioral event; completing a worksheet also serves as a behavioral event.

**Step 3**

Count the number of standards in which a behavioral event (for an alternate assessment task) appears as a performance task or is part of a portfolio or observation. Calculate the percentage by dividing the total number of standards with behavioral events into the total number of standards. The resulting number reflects the degree of categorical concurrence.

**Step 4**

Evaluate each standard having an associated alternate assessment task or behavioral event for the depth of knowledge. These evaluations are listed in the table and can be summarized as percentages: (a) matching, (b) alternate assessments > standards, or (c) standards > alternate assessments.

**Step 5**

Calculate the range of knowledge by counting within each standard the number of objectives having associated behavioral events or alternate assessment tasks divided by the total number of objectives in that standard. This percentage is computed within each standard and can be averaged across the standards.

**Step 6**

Calculate the balance of representation for each standard in which objectives have associated alternate assessment tasks or behavioral events (referred to as a hit). The easiest system for calculating balance is to use a spreadsheet and in successive columns (one for each standard) place sufficient rows to equal the number of objectives in that standard. With each objective, calculate the number of tasks in the assessment (those standards with assessment tasks then represent a hit); then the formula for that objective is  $(1/\# \text{ hits}) - (\# \text{ tasks for that objective} / \# \text{ total tasks})$ . The absolute value is taken of these values summed, which is then divided by 2 and subtracted from 1. The formula described above is used to ascertain that for this objective, the balance (.78) is sufficient, using definitions articulated by Webb (2002).  $\text{Balance Index} = 1 - (\sum |1/(O) - I(k)/(H)|) / 2$ .

**Step 7**

Report the results in a table with both the raw results in an appendix and a summary for critical stakeholders. The alignment system described in this paper can be used with the wide variety of alternate assessment formats that exist. Multiple examples of how to use the alignment system are now presented.

## Section III: Alignment Alternate Assessments Using Performance Tasks

In State One, the department of education has determined that the alternate achievement standards are based on pre-requisite skills (in this example) to align with both the general education standards and the alternate assessments. In this example, only Standards 4, 7, and 8 are explicated.

### *Alignment of Alternate Assessment with Performance Tasks*

**Step 1**

Assemble appropriate standards. Table 4 lists some of State One's essential reading standards for comprehension, demonstrating general understanding of informational text, and developing an interpretation.



Table 4

## Reading Comprehension: Essential Skills in Reading Standards

## Reading Comprehension

**Standard 4: Listening to and Reading Informational and Narrative Text**

Listen to, read, and understand a wide variety of grade-level informational and narrative (story) text including children's magazines and newspapers, dictionaries, other reference materials, online information, classic and contemporary literature, and poetry.

Demonstrate literal and inferential listening comprehension of more complex text through discussions.

Monitor own reading and self-correct when an incorrectly identified word does not fit with cues provided by the letters or context.

Notice when difficulties are encountered in understanding text.

**Standard 5: Vocabulary**

Understand, learn, and use new vocabulary that is introduced and taught directly through orally read stories and informational text as well as student-read stories and informational text.

Develop vocabulary by listening and discussing both familiar and conceptually challenging selections read aloud.

Classify categories of words (e.g., concrete collections of animals, foods, toys).

Use context to understand word and sentence meanings.

**Standard 6: Reading to Perform a Task**

Read written directions, signs, captions, warning labels, and informational books.

Locate the title, name of author, name of illustrator, and table of contents.

Alphabetize a list of words by the first letter.

Read and understand simple one-step written instructions.

Obtain information from print illustrations.

Identify text that uses sequence or other logical order (explain how informational text is different from a story).

**Standard 7: Informational Text—Demonstrate General Understanding**

Describe new information gained from text in own words.

Answer simple written comprehension questions based on material read.

**Standard 8: Informational Text—Develop an Interpretation**

Make connections and discuss prior knowledge of topics in informational texts.

Discuss how, why, and what-if questions in sharing informational texts.

**Step 2**

Assessment tasks are described and listed in Table 5: Note the format of assessment being used (and develop a student sampling plan if necessary for portfolios and observations). Students are expected to complete these tasks as they participate in the alternate reading performance assessment. In aligning these performance tasks with the state standards, a comparison is made between the 10 standards and these 6 tasks. For most of these standards, more than one objective is listed.

Table 5

## Reading Comprehension: Essential Skills in Alternate Assessment

**Read Passages** – Read aloud passages with 250 words.

**Comprehend Oral Text (Listening)** – Answer 6 questions from one of three stories read to the student.

**Comprehend Printed Text** – Answer 6 questions from one of three stories read by the student.

In Table 6, the standards are cross-listed with an alternate assessment. The standards are listed down the first column and the tasks used in the alternate assessment are listed across all subsequent columns. The number



inside each cell depicts the number of items within each task. The three standards are addressed with tasks (Listening to and Reading Informational and Narrative Text has three passages to be read to the student and to be read by the student; Informational Text—Demonstrate General Understanding has six questions from each mode of interaction. The same is true of Informational Text—Develop an Interpretation.

Table 6

**Cross-Map of Standards (first column) and Alternate Assessments (remaining columns) Identifying Number of Items within Tasks**

READING COMPREHENSION	EXTENDED ASSESSMENT READING		
<b>Standard 4: Listening to and Reading Informational and Narrative Text</b>	Read Passages	Comprehend Oral Text	Comprehend Printed Text
Listen to, read, and understand a wide variety of grade-level informational and narrative (story) text including children's magazines and newspapers, dictionaries, other reference materials, online information, classic and contemporary literature, and poetry.	3	3	3
Demonstrate literal and inferential listening comprehension of more complex text through discussions.		3	3
Monitor own reading and self-correct when an incorrectly identified word does not fit with cues provided by the letters in the word or the context surrounding the word.	3		
Notice when difficulties are encountered in understanding text.	3		
<b>Standard 7: Informational Text—Demonstrate General Understanding</b>			
Describe new information gained from text in own words.		6	6
Answer simple written comprehension questions based on material read.		6	6
<b>Standard 8: Informational Text—Develop an Interpretation</b>			
Make connections and discuss prior knowledge of topics in informational texts.		6	6
Discuss how, why, and what-if questions in sharing informational texts.		6	6

### Step 3

Categorical concurrence addresses whether the same content exists in the assessments as in the standards. In this example, every standard has at least one task in the assessment. Overall, for the entire reading assessment, only Standard 12 (not shown) fails to have any tasks represented in the assessment: Literary Text—Examine Content and Structure (by distinguishing fantasy from realistic text). Therefore, 88 percent (11 out of 12) of the standards have associated assessments. Some objectives within the standards depicted are missing assessment tasks, but each standard is represented by an assessment task. A lack of correspondence, however, is not part of categorical concurrence, but enters into the alignment process when addressing range of knowledge and balance of representation.

### Step 4

Depth of knowledge refers to levels of cognitive demands reflected in the standards and assessments.

- 1) A behavioral event with 1:1 correspondence completed in single context.
- 2) A behavioral event with more than 1:1 correspondence in more than one context with correct or incorrect responses.
- 3) 3 –correspondence and with partial correct scoring of responses.
- 4) A multiple step behavioral event executed as an approach (of many) to completing a task that occurs in multiple settings.

In Table 7, each of the standards-objectives has been assigned a value for DOK levels; where behavioral events in the alternate assessment are present, the task also has been assigned a value. In nearly all standards-objectives and associated alternate assessment performance tasks in this particular state, the DOK level is rated the same; for most of them, the rating is 3 or 4 with very few 2s.

**Table 7**

**Comprehension: Categorical Concurrence and Depth of Knowledge (Range, and Balance)**

READINGS COMPREHENSION					
Standard 4: Listening to and Reading Informational and Narrative Text	DOK STD AA		No. AA Tasks	Balance Index [BI]	
Listen to, read, and understand a wide variety of grade-level informational and narrative (story) text including children's magazines and newspapers, dictionaries, other reference materials, online information, classic and contemporary literature, and poetry.	3	3	3	1/4-3/7	
Demonstrate literal and inferential listening comprehension of more complex text through discussions.	3	3	2	1/4-2/7	
Monitor own reading and self-correct when an incorrectly identified word does not fit with cues provided by the letters in the word or the context surrounding the word.	4	4	1	1/4-1/7	
Notice when difficulties are encountered in understanding text.	4	4	1	1/4-1/7	
The AA addresses all 4 of the 4 objectives in Standard 4 (100%).			O=4 H=7	$\Sigma = .43$ $43/2 = .21$ $1 - .21 = .79$ BI = .79	
Standard 7: Informational Text—Demonstrate General Understanding	DOK STD AA		No. AA Tasks	Balance Index [BI]	
Describe new information gained from text in own words.	3	3	2	1/2-2/4	
Answer simple written comprehension questions based on material read.	3	3	2	1/2-2/4	
The AA addresses both of the 2 objectives in Standard 7 (100%).			O=2 H=4	$\Sigma = 0$ $0/2 = 0$ $1 - 0 = 1$ BI = 1	
Standard 8: Informational Text—Develop an Interpretation	DOK STD AA		No. AA Tasks	Balance Index [BI]	
Make connections and discuss prior knowledge of topics in informational texts.	3	3	2	1/2-2/4	
Discuss how, why, and what-if questions in sharing informational texts.	3	3	2	1/2-2/4	
The AA addresses both of the 2 objectives in Standard 8 (100%).			O=2 H=4	$\Sigma = 0$ $0/2 = 0$ $1 - 0 = 1$ BI = 1	

### Step 5

Range of knowledge is summarized using this state's alternate assessment in reading (comprehension). In "No. AA Tasks" (above in Table 7), the number of tasks is presented addressing each of the standards/objectives. For example, for the first standard (Listening to and Reading Information and Narrative Text), assessment tasks appear within all four objectives. This number and percentage is summarized in Table 8 (below) for all of the standards (4 – 8) and objectives (n = 8) in this state.

Table 8

## Reading Comprehension: Range of Knowledge

Standard #	Standard Content	No. of Objectives with AA	Percent
Reading Comprehension			
4	Listening to and Reading Information and Narrative Text	4 of 4	100%
5	Vocabulary	1 of 4	25%
6	Reading to Perform a Task	1 of 6	17%
7	Informational Text – Demonstrate General Understanding	2 of 2	100%
8	Informational Text – Develop an Interpretation	2 of 2	100%

In range of knowledge, this state's alternate assessment is quite strong (except in vocabulary and reading to perform a task, which are not shown).

## Step 6

Balance (of representation) is summarized in Table 9, using State One's alternate assessment. As mentioned earlier, a formula is used to calculate the Balance Index:  $1 - (\sum |1/(O) - I(k)/(H)|) / 2$ .

Table 9

## Reading Comprehension: Balance of Representation

Standard #	Standard Content	No. of Objectives with Hits	No. Tasks	Balance
Reading Comprehension				
4	Listening to and Reading Information and Narrative Text	4	7	0.79
5	Vocabulary	1	1	1.00
6	Reading to Perform a Task	1	2	1.00
7	Informational Text-Demonstrate General Understanding	2	4	1.00
8	Informational Text-Develop an Interpretation	2	4	1.00

In summary, this example of an alternate reading (using essential skills) assessment reflects consistently strong balance of representation. In great part, the balance is high because of the few number of tasks in each of the standard-objective areas.

In this state, the alignment of alternate assessment in reading with state standards is quite adequate on all four dimensions (categorical concurrence, depth-of-knowledge consistency, range of knowledge, and balance of representation). All of the standards (except one) have assessment tasks reflecting 88% categorical concurrence. Depth-of-knowledge consistency is very well matched and reflects a range from 3 (A multiple step "behavioral event" executed in more than one context with more than 1:1 correspondence and with partial correct scoring of responses) to 4 (A multiple step "behavioral event" executed as one approach—of many—to completing a task that occurs in multiple settings). The range of knowledge (a more granular look within the standards) is generally quite high across the standards, with most of the objectives within standards having tasks associated with them. Finally, balance of representation is adequate, in part because of the limited number of tasks within each standard and assessment.

### Alignment of Alternate Assessment with Portfolios

In State Two, alternate achievement standards have been developed for every standard appearing in general education. In this state, a portfolio format was used for the alternate assessment.

### Step 1

Assemble appropriate standards. These standards were written in a more general and broader manner so that students with significant disabilities can participate in learning many of the same skills and knowledge as their peers. They reflect a loose essence of skill at each grade level. (See Table 10.)

**Table 10**

**State Two Standards**

<b>Mathematics, Science, and Technology</b>
Standard 1 -- Analysis, Inquiry, and Design
Standard 2 -- Information Systems
Standard 3 -- Mathematics
Standard 4 -- Science
Standard 5 -- Technology
Standard 6 -- Interconnectedness: Common Themes

### Step 2

Note the format of assessment being used (and develop a student sampling plan, if necessary, for portfolios and observations). In this state, a portfolio system is used within the alternate assessments in which work samples and pictures depict essential "learning" from the student. These work samples are displayed in a notebook that contains a number of forms on which the teacher (a) records important demographic and program information, (b) introduces the student, and (c) displays outcomes in individual sections organized by standards. We sampled only one student in this state; therefore, it is not possible to ascertain the degree to which these results can be generalized. If an actual alignment study had been conducted, more students would have to be sampled (representing different geographic areas of the state, reflecting different disabilities, and attending different types of programs).

### Step 3

We analyzed every element of a student's portfolio and categorized everything that could be mapped onto a standard. Using both information from the forms at the front of the portfolio (in which the teacher explicitly noted tasks oriented to specific standards) and individual work samples appearing throughout the portfolio (often labeled with language oriented to standards), we counted the number of these "behavioral events." In many instances, the work sample was a picture of the student engaged in an activity, in which case each picture was counted as an event. In other instances, the sample was a worksheet with many items completed (e.g., individual words copied or numbers displayed). In these instances, we counted the entire worksheet as a behavioral event.

### Step 4

*Categorical concurrence addresses whether the same content exists in the assessments as in the standards.* See Table 11 for a summary of Categorical Concurrence appearing at the end of all objectives within each standard. We found the portfolio varied considerably in this dimension. Mathematics, science, and technology standards were marginally considered in this student's portfolio. For analysis, inquiry, and design, one of 10 standards had a work sample or behavioral events; no information system standards were addressed; 23 percent of the mathematics standards, but no science standards, were present in the portfolio; only 5 percent of the technology standards were covered, and no standards dealing with interconnectedness were present in the portfolio. In the end, only 5 of 70 (7 percent) standards were present in the work sample or behavioral events of the portfolio. Standards from career development and occupational studies were not addressed by the behavioral events or work samples in the portfolio.

Table 11

## Concurrence, Depth of Knowledge, Range, and Balance for Portfolio Assessment

MATHEMATICS, SCIENCE AND TECHNOLOGY MST Student Evidence Provided: Use calculator to add/subtract single and double digit numbers; identify, sort, count, and wrap coins; weigh bulk foods at a grocery store.	Depth of Knowledge (DOK) Levels		Behavioral Events	Balance Index (BI)
<b>Standard 1: Analysis, Inquiry, and Design</b>	<b>STD</b>	<b>AA</b>		
Use mathematics and symbolism	3			
Compare and describe quantities	3			
Mathematical relationships	3			
Relate math to immediate environment	2	2	1	$ 1/1-1/1 =0$
Learn to ask "why" questions	4			
Activate devices	2			
Recognize why an object or choice is not working	3			
How to fix a simple object or device	4			
Improve performance of a simple device	4			
Design a structure using modeling materials	2			
MST Student Evidence has one of the ten (10%) possible standards within the Analysis, Inquiry, and Design Standard Domain			O = 1 H = 1	$\Sigma = 0$ $0/2 = 0$ $1 - 0 = 1.0 = BI$
<b>Standard 2: Information Systems</b>	<b>STD</b>	<b>AA</b>	<b>Behavioral Events</b>	<b>Balance Index (BI)</b>
Using software to communicate information	4			
Access information from media, databases	3			
Use communication systems to satisfy needs	2			
MST Student Evidence has 0% of the Information Systems objectives.				BI = N/A
<b>Standard 3: Mathematics</b>	<b>STD</b>	<b>AA</b>	<b>Behavioral Events</b>	<b>Balance Index (BI)</b>
Single digit whole numbers for identification	4			
Use concrete materials to model numbers	3			
Relate counting to grouping using manipulatives	3	1	1	$ 1/3-1/10 =0.23$
Recognize order of whole numbers to 12	2			
Recognize coins and dollars and their value	2			
Add and subtract whole numbers under 12 with calculator	3	3	8	$ 1/3-8/10 =0.47$
Use appropriate measurement tools	2			
Length, weight, volume, time, and temperature	2	4	1	$ 1/3-1/10 =0.23$
Measure length or volume of an object	2			
Collect and display simple data	3			
Recognize and duplicate simple patterns	3			
Explore patterns	3			
Recognize patterns in nature, art, music, literature	4			
MST Student Evidence has three of the 13 (23%) possible objectives within the Mathematics Standard Domain.			O = 3 H = 10	$\Sigma = .93$ $.93/2 = .47$ $1 - .47 = .53 = BI$
<b>Standard 4: Science</b>	<b>STD</b>	<b>AA</b>	<b>Behavioral Events</b>	<b>Balance Index (BI)</b>
Patterns in daily, monthly, and seasonal change	3			
Relationships among air, water, and land	1			
Describe properties of materials	4			
Observe chemical and physical change	1			
Observe energy forms and changes to objects	1			
Use of common forces on objects	3			

Living and nonliving things	3			
Identify simple life processes of all living things	4			
How living things change	1			
Differences within species and survival	1			
Major stages in life cycles	1			
Observe growth, repair, and maintenance	1			
Identify life functions	3			
Identify survival behaviors	3			
Activities that promote good health and growth	3			
Plants and animals depend on each other	4			
Sun as an energy source	4			
Environmental changes and their effects	3			
MST Student Evidence satisfies 0% of the possible objectives within the Science Standard Domain.				BI = N/A
<b>Standard 5: Technology</b>	<b>STD</b>	<b>AA</b>	<b>Behavioral Events</b>	<b>Balance Index (BI)</b>
Recognize an object is not working properly	2			
How a simple object might be fixed	4			
Manipulate components of a simple device	4			
Tell how device has been improved	4			
Design structure or environment	2			
Describe design in words or drawings	3			
Use variety of materials to construct things	3			
Fasten components	2			
Process materials into more useful forms	4			
Safety and ease of use in selecting tools	3			
Basic skills using hand tools	2			
Manufacturing processes to produce a product	3			
Use computer as a tool	2			
Identify and operate familiar systems	2	3	1	$ 1/1-1/1 =0$
Assemble simple systems	3			
Technology safety issues	3			
Responsible disposal of materials	2			
Work cooperatively with others	3			
Plan event or activity	3			
MST Student Evidence satisfies one of the nineteen (5%) possible objectives within the Technology Standard Domain.		O = 1 H = 1		$\Sigma = 0$ $0/2 = 0$ $1 - 0 = 1.0 = BI$
<b>Standard 6: Interconnectedness: Common Themes</b>	<b>STD</b>	<b>AA</b>	<b>Behavioral Events</b>	<b>Balance Index (BI)</b>
API 152: Construct and operate models	4			
API 153: Models can be used to study real thing	4			
API 154: Use models to represent aspects of real world	3			
API 155: Things with different measurements	3			
API 156: Identify biggest and smallest values	3			
API 157: Observe a balance	3			
API 158: Record body temperature	2			
MST Student Evidence satisfies 0% of the objectives within the Interconnectedness: Common Themes Standard Domain.				BI = N/A
MST Student Evidence satisfies 5 of the 70 (7%) possible objectives within the domain of Math, Science, and Technology.				

### Step 5

Depth of knowledge levels for cognitive demands reflected in the standards and made by the assessments use the following scale in special education:

- 1) A behavioral event with 1:1 correspondence completed in single context
- 2) A behavioral event with more than 1:1 correspondence in more than one context with correct or incorrect responses
- 3) A multiple-step behavioral event executed in more than one context with more than 1:1 correspondence and with partial correct scoring of responses
- 4) A multiple step behavioral event executed as an approach (of many) to completing a task that occurs in multiple settings

We have displayed depth-of-knowledge (DOK) levels in Table 11 for alternate standards and then for all work samples or behavioral events present in the portfolio (with very few entries). In the English Language Arts standards, the DOK levels are quite varied across the objectives, but primarily tap rote behavioral chains, with only an occasional discrimination behavior in the alternate assessment. In all of the remaining areas with hits, the DOK levels in the alternate assessment are equal to or exceed the level noted in the standard.

### Step 6

Range of knowledge refers to the percentage of objectives having at least one related task (work sample or behavioral event). In Table 12, the alignment between the objectives from the state and the presence of behavioral events from the student is summarized. Considerable variation exists within each of the standards. Every standard has at least one objective with no behavioral events present in the portfolio. Even when an objective has a behavioral event, it frequently is represented by a small number of unique behavioral events displayed (often by only one). Occasionally, on a few objectives, many different behavioral events are presented.

**Table 12**  
**Range of Knowledge for Portfolio Approach**

Standard #	Standard Content	No. of Objectives with AA	Percent
Mathematics, Science, and Technology			
1	Analysis, Inquiry, and Design	1 of 10	10%
2	Information Systems	0 of 3	0%
3	Mathematics	3 of 13	23%
4	Science	0 of 18	0%
5	Technology	1 of 19	5%
6	Interconnectedness: Common Themes	0 of 7	0%

### Step 7

Balance of representation refers to the degree to which objectives have been more or less emphasized on an assessment (given only those that are addressed in the first place). We calculated balance for each standard (with a portfolio entry) and displayed these results in Table 13. Below, we summarize the results by objectives grouped into the five standards. In general, balance is spuriously high because of the few behavioral events that are spread across the few standards.

**Table 13**  
**Summary of Balance of Representation Analysis for Portfolio Approach to Alternate Assessment**

Math, Science and Technology				
1	Analysis, Inquiry, and Design	1	1	1.0
2	Information Systems	0	0	
3	Mathematics	3	10	.53
4	Science	0	0	
5	Technology	1	1	1.0
6	Interconnectedness: Common Themes	0	0	

## Summary of Alignment for State Two

---

In summary, the portfolio for this student is not well aligned within or across the alternate standards and objectives. Categorical concurrence is varied; depth-of-knowledge level is low, with an emphasis on rote behavioral chains; range of knowledge is varied, with most standards having many objectives not addressed by any assessment tasks (behavioral events or work samples); finally, some objectives have many assessment tasks and many have just a few, preventing balance of representation from being achieved.



## References

- Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3), 21-29.
- Council of Chief State School Officers. (2002). *Models for alignment analysis and assistance to states*. Paper prepared by the Council of Chief State School Officers. Washington, DC: Author.
- Hansche, L. N. (1998). *Meeting the requirements of Title 1: Handbook for the development of performance standards*. Washington, DC: U.S. Department of Education.
- La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, 7(21). Retrieved June 4, 2003, from <http://ericae.net/pare/getvn.asp?v=7&n=21>
- La Marca, P. M., Redfield, D., Winter, P. C., Bailey, A., & Despriet, L. H. (2000). *State standards and state assessment systems: A guide to alignment*. Washington, DC: Council of Chief State School Officers.
- Long, V. M., & Benson, C. (1998). Re. Alignment. *The Mathematics Teacher*, 91(6), 503-508.
- Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003-2004). Benchmarking and alignment of standards and testing. *Educational Assessment*, 9(1 & 2), 1-27).
- Webb, N. L. (1997). *Determining alignment of expectations and assessments in mathematics and science education. NISE Brief*. National Center for Improving Science Education, University of Wisconsin-Madison.
- Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Washington, DC: Council of Chief State School Officers (CCSSO): State Collaborative on Assessment and Student Standards (SCASS), Technical Issues in Large Scale Assessments (TILSA).
- Quenemoen, R., Thompson, S., & Thurlow, M. (2003). *Measuring academic achievement of students with significant cognitive disabilities: Building understanding of alternate assessment scoring criteria* (Synthesis Report 50). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Synthesis50.html>



# Alignment Report 3

---

## Assessing Vertical Alignment

Lauress L. Wise  
Human Resources Research Organization (HumRRO)

Meredith Alt  
Wisconsin Center for Education Research

September 2005



## Table of Contents

<b>Assessing Vertical Alignment</b>	<b>61</b>
Project Goals .....	61
<i>Background</i> .....	61
Vertical Alignment .....	61
Vertical Scaling .....	62
<i>Rationale for measuring growth and scaling</i> .....	62
<i>Issues with Vertical Scaling</i> .....	62
<i>Moving Forward with Vertical Scaling</i> .....	63
Approach to Assessing Vertical Alignment .....	63
<i>Developing a Vertical Alignment Process</i> .....	63
<i>Building on Webb's Model of Alignment</i> .....	63
<i>Questions to be Addressed</i> .....	64
<i>Procedure for Mapping Standards and Objectives Across Grades</i> .....	64
<i>Judging Alignment Quality</i> .....	65
<i>Who Should Judge?</i> .....	65
<i>Reporting Standards Alignment</i> .....	65
<i>What Constitutes Good Alignment?</i> .....	66
<i>Uses of Content Alignment Information</i> .....	66
<i>Mapping Tests onto Tests</i> .....	67
<i>Alternatives to Vertical Scaling</i> .....	67
Next Steps .....	68
<b>References</b>	<b>69</b>
 <b>Appendix A: Sample Instruments and Instructions</b>	 <b>71</b>



# Assessing Vertical Alignment

---

## Project Goals

---

This paper describes results of a project conducted by Council of Chief State School Officers (CCSSO) under subcontract to the Oklahoma Department of Education to extend prior work on the alignment of assessments to content standards. The portion of the project described here concerns the alignment of content objectives from one grade to the next.

The work on vertical alignment was inherently exploratory. Part of that exploration involved determining what states most want to know about the alignment of standards, tests, and even curriculum from one grade to the next. Potential uses range from articulating a logical progression of the curriculum from one grade to the next to building a vertical measurement scale for assessing and reporting individual student progress from one year to the next.

In addition to understanding potential uses of vertical alignment information, the work reported here involved building a prototype process for assessing and reporting the alignment of content standards across a range of grades.

### **Background**

The alignment of expectations for student learning with assessments for measuring students' attainment of these expectations is an essential attribute for an effective standards-based education system. Alignment is defined as the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide an education system toward students learning what they are expected to know and do. Alignment describes the match between expectations and assessment that can be legitimately improved by changing either student expectations or the assessments. As a relationship between two or more system components, alignment is determined by using the 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 more recent call for alignment through No Child Left Behind (NCLB) has also brought attention to vertical alignment, and the degree to which state standards and assessments are aligned across grades. Seeking to create a coherent way of measuring student progress across grades, several states and testing companies are developing vertical scales.

## Vertical Alignment

---

The requirements of NCLB provide one rationale for vertical alignment. Pointing to NCLB requirements for reporting on schools "adequate yearly progress," Margaret Jorgensen (2004) suggested "there will be an increasing emphasis on documenting student progress within selected subject areas" (p. 1). Identifying this progress could be aided by the tests. The requirements for grade-by-grade testing can on the one hand allow

parents and educators to monitor student learning more closely each year and adjust curriculum accordingly (Achieve, 2002). However, Achieve (2002) notes that vertical alignment must pay close attention to both aligning assessments to state standards and aligning the assessments to the grades above and below. The degree to which vertical scaling is an adequate measure of student growth will depend on this alignment of assessments to standards and to assessments in other grade levels. 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).

## Vertical Scaling

According to Michael Kolen (2001), vertical scaling “places scores from tests intended for different educational levels on the same scale, and allows for the tracking of student growth from one educational level to another” (p. 6). A vertical scale (or “developmental scale”) is an expanded normalized standard score scale that spans a range of grades (Nitko, 2004). The expanded scale may either draw on the traditional raw score for students (i.e., the number of items answered correctly), or it may draw on item response theory. In the latter case, an equation is fit to the test publisher’s sample of students’ item responses and the scoring considers the level of difficulty of questions students answered, as well as whether those items distinguish students of different achievement levels (Nitko, 2004). In either case, the vertical scale is designed to be used to interpret student development in a particular subject area, and can provide “a common vocabulary and metric for describing a student’s progress throughout his or her educational journey” (Jorgensen, 2004, p. 1).

### Rationale for Measuring Growth and Scaling

In “Charting the Course of Student Growth,” Dennie Palmer Wolf and Ann Marie White (2000) identify some of the main benefits and issues with monitoring and reporting student achievement over time. Writing in 2000, the authors noted that the cross-sectional design of many large-scale assessments meant that different students are examined at each testing point. Policymakers were thus able to compare performances across cohorts (4th graders in 1997–98 versus 4th graders in 2000–2001), but were unable (without an additional scale) to look at how individual students grow over time. According to Wolf and White, teachers need an assessment system that includes information about whether their teaching is producing student growth (p. 9). For the authors, examining growth requires following representative populations of the same students across grades.

Wolf and White discuss developmental scales in particular as a way to indicate individual growth over time. Rather than using broad achievement levels (such as “failing,” “needs improvement,” “proficient,” or “advanced”), developmental scales are able to go beyond establishing levels of performance that students should achieve at particular points of schooling. The scales, which the authors compare to proficiency scales developed by the foreign language community, could outline the quality of student performance over time, rather than showing only age or grade-based success. These scales would help both teachers and families measure students’ success in working toward state standards. The main obstacles Wolf and White envision to assessing longitudinal growth are financial (i.e., the expense of following individual students over time) and assessment problems associated with selecting valued activities that increase in sophistication rather than cover the domain of educational activities. However, their financial consideration has changed in some ways due to the inevitability of grade-by-grade testing.

### Issues with Vertical Scaling

One of the issues with vertical scaling is the degree to which yearly assessments are themselves aligned. For instance, if the state uses different tests in successive grades—its own standards-based test in some years and off-the-shelf tests in others—the results will not be consistent from grade to grade (Achieve, 2002). In terms of scaling, Brennan (2001) states that examinees and users of test scores are entitled to know how a test is scored and that anything other than total number of items correct/total number of points is likely to be challenged by examinees and users of test scores (p. 11). These two assertions lead Brennan to question pattern scoring as a basis for reported scores for individual students. Brennan also notes that the transformation of raw scores to a scale necessarily involves making assumptions that are usually not value-free and may be masked by statistical complexity (p. 11). Finally, in some states, efforts to align curriculum/assessment to standards and to create a common vertical scale have coincided with efforts to rank schools according to their



perceived success in educating children. In Indiana, Public Law 221 has included each of these elements, stirring controversy about the ways such rankings may affect teachers and communities (Sheese & McDaniel, 2002).

### **Moving Forward with Vertical Scaling**

Despite some concerns about methodological issues with scaling or the potential use of vertical scales, both test developers and states are moving forward with efforts to create vertical scales. Harcourt Assessment, Inc., for instance, has developed a vertical scale for the Stanford Achievement Test and its other assessments, the common vertical scale (for each subject within a test) being one that can be used with multiple testing products (Jorgensen, 2004). The state of Florida has also created a developmental scale score in reading and mathematics that links adjacent grades and attempts to track progress over time (Florida Department of Education, 2004). Florida students' test scores are reported in three ways: by achievement level, by scale score, and by developmental scale score (which ranges from 0-3000 across grades 3–10). Students are told the scale score and developmental scale score for required proficiency in particular grades, and their developmental scale maps their progress and underpins their study plan to meet those proficiency levels (Florida Department of Education, 2004). North Carolina also reports a developmental scale score for reading and mathematics—in this case, through its end-of-grade tests. The raw score is converted to a developmental scale score, and students receive developmental scale scores, percentiles that compare that year's individual student performance to other North Carolina students of the same grade, and students' achievement levels on a scale (ranging from a low of 1 to a high of 4) (North Carolina Department of Public Instruction, 2004). Such scales are increasingly likely to be implemented as states seek to measure the alignment of curriculum and standards across grade levels.

## **Approach to Assessing Vertical Alignment**

### ***Developing a Vertical Alignment Process***

The approach taken in the current effort involved taking a step back in the rush to create vertical scales for describing growth in individual student achievement. The focus is on a key question that should be answered first: What is the domain that we want our vertical scale to assess?

The approach to assessing vertical alignment described in this report was developed by the authors in conjunction with a working group of the CCSSO collaborative on Technical Issues in Large Scale Assessments (TILSA). Members of the working group participated in a brief pilot, conducted during the October 2003, TILSA meeting (using content standards from West Virginia). Subsequently, Oklahoma Department of Education staff members used an updated process to collect vertical alignment judgments for their state's content standards. Refinements from each of these exercises have been incorporated into this final concept paper.

### ***Building on Webb's Model of Alignment***

The proposed approach to assessing and describing the alignment of content objectives across grades builds on Webb's (1997) framework and methodology for assessing the alignment of tests to content standards within a given grade. Within this system, content standards are organized into two or more hierarchical levels. The term "standards" is used for the highest level in this hierarchy, while the term "objective" is consistently used for the lowest (or all lower) levels within the content framework hierarchy.

Assessing the alignment of tests to content standards within Webb's frameworks involves the following dimensions of alignment:

- 1) **Categorical Concurrence:** Is the same content covered?
- 2) **Depth of Knowledge:** Are cognitive requirements between the assessments and standards consistent for each of the content areas covered? Is the same complexity of knowledge (and skill) sought/required?
- 3) **Range of Knowledge:** Is the range of content covered under each of the major content standards similar?

- 4) **Balance of Representation:** Are objectives for a particular standard given the same emphasis?
- 5) **Source of Challenge:** Does test performance actually depend on mastering the target objectives and not on irrelevant knowledge or skills?

### ***Questions to be Addressed***

Our assessment of content alignment across grades focuses on two key questions:

- 1) What is the nature of content linkages from one grade to the next?
- 2) What is the clarity or quality of these linkages?

The first question may be addressed in terms of the first four dimensions of Webb's model. Specifically,

- What level of concurrence is there between objectives for the two grades?
- To what extent do comparable objectives increase in depth from one grade to the next?
- To what extent does the range of content increase from one grade to the next?
- How does the balance of representation change from one grade to the next?

The second question is aimed at identifying ways in which the specification of content objectives at each grade might be improved, both to clarify what is expected within a grade and to make clearer the nature of expected growth from one grade to the next.

The Content Linkage Checklist described below solicits information to answer these questions. Options for summarizing the information gathered are described under reporting options below.

### ***Procedure for Mapping Standards and Objectives Across Grades***

Mapping the content standards for a grade onto the content standards for one or more prior grades can be achieved in two steps. The first step is to line up the major content areas (referred to as standards in Webb's system). In most cases, these will be broad areas of content, such as number operations in mathematics, and will be identical across a wide range of grades. In other instances, the higher-level organization of the content standards may be expanded in succeeding grades. By lining up the major areas of content objectives from one grade to the next, we can limit the number of objectives searched in finding corresponding objectives at each higher or lower grade.

Within each broad content area, the mapping of specific objectives is accomplished by asking for each higher grade standard: "Is this standard related to one of the standards for the prior grade(s)?" Possible responses and follow-up actions are

- Yes. The corresponding standard is then linked.
- No, not at all. The new content standard is added to a combined list of content standards.
- The higher grade standard is partly covered by one or more lower-grade standards. In this case, the task is to create a combined standard that encompasses the objectives of all of the standards involved. In most cases, the higher-grade standard encompasses the lower-grade standard, so the higher-grade standard is the combined standard. This effort will break down, however, if the higher-grade standards are mostly a rearrangement, rather than an expansion, of the lower-grade standards.

The result of the first step will be expanded or reorganized groupings of specific content objectives for each grade.

The next step is to compare the specific objectives (detailed content standards) for the higher grade to the objective for the lower grade(s). Using results from the first step, these comparisons can proceed separately for each content standard (general content area), thereby limiting the complexity of the comparisons. The question, for each higher-grade objective, is whether it is related to one or more of the lower-grade objectives. Several types of relationships are possible including

- **Broader:** The higher-grade standard reflects a broader application of the target skill or knowledge (generalizing from specific to additional applications).
- **Deeper:** The higher-grade standard reflects deeper mastery of the target skill or knowledge (e.g., application rather than recognition).
- **Prerequisite:** The lower-grade standard reflects a different, but prerequisite skill for mastery of the higher-grade standard.
- **New:** The higher-grade standard is a new skill or knowledge unrelated to skills or knowledge covered at prior grades.

In preliminary pilot tests, another possible relationship was encountered:

- **Identical:** The higher-grade standard appeared to be identical to one of the lower-grade standards.

The working group was hard pressed to identify an instance where simply repeating a standard at the next higher grade was a good thing. In most cases, further clarification of the standards would be needed to indicate what additional mastery was required at the higher grade.

### **Judging Alignment Quality**

As specific linkages are identified, the next step is to rate the quality of these linkages. The working group developed a two-part process consisting of a summary rating and a checklist of sources of potential ambiguity in the linkage. The summary rating is on a three-point scale: (1) serious ambiguity in the relationship, (2) minor ambiguity in the relationship, and (3) no ambiguity.

The Content Linkage Checklist includes specific issues encountered in pilot test work. Examples are: ambiguity in the individual objectives, using different terms for the same thing, using the same term to mean different things, and lack of detail on differences in expectations. A current draft of the checklist is included in Appendix A. With further piloting, it is likely that additional sources of challenges will be identified and the alignment quality checklist will be correspondingly expanded.

### **Who Should Judge?**

Another important question considered by the working group was who should be asked to provide the required judgments. During the initial pilot, many of the working group members, most of whom were testing experts rather than content experts, expressed the belief that they were not themselves sufficiently familiar with the content standards to provide the best ratings of alignment. This led to a discussion of how best to constitute alignment rating panels.

The primary recommendation from the working group was that the individuals involved in developing or recommending the initial standards should be heavily represented on the alignment panels. The panels should also include expert teachers from the grades covered by the alignment process who were not involved in standards development. These teachers would provide the best assessment of the degree to which the existing standards communicate effectively to others not in the room when the standards were developed. Expert teachers are also critical because enactment of the standards depends heavily on their understanding and participation.

In many states, efforts are made to include parents, representatives of the business community, and higher-level policy-makers in developing and adopting content standards that define our expectations for students. It would seem prudent to include these other stakeholders in alignment panels to the same extent that they were represented in the initial development of the standards.

### **Reporting Standards Alignment**

The alignment judgments, particularly at the objective level, are detailed and somewhat complex. A few simple summary indicators are needed to communicate and evaluate the overall results. Webb's first four criteria are most directly relevant to comparing standards to standards. The fifth criterion, Source of Challenge, will be modified to cover issues raised about the clarity of grade-to-grade linkages.

Results from the standards-to-standards mapping will be reported in the following ways:

- **Categorical Concurrence.** Results for each standard are summarized by a simple listing of the objectives common to or at least partially covered in both grades and the objectives unique to the

lower or higher grades. The ratio of the number of common standards to the total number of standards serves as an overall indicator of alignment.

**Depth of Knowledge.** Increases in depth of knowledge can be assessed by computing the proportion of objectives at one grade that are judged to be covered in greater depth by corresponding objectives at the next higher grade. Note that simple comparisons of the depth-of-knowledge ratings from a Webb assessment of test alignment may be misleading. A skill which involves deduction or novel application at one grade can become much more routine at a higher grade. In this case, the same skill would be judged to be lower depth at the higher grade, running counter to expectations that depth should increase over grades.

**Range of Knowledge.** Increases in range of knowledge can be assessed by computing the proportion of objectives at one grade that are judged to be covered more broadly by corresponding objectives at the next higher grade and also the proportion of objectives at the higher grade that are new.

**Balance of Representation.** Grade-to-grade changes in the balance of representation can be assessed by counting the number of objectives in each of the broad area of standards for each grade. Alternatively, representation could be assessed in terms of the number of test questions targeted for each of the broad standards.

Beyond these summary indicators, an important way of reporting standards alignment is to simply lay out the objectives for adjacent grades, side-by-side, showing the judged relationships for related objectives within each standard. Across a range of grades, these displays would be similar to “scope and sequence” charts used by developers of integrated curricula.

The final reporting element concerns Sources of Challenge. The clarity of grade-to-grade linkages can be reported by listing specific ratings and comments on the grade-to-grade linkages and by indicating the proportion of linkages flagged for specific challenges.

### ***What Constitutes Good Alignment?***

Criteria for judging the alignment of tests to content standards are relatively clear. Tests should cover the content standards, completely, evenly, and at an appropriate depth of knowledge. Criteria for determining the quality of the alignment of content standards across grades are considerably less obvious.

Some tentative criteria for good content alignment across grades are

- Standards and objectives for each grade should show logical relationships to content covered in preceding grades.
- Coverage of each subject should, on average, increase evenly in depth across grades.
- Coverage of each subject should, on average, increase evenly in breadth across grades.
- Objectives at one grade should not be identical to objectives for lower grades.

### ***Uses of Content Alignment Information***

Many states consider the vertical alignment of content standards as a step in constructing a common (vertical) scale for measuring growth in student achievement from one grade to the next. Review of content alignment across grades is, however, a useful end in itself. Some of the uses of this information include:

- identifying areas where content standards may need to be clarified to make the progression of expected skill or knowledge more evident
- providing a basis for development of integrated curriculum which progresses smoothly from one grade to the next
- providing a framework that teachers can use as they communicate an individual student’s strengths and weaknesses to teachers at the next higher grade

In an early pilot of content alignment procedures, there was considerable resistance from panelists who had been involved in developing the existing standards. They were rightfully concerned with the need for stability in the content standards and did not want to have to go through the debate and compromise that would be required if the alignment analyses were to suggest a need to change the standards. They did, however,

respond positively to the opportunities to clarify the existing standards where the grade-to-grade progression was not clear to all. They also saw clear advantages to helping teachers understand increases in specific expectations from one grade to the next.

### ***Mapping Tests onto Tests***

At the outset of this project, we planned to consider ways of comparing the content alignment of tests from one grade to the next. As a result of our consideration and pilot testing of methods for comparing content standards across grades, we have reservations about such an approach. Within-grade alignment of tests to the content objectives continues to be paramount in assessing and reporting student achievement. The degree of overlap of each grade-level assessment to the objectives for other grades is of lesser importance. The bottom line is that if the tests are well-aligned to objectives for their targeted grade, then the alignment of tests across grades should follow closely the alignment of content standards across grades. If the tests do not align well with objectives targeted for their grade, then there is a serious problem and further discussion is unwarranted until that problem is resolved.

### ***Alternatives to Vertical Scaling***

Where content standards are not closely related from one grade to the next, many argue that a vertical scale would provide misleading information. A student's score from an assessment of one grade's content would be interpreted as indicating the degree to which he or she has mastered the next grade's content, even though the student was not assessed on that content at all. Some alternatives for assessing individual student progress include

- **Changes in normative rank.** A student at the 30th percentile would make normal progress if he or she remained at the 30th percentile in results from the next year's assessment and would show growth above expectation if achieving a higher percentile score. This approach is similar to prior models of "grade equivalent" scores, where percentile scores from each grade's assessment were transformed to "normal curve equivalents" mapping the 50th percentile onto the expected score for the grade. One problem with normative approaches is that it is difficult to show overall growth over time. There will never be more than 50 percent of the students above the median. This problem might be overcome by using normative information to develop an initial, grade-related scale and then holding this scale fixed over time.
- **Changes relative to standards.** An alternative approach would be to chart individual student progress in terms of changes in student scores relative to performance standards for each grade. Thus, a student halfway between basic and proficient cutoffs in one grade would be making normal progress if she were also halfway between the basic and proficient cutoffs at the next higher grade. One problem with this approach is that the judgments about proficiency may vary across grades, so that average growth artificially appears much greater at some grades than at others. Increasingly, states are using normative information to judge the relative aggressiveness of performance standards at different grades and adjusting cutoff scores (performance level definitions) where needed for greater consistency.
- **Diagnostic Profiling.** Rather than an overall score, growth in individual student achievement might be communicated by listing the additional standards that a student has mastered each year (or the ones he or she has failed to master). Accurate assessment of mastery of individual standards is not generally feasible within the framework of a once-a-year test. Knowing which standards a student has mastered is, however, the job of the classroom teacher. Classroom or end-of-unit assessments can supplement annual assessment information and be used to support teacher judgments of individual student strengths (skills mastered) and weaknesses (skills not yet mastered). Coordinated frameworks from the assessment of the vertical alignment of content standards might be used to organize communication of these judgments to students, parents, and teachers at the next grade.

## Next Steps

---

A full-scale pilot of the proposed vertical alignment procedures is planned for April 2006. The pilot will result in suggested revisions to data collection instruments and more detailed examples of report formats. Following these revisions, several steps are needed to make vertical alignment procedures fully operational. These include

- Develop expanded training materials. These should include instructions for workshop leaders as well as for the panelists themselves.
- Provide detailed instructions for the analysis of panelist ratings.
- Specify reporting procedures, providing options to aid interpretation where possible.
- Consider automating the data collection and analysis procedures through interactive software.

The final step will be to work with a range of states to extend and adapt the alignment procedures to meet their specific needs.

## References

- Achieve, Inc. (2002). No Child Left Behind: Meeting challenges, seizing opportunities, improving achievement. Achieve Policy Brief, Issue Number 5.
- Anderson, L. W., & Krathwohl, (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives. New York: Longman.
- Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice*, 20(4): 6-18.
- Florida Department of Education. (2004). Understanding FCAT Reports 2004. June 2, 2004: [http://www.firn.edu/doe/sas/fcat/pdf/fc\\_ufr2004.pdf](http://www.firn.edu/doe/sas/fcat/pdf/fc_ufr2004.pdf)
- Jorgensen, M. (2004). The value of the Stanford Scale as a common metric. Harcourt, Inc.
- Kolen, M. (2001). Linking assessments effectively: Purpose and design. *Educational Measurement: Issues and Practice*, 20(1): 5-9.
- Nitko, A.J. (2004). Educational assessment of students. Upper Saddle River, NJ: Pearson/Merrill Prentice Hall.
- North Carolina Department of Public Instruction. (2004). Understanding your child's end-of-grade test scores. June 2, 2004: <http://www.ncpublicschools.org/accountability/testing/eog/>
- Palmer Wolf, D. & White, A. M. (2000). Charting the course of student growth. *Educational Leadership*, 57(5): 6-11.
- Sheese, J., & McDaniel, T. (2002). Assessing schools: Not child's play. *Kappa Delta Pi*, 38(2): 68-72.
- Webb, N. L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.





## Appendix A: Sample Instruments and Instructions

### Sample Content Linkage Checklist Rating Sheet

#### Target Objective:

4.1.1	Identify level appropriate vocabulary (e.g., multiple meaning words; synonyms; antonyms; homonyms; content area vocabulary; context clues).
-------	---

Most Similar Grade 3 Objective (No.): \_\_\_\_\_

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deeper, Same, Prerequisite, New)

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear)

Source(s) of ambiguity:

- ☐ Different words are used for the same knowledge or skill  
Which words? \_\_\_\_\_
- ☐ The same word(s) is (are) used in different ways  
Which words? \_\_\_\_\_
- ☐ Differences between the two standards are not clear.  
(They appear to be essentially the same.)
- ☐ The higher grade standard is unclear or imprecise.
- ☐ The lower grade standard is unclear or imprecise
- ☐ Other (explain):  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

Alternate Grade 3 Standard (No.): \_\_\_\_\_

Nature of Linkage Code(s): \_\_\_\_\_ (Extend, Deeper, Same, Prerequisite, New)

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality to 1-Not Clear)

Source(s) of ambiguity:

- ☐ Different words are used for the same knowledge or skill  
Which words? \_\_\_\_\_
- ☐ The same word(s) is (are) used in different ways  
Which words? \_\_\_\_\_
- ☐ Differences between the two standards are not clear.  
(They appear to be essentially the same.)
- ☐ The higher grade standard is unclear or imprecise.
- ☐ The lower grade standard is unclear or imprecise
- ☐ Other (explain):  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

### Instructions for Completing the Content Linkage Checklist

You will be asked to provide information on both the nature and quality of linkages between 3rd and 4th Grade Reading objectives. Response options are listed on this page. Further explanation of the different codes and, in some cases, examples are provided on the following pages.

#### Nature of Linkage:

The higher grade objective (check all that apply):

- ☐ **broadens (B)** the lower grade knowledge or skill to different/broader applications (applied to novel situations; or simply different domains)
- ☐ requires a **deeper mastery (D)** of the lower grade knowledge or skill (performance of skill is more precise, more automatic, or requires less cuing)
- ☐ is essentially **the same (S)** as the lower grade knowledge or skill.
- ☐ describes a different knowledge or skill for which mastery of the lower grade knowledge or skill is **prerequisite (P)**.
- ☐ is **New (N)** and does not match any of the prior grade objective.

#### Quality of the Linkage:

The quality of each linkage is rated on a 3-point scale, with explanatory text when there are issues.

3. The relationship of the higher and lower grade standards is clear and precise.
2. There is some ambiguity in the relationship of the two standards.
1. The relationship is not at all clear because (check all that apply):

Source of ambiguity:

- ☐ Different words are used for the same knowledge or skill  
Which words? \_\_\_\_\_
- ☐ The same word(s) is (are) used in different ways  
Which words? \_\_\_\_\_
- ☐ Differences between the two standards are not clear.  
(They appear to be essentially the same.)
- ☐ The higher grade standard is unclear or imprecise.
- ☐ The lower grade standard is unclear or imprecise
- ☐ Other (explain):  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

## Explanation and Examples

1. **Broadens (B)** the lower grade knowledge or skill to different/broader applications (applied to novel situations; or simply different domains)

Extension covers situations where students are expected to know additional related facts or to generalize a skill to a wider range of settings.

Example: Grade 3 and 4 Writing, Punctuation Objectives

**Grade 3 Objective:** Students are expected to demonstrate punctuation in writing including:

- a. Periods in abbreviations and sentence endings.
- b. Question marks.
- c. Commas in dates, addresses, locations, quotes, introductory words, words in a series, greetings, and closings in a letter.
- d. Apostrophes in contractions and possessives.
- e. Colon in notation of time, formal letter writing, and the introduction of words or concepts in a series.
- f. Quotation marks around direct quotations, the titles of individual poems, and short stories.

**Grade 4 Objective:** Students are expected to demonstrate appropriate punctuation in writing including:

- a. Parentheses
- b. Quotation marks.
- c. Terminal punctuation.
- d. Apostrophes in contractions and possessives.
- e. Commas.
- e. Colons and semicolons.

The 4th Grade objective repeats most of the material covered in the 3rd grade objective and adds additional types of punctuation that the student is expected to use appropriately (parentheses and semicolons). Descriptions of repeated material (e.g., commas and colons) are more abbreviated and, in one case, combined into a single category (terminal punctuation rather than periods and question marks separately).

2. **Requires a deeper mastery (D)** of the lower grade knowledge or skill (performance of skill is more precise, more automatic, or requires less cuing).

Depth of mastery is intended to reflect a deeper level of cognitive processing as described in Bloom's taxonomy or similar schema (e.g., Anderson and Krathwohl's revision). Depth of mastery ranges from simple recall of facts to understanding the significance of the facts, applying this understanding in carrying out appropriate procedures, analyzing new situations or information through the understanding of relevant facts and theories, to evaluation or synthesis of new material.

Example: Grade 4 and 5 Reading, Summarize Information:

**Grade 4 Objective:** Represent text information in different ways such as outline, timeline, or graphic organizer.

**Grade 5 Objective:** Organize text information in different ways (e.g., timeline, outline, graphic organizer) to support and explain ideas.

The grade 4 Objective requires students to demonstrate understanding by representing information in different formats. The grade 5 Objective requires students to use alternative displays to synthesize or evaluate the information presented.

3. **The same (S)** as the lower grade knowledge or skill.

The standards for the different grades are not discernibly different.

4. A different knowledge or skill for which mastery of the lower grade knowledge or skill is **prerequisite (P)**.

This type of linkage should be used sparingly, as it would be easy to say that a great many skills at one grade are prerequisite to more specific skills at the next grade. The most appropriate use is where foundational skills at one grade are not restated at the higher grade, but are clearly needed for mastery of one or more new objectives introduced for the first time at the higher grade.

5. **New (N)** and does not match any of the prior grade objective.

The degree of new and largely unrelated material introduced at each grade will vary considerably from one subject to another. The introduction of new material is a second way of broadening the range of content.



## Alignment Report 4

---

### Vertical Alignment of Grade-Level Expectations for Student Achievement: Report of a Pilot Study

Lauress L. Wise, Human Resources Research Organization, (HumRRO)  
Liru Zhang, Delaware Department of Education  
Phoebe Winter, independent consultant for CCSSO  
Leslie Taylor, Human Resources Research Organization, (HumRRO)  
D. E. (Sunny) Becker, Human Resources Research Organization, (HumRRO)

Dec. 2, 2005

Prepared for:  
Delaware Department of Education  
Dover, DE  
  
Council of Chief State Schools Officers  
Washington, DC  
Contract Number: C03-16

## Acknowledgments

---

The authors acknowledge the support of staff at the Oklahoma Department of Education (Ronda Townsend) and at CCSSO (Art Halbrook) who managed the grant under which the vertical alignment procedures described here were developed. We also acknowledge and thank members of the CCSSO State Collaborative on Assessment and Student Standards (SCASS) project for Technical Issues in Large Scale Assessment (TILSA) who participated in the development of these procedures.

In addition, we are deeply grateful to Wendy Roberts and other members of the Delaware Department of Education staff who agreed to take on this pilot and provided essential sponsorship and logistics.

## Table of Contents

<b>Chapter 1: Introduction</b>	<b>79</b>
Establishing Explicit Expectations for Student Performance under NCLB .....	79
<i>Delaware Efforts to Meet NCLB Requirements</i> .....	80
The CCSSO Alignment Project(s) .....	80
<i>Description of the Overall Grant</i> .....	80
<i>Concept Paper on Vertical Alignment</i> .....	80
Goals for the Pilot Study .....	80
<i>Opportunity to Try Out and Improve the Vertical Alignment Process</i> .....	80
<i>Implications for NCLB Requirements</i> .....	80
Organization of the Report .....	80
<b>Chapter 2: Methods</b>	<b>81</b>
Workshop Participants .....	81
<i>Identifying and Recruiting Appropriate Panelists</i> .....	81
<i>Selection and Training of Table Facilitators</i> .....	82
<i>Rating Tables and Review Assignment</i> .....	82
Vertical Alignment Ratings .....	82
<i>Rating Forms</i> .....	83
Discussions of Additional Dimensions for Grade-Level Expectations .....	83
Debriefing and Evaluation .....	84
Analysis of Results .....	84
<b>Chapter 3: Results</b>	<b>85</b>
Agreement across Independent Groups .....	85
<i>Mathematics</i> .....	85
<i>English Language Arts</i> .....	86
<i>Summary of Agreement Rates</i> .....	87
Nature of Alignment from One Grade to the Next .....	88
<i>Mathematics</i> .....	88
<i>English Language Arts</i> .....	88
Ratings of the Importance of Increases in Expectation .....	89
<i>Mathematics</i> .....	89
<i>English Language Arts</i> .....	90
Clarity of Alignment of Expectations .....	91
<i>Mathematics</i> .....	91
<i>English Language Arts</i> .....	92
Detailed Workshop Results .....	92
Debriefing Session .....	93
<b>Chapter 4: Recommendations</b>	<b>95</b>
Improvements to the Vertical Alignment Process .....	95
Suggestions for Delaware's Grade-Level Expectations .....	95
<b>References</b>	<b>96</b>

### List of Appendices

Appendix A:	Delaware Content Standards	97
Appendix B:	Sample Rating Sheet for English Language Arts	99
Appendix C:	Sample Rating Sheet for Mathematics	100
Appendix D:	Sample Reference Sheet for English Language Arts	101
Appendix E:	Sample Reference Sheet for Mathematics	102
Appendix F:	Table Leader Orientation Slides	103
Appendix G:	Vertical Alignment Workshop Training Slides	116
Appendix H:	Summary of Responses to Debriefing Questions	128
Appendix I:	Vertical Alignment Workshop Evaluation Survey	130

<b>List of Tables</b>
-----------------------

Table 1.	Characteristics of the Workshop Participants	81
Table 2.	Standards Rated by Each Table	82
Table 3.	Grade Ranges Rated by Each Teacher	82
Table 4.	Agreement of Nature of Link Ratings from Independent Groups-Mathematics	85
Table 5.	Agreement of Clarity of Link Ratings from Independent Groups-Mathematics	85
Table 6.	Agreement of Importance Ratings from Independent Groups-Mathematics	86
Table 7.	Agreement on Expectation Matched across Independent Groups-Mathematics	86
Table 8.	Agreement of Nature of Link Ratings from Independent Groups-English Language Arts	87
Table 9.	Agreement of Clarity of Link Ratings across Independent Groups-English Language Arts	87
Table 10.	Agreement of Importance Ratings across Independent Groups-English Language Arts	87
Table 11.	Nature of Increases in Mathematics Expectations for Each Standard	88
Table 12.	Nature of Increases in Mathematics Expectations by Grade	88
Table 13.	Nature of Matches for Each Language Arts Standard	89
Table 14.	Nature of Increases in Language Arts Expectations by Grade	89
Table 15.	Importance of Expectation Increases for Each Mathematics Standard	90
Table 16.	Importance of Increases in Mathematics Expectations by Grade	90
Table 17.	Importance of Expectation Increases for Each Language Arts Standard	90
Table 18.	Importance of Increases in Language Arts Expectations by Grade	91
Table 19.	Clarity of Expectation Increases for Each Mathematics Standard	91
Table 20.	Clarity of Mathematics Expectation Increases by Grade	91
Table 21.	Clarity of Expectation Increases for Each Language Arts Standard	92
Table 22.	Clarity of Language Arts Expectation Increases by Grade	92



# Vertical Alignment of Grade-Level Expectations for Student Achievement: Report of a Pilot Study

---

## Chapter 1: Introduction

---

### ***Establishing Explicit Expectations for Student Performance under NCLB***

A key, albeit difficult, component of the No Child Left Behind Act of 2001 (NCLB) is that each state must establish explicit standards for student performance at each grade level in reading and mathematics (and soon also science). For grades three through eight and at least once in high school, states must establish: (1) specific content objectives in English language arts and mathematics, (2) an assessment aligned to these content objectives, and (3) standards for meeting at least three levels of performance on the assessment, representing below proficient, proficient, and advanced achievement.

Prior to NCLB, many states were already engaged in standards-based reform efforts that involved public and specific content standards and assessments of performance against those standards. Few states, however, had implemented either standards or assessments for the complete range of grades and subjects required by NCLB.

The vertical alignment process described here examines the relationship of content standards and expectations from grade to grade. Through this process, panelists evaluate how the standards and expectations at one grade are related to standards and expectations at the next grade. In addition, they are asked to rate the clarity of increases in expectations from one grade to the next. A review of the consistency of expectations across grades can provide validity evidence for content standards used for accountability under NCLB.

### **Delaware Efforts to Meet NCLB Requirements**

With the adoption of the rigorous content standards in English language arts, mathematics, science, and social studies in 1995, Delaware educators have continued the efforts to implement standards-based curriculum and assessment in order to meet the goal of improved achievement for all students. According to state regulations, the Delaware Student Testing Program (DSTP) is the official measure of student progress toward the standards in reading, writing, and mathematics by the end of each grade cluster (K–3, 4–5, 6–8, and 9–10).

NCLB requires that states report student progress toward meeting the state standards by grade for grades three through eight even though the content standards are developed by grade cluster. To meet the NCLB requirements, Delaware developed the Grade-Level-Expectations (GLE) in early 2005 based on the Delaware content standards for every grade from kindergarten through grade 10 in English language arts and from kindergarten through grade 11 in mathematics. These expectations will be used to develop curriculum and instruction at the local level and develop or identify items for inclusion on the DSTP at the state level. Although the expectations for students by the end of a grade cluster have few changes, the goals and expectations are clearly specified by the end of each grade. The GLEs have been reviewed by Delaware teachers and educators and recently finalized.

It is important to be aware that at the time the vertical alignment workshop was conducted, the Grade-Level-Expectations were just available as a draft version. As described below, the vertical alignment workshop provided the panels an opportunity to read/review the newly developed GLEs and collect inputs to the process for the upcoming workshops to review/set the performance standards (cut scores) for all tested grades. Specifically, the alignment of expectations across grades was intended to help the development of the performance descriptors.

### ***The CCSSO Alignment Project(s)***

#### **Description of the Overall Grant**

The workshop design stemmed from a project conducted under a grant from the U.S. Department of Education (ED) to the Oklahoma Department of Education, in cooperation with CCSSO. The project's three primary goals were: 1) to produce an electronically-based test-to-standards alignment analysis process based on Webb's (2001) alignment method; 2) to use Webb's method to develop procedures appropriate to aligning alternate assessments for students with disabilities to content standards; and 3) to expand the method to inform vertical alignment and scaling of assessments and standards.

#### **Concept Paper on Vertical Alignment**

The CCSSO work on alignment resulted in a concept paper describing a process that states might use to check the alignment of content standards across grades (Wise & Alt, 2005). This paper described the types of judges who might review the vertical alignment of a state's content standards, the types of ratings these judges would be asked to make, and how the results of these ratings could be summarized and reported. The pilot test reported here was designed to try out the procedures described in the concept paper, and it included a number of extensions and enhancements of the procedures to better meet Delaware's goals for the workshop. These goals are described next; modifications and extensions to the vertical alignment procedures are discussed in the next chapter.

### ***Goals for the Pilot Study***

#### **Opportunity to Try Out and Improve the Vertical Alignment Process**

This report describes a workshop held to review the alignment of grade-level expectations across grades. The workshop served as a full-scale pilot study of the vertical alignment process described in the concept paper developed by CCSSO. Expenses for HumRRO and CCSSO staff to conduct this workshop were covered by funds remaining in the initial grant. Thus, in addition to meeting Delaware's goals for this workshop, this report also describes further enhancements to the vertical alignment process based on planned and unplanned results from this pilot.

#### **Implications for NCLB Requirements**

The workshop supported Delaware's efforts to meet NCLB requirements. This workshop created an opportunity for Delaware teachers and educators to review the newly developed Grade-Level-Expectations (GLEs) in English language arts and mathematics. Feedback, comments, and suggestions provided by the panelists were helpful to the Department of Education in revising and finalizing the GLEs. The workshop also provided input to the process for developing the performance level descriptors—what students are expected to know and be able to do at each performance level.

### ***Organization of the Report***

The methods used with the pilot vertical alignment workshop are described in Chapter 2, including a brief description of the panelists who participated in the workshop, the rating tasks the panelists were asked to perform, and the methods used to analyze the ratings and summarize the feedback provided by the panelists.

The results from the workshop are presented in Chapter 3, which includes the nature of alignment across grades, an examination of agreement of ratings across tables (groups of panelists), a summary of the ratings of Importance and Clarity, and a description of detailed results for the Delaware Department of Education.

Recommendations based on results from the workshop are presented in Chapter 4. These recommendations focus on continued enhancement of the process and criteria for vertical alignment and the further improvement of the Delaware Grade-Level-Expectations

## Chapter 2: Methods

The vertical alignment process is a way of making explicit the assumptions about grade-to-grade growth in knowledge and skills that are implicit in the content standards, objectives, and expectations for each separate grade. The process involves assembling content experts to identify objectives that are new at each grade and to link other objectives to related objectives from earlier grades. Where objectives are linked across grades, the panelists are asked to rate how knowledge and skill requirements increase across the grades.

Different states use different terminology for their content standards. In this report, we have adopted the terminology used in Delaware. The Delaware content standards are structured with four broad statements in English language arts and eight in mathematics across grades. Objectives that specify the knowledge and skills are under the general statements by the end of each grade cluster. The newly developed Grade-Level-Expectations further specify what students are expected to know and be able to do by the end of each grade.

As noted in Chapter 1, the methods used in the Delaware vertical alignment workshop included modifications, enhancements, and extensions to the procedures proposed in the vertical alignment concept paper developed for CCSSO. In this chapter, we describe (a) the recruitment of panelists for the workshop, (b) the rating tasks used in this workshop, and (c) the methods used to analyze the results. The results themselves are presented in Chapter 3.

### Workshop Participants

#### Identifying and Recruiting Appropriate Panelists

The Delaware content experts served on the two panels, one for English language arts and one for mathematics. A total of 56 classroom teachers and curriculum specialists throughout the state participated in the vertical alignment workshop at the elementary, middle, and high school levels. Table 1 shows the demographic characteristics of panelists. Over 80 percent of the panelists were experienced teachers and content experts (over six years), and 14–15 percent of the panelists were new teachers. Many had very extensive teaching careers with over 10 years of service.

Table 1. Characteristics of the Workshop Participants

Demographic Category	Number (Percent) of Panelists	
	ELA	Mathematics
<b>Gender</b>		
Female	25 (93%)	23 (79%)
Male	2 (07%)	6 (21%)
<b>Race Ethnicity</b>		
Caucasian	23 (85%)	24 (83%)
African American	3 (11%)	3 (10%)
Other/Unknown	1 (04%)	2 (07%)
<b>Grade Taught</b>		
Primary School	9 (33%)	10 (34%)
Middle School	9 (33%)	8 (28%)
High School	9 (33%)	9 (31%)
<b>Years of Experience</b>		
2–5	4 (15%)	4 (14%)
6–9	9 (33%)	3 (10%)
10+	14 (52%)	21 (72%)

### Selection and Training of Table Facilitators

Table facilitators were selected from panelists according to their experience and expertise, three for English language arts and four for mathematics. The table facilitator served to facilitate discussion and direct activities. A training session was held prior to the workshop. The training included an overview of the workshop activities and the role and responsibilities of the process.

### Rating Tables and Review Assignment

The panelists were split into seven groups, three for English language arts and four for mathematics. Each group included 6–9 panelists, 2–3 each for elementary (grades 2–5), middle (grades 5–8), and high school level (grades 8–10 or 11), respectively. Each group was seated at a separate table and led by a table facilitator. The review and rating started with subgroup discussion within each grade level. This was followed by the groupwide discussion across grades. The overlapping arrangement of panelists for the subgroups created the opportunity for communication between elementary, middle, and high school levels and examined the consistency of ratings for grades 5 and 8.

Each group was assigned a standard for the primary review. (See appendix A for a list of the standards reviewed.) Each standard also received a secondary review by another group to provide additional results and allow for examining the consistency of groupwide results. The four mathematics groups were each assigned one of the four content standards for the first round and another standard for the second round. The English language arts groups started with either Reading Standard 2 or Writing Standard 4. Because of relatively small number of objectives/expectations of Reading Standard 4, connections, this standard was not reviewed until the second round. Table 2 shows the standards reviewed by each group in each round of ratings.

**Table 2. Standards Rated by Each Table**

Group (Table Number)	Standards Rated in Each Round		
	Round 1	Round 2	Round 3
<b>Mathematics</b>			
1	1. Numbers and Operations	2. Algebra	
2	2. Algebra	3. Geometry	
3	3. Geometry	4. Data and Probability	
4	4. Data and Probability	1. Numbers and Operations	
<b>Language Arts</b>			
1	1. Writing	2. Comprehension	
2	2. Comprehension	4. Connections	1. Writing
3	1. Writing	4. Connections	2. Comprehension

### Vertical Alignment Ratings

The alignment ratings were collected in two phases. First, panelists in each grade range subgroup reviewed the GLEs for a subset of grades as shown in Table 3. The middle and high school panelists were asked to rate one or two grades below their normal grade levels to provide overlap in the ratings at key transition points, which were thus labeled as consensus grades. Next the three subgroups at each table came together and made consensus ratings for the transition grade pairs (4 to 5 and 7 to 8).

**Table 3. Grade Ranges Rated by Each Teacher**

Subgroup	Grade Linkages Rated							
	3 to 2	4 to 3	5 to 4	6 to 5	7 to 6	8 to 7	9 to 8	10 to 9
Elementary	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>				
Middle			<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>		
High School					<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>

Note: Columns shown in bold are consensus grades.

## Rating Forms

Because of the nature of each content area, different forms were developed for English language arts and mathematics. For mathematics, the rating forms generally followed the design described in the CCSSO vertical alignment concept paper. For each grade, panelists were asked to identify one or possibly two expectations at the previous grade that were most related to each expectation for the target grade. Reference sheets were provided that listed all the expectations by grade and standard as shown in the examples provided in appendices D and E. If no related expectations were found, the panelist recorded “new”. Where a matching expectation was found, panelists were asked to rate

- the **Nature** of the increase in expectation from one grade to the next (Broadened, Deepened, Same, or New)
- the **Importance** of the increase in expectation
- the **Clarity** of the increase in expectation

Appendix C contains an example of the ratings sheets used for mathematics. Additional information about the scales used for these ratings is provided in the training materials included in appendix F.

For English language arts, the expectations across grades are similar. Most of the significant differences from grade to grade are the complexity of text. Because many of the expectations for adjacent grades are otherwise identically worded, it was not necessary to ask the panelists to locate matching expectations. Instead, panelists were shown matching standards for the adjacent grades where a particular expectation was worded differently from the corresponding expectation at the previous grade. The rating forms for English language arts showed the matching expectations at the two adjacent grades and ask for ratings of the nature, importance, and clarity of the differences. See appendix B for an example of an ELA Rating Sheet.

## Discussions of Additional Dimensions for Grade-Level Expectations

While grade-level expectations had been developed for purposes of assessment and accountability, there were aspects of the standards for each content area where variation by grade was not specific. For language arts, the expectations consistently referred to grade-level texts, but did not include descriptions of the specific ways in which texts might vary across grades. For mathematics, there are four process standards in addition to the four content standards. For process standards such as problem solving or mathematical reasoning, expectations are the same for all grades, but the complexity with which these standards are assessed, as evidenced by specific test questions, increase from one grade to the next.

After the vertical alignment ratings were completed, panelists in each subject area met as a group to discuss ways of characterizing increased complexity demands in their subject area. The language arts panelists reviewed reading passages used in assessments at different grade levels. The mathematics panelists reviewed mathematical reasoning problems from different grade levels. A brief description of these discussions is provided here although it was not a formal part of the vertical alignment process.

**Reading.** To capture the complexity of grade-level reading materials and provide information of increased complexity from grade to grade, a session called Rating the Narrative Complexity of Materials was developed by a consultant, Dr. Charles W. Peters, with the content staff of the Delaware Department of Education. The purpose of the narrative complexity scale was to capture such changes in complexity as themes, setting, characters, plot, and author’s craft (e.g., figurative language, rhetorical and literary devices) that occurs across grade levels. The English language arts panelists were grouped with grade level to compare selected reading passages between grades 2 and 4, 4 and 6, 6 and 8, and 8 and 10. The participants were engaged in a three-hour activity to determine where and how the change in narrative complexity occurred and rate the complexity of narrative structures across various grade-level passages that appeared on the Delaware Student Testing Program reading test. They also compared the changes that occurred from one grade level to another to determine the changes in the narrative structure that occurred from level to level and how the narrative structure becomes more complex from lower grades to higher grades.

**Mathematics.** Unlike the content standards, the four process standards in mathematics use general language across grades rather than being described progressively from one grade to the next. Nevertheless, the process standards form the backbone of the mathematics grade-level expectations and should therefore not be ignored in discussions of vertical alignment. Therefore, an additional session was allotted to explorations of how the process standards evolve across the grades, which was led by a consultant, Dr. Linda Wilson, with the content staff of the Delaware Department of Education. Participants were grouped according to their teaching grade level (elementary, middle, or high) and the standard of mathematical reasoning chosen for consideration. Using

a very simple task of two sets of five pennies arranged in different configurations on two sides of a mat, the participants explored how this question would be addressed by children at very early grades, then how middle-grades students might approach it, and finally how a mathematician might write a proof. The example was used to show how reasoning develops in students from very informal precounting to formal proof. And then the participants discussed how the given task might be adapted for a grade level higher or lower than the grade for which it is presently intended.

The panel agreed that some of the aspects of reasoning that evolve over the grades should include

- 1) types of explanations required
- 2) number of clues provided to the student in the problem
- 3) number of steps required to solve the problem
- 4) level of difficulty of patterns
- 5) level of complexity of the problem
- 6) level of sophistication of the vocabulary used
- 7) abstraction
- 8) use of symbols and technical mathematical representations

A recommendation was made that more specificity be included in the GLE documents to include descriptions of the evolution of the process standards across the grade bands.

### ***Debriefing and Evaluation***

At the conclusion of the workshop, a debriefing session was held for each content area. The content leaders led the discussion and collected comments and reactions to the processes and outcomes of the workshop. Specific debriefing questions are shown in appendix H. In addition, panelists completed an evaluation form that provided more quantitative feedback on various aspects of the workshop. See appendix I.

### ***Analysis of Results***

Nature, importance, and clarity ratings were recorded for each standard reviewed by each of group of panelists. Groupwide ratings, generally just for the consensus grades, were entered by the table facilitators during the workshop. Ratings for additional grades were entered from the paper rating sheets completed by the grade-range subgroups. Results for grades 3 and 4 were taken from the elementary subgroup ratings, results for grades 6 and 7 were taken from the middle school subgroup ratings, and results from grades 9 and 10 were taken from the high school subgroup ratings. Analyses of these ratings are reported in the next chapter.

In addition to the results reported here, panelist comments and feedback were provided to the department in two ways. All of the rating booklets were returned to the department so that they could see the comments in their original form. Second, for mathematics, alignment charts showing how expectations were matched across grades were recorded in an Excel spreadsheet. This spreadsheet provided additional feedback to the department for use in organizing and summarizing information for the development of performance level descriptions.

## Chapter 3: Results

Results from analyses of the workshop data are presented in this chapter, beginning with analyses of the consistency of the main ratings as measured by the consistency in results across independent groups. This is followed by a discussion of specific findings from the nature, importance, and clarity ratings for each content area. Then it continues with a discussion of more detailed workshop results and concludes with a summary of the debriefing session comments.

### Agreement across Independent Groups

The vertical alignment ratings collected reflected consensus judgments at either the subgroup or whole table level. In this context, it was neither possible nor meaningful to examine agreement at the level of individual panelists. Instead, the design of the workshop review of each standard by two independent groups of raters, seated at different tables. Analyses of the consistency of rating results from the independent groups is presented here before turning to the substantive findings from the workshop.

#### Mathematics

Table 4 shows the relationship of judgments on the nature of linkage ratings across independent groups of panelists. Group 1 indicates the table of panelists who rated the standard on the first round and Group 2 indicates the group that rated that same standard on the second round of ratings. Across all standards and grades, the agreement was found for 164 of the 344 links (48 percent). Even though broadened and deepened are highly related in some cases, the panelists could only code the expectation with one category, broadened or deepened. If these two categories are combined into a single category of “extended” an additional 86 links would be added to the agreement count, yielding an agreement rate of 73 percent (250 of 344 links).

**Table 4. Agreement of Nature of Link Ratings from Independent Groups—Mathematics**

Group 1 Judgment	Group 2 Judgment				
	Broadened	Deepened	New	Same	Total
Broadened	61	45	13	4	123
Deepened	41	54	16	6	117
New	9	20	33	2	64
Same	16	8	0	16	40
Total	127	127	62	28	344

Note: 48 percent agreement (73 percent if broadened and deepened are combined )

Table 5 shows the relationship of judgments between round 1 and round 2 about the clarity increases in expectations from one grade to the next. Note that clarity was not rated for the target expectation if the panelists considered the expectation was new and not related to any expectations from the previous grade. The independent groups agreed on 189 out of 264 links (72 percent) largely because most expectations were rated as clear. For the most part, the independent groups identified clarity issues with different links. Since all links identified by any of the groups were forwarded for further review, exact agreement across groups was unnecessary. Because the links in expectations were reviewed by multiple groups, it was more likely that any significant problem would be flagged by at least one of these groups.

**Table 5. Agreement of Clarity of Link Ratings from Independent Groups—Mathematics**

Group1 Judgment	Group 2 Judgment			
	1. Not Clear	2. Minor	3. Clear	Total
1. Not Clear	1	1	18	20
2. Minor Issues	0	3	35	38
3. Clear	7	14	185	206
Total	8	18	238	264

Note: 72% overall agreement

Table 6 shows the relationship across independent groups of judgments about the Importance of increases in expectations from one grade to the next. Again, importance was not rated where an expectation was considered to be new and not related to any of the expectations at the previous grade. Agreement was found for only 173 of the 257 links (44 percent agreement). As with the clarity ratings, the primary use of this information was to flag instances where increases in expectations were judged to be of low importance

**Table 6. Agreement of Importance Ratings from Independent Groups—Mathematics**

Group 1 Judgment	Group 2 Judgment			
	1. Low	2. Medium	3. High	Total
1. Low	14	16	21	51
2. Medium	23	37	35	95
3. High	13	36	62	111
Total	50	89	118	257

Note: 44 percent overall agreement

Table 7 shows the percentage of agreement across independent groups on the lower-grade expectations matched to each expectation at the target grade. Overall the independent groups found the same matching expectation 62 percent of the time and found different matching expectations 21 percent of the time. For the remaining 17 percent of the target expectations, no matching lower-grade expectation was found by one or both of the independent groups. Across the four mathematics standards, the highest agreement found for Data and Probability (77 percent). The agreement was considerably lower for the Algebra expectations (48 percent) than the other standards.

**Table 7. Agreement on Expectation Matched across Independent Groups—Mathematics**

Standard	Number of Expectations	Agreement across Groups		
		% Same	% Different	% No Match
1. Number and Operations	120	61%	23%	21%
2. Algebra	81	48%	25%	27%
3. Geometry	108	67%	18%	16%
4. Data and Probability	47	77%	17%	6%
Total	356	62%	21%	17%

"No Match" means one group did not find a match while the other did.

### English Language Arts

Table 8 shows the relationship of judgments on the nature of linkage ratings across independent groups of panelists. It also shows the relationship between type of linkage judgments from the table that rated a given standard in the first round and a second table rating expectations from that same standard in round 2. For English language arts, only matched expectations were presented, so there was not a "new" category. A few panelists, however, insisted that some expectations were both broadened and deepened at the next grade and created a new response category ("both") to cover this situation.

Across all standards and grades, there was a clear relationship between the nature of link ratings from independent tables, with agreement for 284 of the 421 linkages (67 percent). Unlike mathematics, judges were better able to distinguish between broadening and deepening for the English language arts expectations.

As with mathematics, about 10 percent of the time, panelists did not see a clear difference in an expectation from one grade to the next and rated the type of linkage as the "same." Each of these instances was flagged for further review by the Department to see where further clarification of intended differences might be needed.



**Table 8. Agreement of Nature of Link Ratings from Independent Groups—English Language Arts**

Group 1 Judgments	Group 2 Judgments				
	Broaden	Deepen	Both	Same	Total
Broaden	84	30	4	15	133
Deepen	12	189	3	11	215
Both	8	27	0	6	41
Same	18	3	0	11	32
Total	122	249	7	43	421

Note: 67% agreement (77% if Broaden and Deepen are combined)

Table 9 shows the relationship across independent groups in ratings of the clarity of increases in expectations from one grade to the next. As with mathematics, different rating tables were likely to raise different issues about their understanding of increases in expectation from one grade to the next. The primary concern is not so much with agreement as with the likelihood that at least one panelist spotted an issue that might trouble other teachers. With more than one group of panelists reviewing each standard, the chances of finding potential problems were increased.

**Table 9. Agreement of Clarity of Link Ratings across Independent Groups—English Language Arts**

Round 1 Judgments	Round 2 Judgments			
	1. Not Clear	2. Minor	3. Clear	Total
1. Not Clear	6	13	32	51
2. Minor Issues	9	35	86	130
3. Clear	15	67	168	250
Total	30	115	286	431

Note: 48 percent overall agreement

Table 10 shows the relationship across independent groups of panelists about the importance of increases in expectation from one grade to the next. Across all standards and grades, there was agreement for only 193 of the 419 linkages (46 percent) for which importance ratings were provided. As with the clarity ratings, expectations where increases from one grade to the next were given low importance ratings were flagged for further review.

**Table 10. Agreement of Importance Ratings across Independent Groups—English Language Arts**

Round 1 Judgments	Round 2 Judgments			
	1. Low	2. Medium	3. High	Total
1. Low	3	2	14	19
2. Medium	6	46	67	119
3. High	39	98	144	281
Total	48	146	225	419

Note: 46% overall agreement.

### Summary of Agreement Rates

Overall agreement rates were modest. The concept paper called for using judges who were thoroughly familiar with the expectations being aligned. It was envisioned that panelists would either be teachers responsible for teaching to the targeted expectations or content experts who had participated in development of the content frameworks. While Delaware's overall content standards were well-established, the grade-level expectations were brand new. It was thus not possible to find people outside of the department who had established of familiarity with these expectations. This may have suppressed the agreement rates considerably.

Given only modest agreement, some caution is advised in interpreting the summary information presented next. The greater value of the workshop was very likely in the detailed feedback indicating where procedures might be improved and indicating areas where the grade-level expectations might be explained more fully.

## Nature of Alignment from One Grade to the Next

### Mathematics

Table 11 shows the distribution of the type of match ratings for each of the four mathematics content standards. Overall, about 20 percent of the expectations were judged to be new for the target grade. The frequencies of broaden versus deepen ratings were nearly equal at about 35 percent each, although recall that distinctions between these two categories were not highly reliable (as evidenced by the cross-table agreement analyses shown in table 4.). In roughly 10 percent of the cases, panelists could not see an increase in expectations from one grade to the next and rated the related expectations as the same. Almost by definition “grade-level” expectations should require more of students in each successive grade, so the panelists’ difficulty in seeing the intended increase is a potential problem. While relatively infrequent, this type of problem was slightly more prevalent for algebra expectations and less so for geometry expectations.

**Table 11. Nature of Increases in Mathematics Expectations for Each Standard**

Standard	Type of Match			
	% Broadened	% Deepened	% Same	% New
1. Number and Operations	40%	34%	9%	17%
2. Algebra	26%	33%	16%	24%
3. Geometry	36%	36%	5%	23%
4. Data and Probability	44%	41%	10%	5%
Total	36%	35%	10%	19%

Table 12 shows the distribution of the type of match ratings across grades. Most striking is the result that the percent of expectations judged to be “new” was much higher for the 9th grade expectations (40 percent) compared to other grades, suggesting a possible disconnect between middle and high school expectations. The biggest proportion of “same” judgments was found for the 8th grade expectations. Significant differences in the proportion of “same” judgments or “new” judgments from one grade to the next could lead to discontinuities in describing target performance levels. Where many expectations were perceived to be similar from one grade to the next, performance level descriptions could also end up being relatively similar. When a number of expectations are new, performance level descriptions are more likely to suggest large increases in competencies.

Finally, the expectations were thought to broaden more frequently at the 5th and 7th grade and deepen more frequently at the 6th and 10th grades.

**Table 12. Nature of Increases in Mathematics Expectations by Grade**

Grades	Number of Ratings	Type of Match			
		% Broadened	% Deepened	% Same	% New
2 to 3	34	41%	44%	12%	3%
3 to 4	45	38%	24%	16%	22%
4 to 5	54	56%	33%	7%	4%
5 to 6	47	21%	47%	13%	19%
6 to 7	58	47%	24%	7%	22%
7 to 8	55	29%	33%	18%	20%
8 to 9	47	21%	28%	9%	40%
9 to 10	16	13%	56%	6%	25%
Total	356	36%	35%	10%	19%

### English Language Arts

For language arts, panelists only rated matched pairs of expectations. Table 13 shows the distribution of type of match ratings for the pairs rated under each language arts standard. In a few cases, panelists disagreed with the match and reported that the target-grade expectation represented a new skill rather than an increase in the prior grade expectation. Nonetheless, relatively few expectations were rated as new (2 percent overall) compared to mathematics, where 19 percent of the expectations are rated as new. The most striking result is that the majority of the expectations for writing (Standard 1) were rated as “broadening” from one grade to the next. Students are expected to perform the same writing tasks but with an extended range of prompts. For the two reading standards, expectations were mostly judged to deepen from one grade to the next (66 percent for comprehension and 72 percent for connections). This suggests that cognitive requirements go from simple recognition to descriptions, understanding, analysis, and even evaluation as students move to higher grades.

**Table 13. Nature of Matches for Each Language Arts Standard**

Standard	Type of Match				
	% Broadened	% Deepened	% Both	% Same	% New
1. Writing	54%	28%	4%	12%	1%
2. Comprehension	18%	66%	5%	9%	2%
4. Connections	20%	72%	0%	5%	4%
Total	31%	53%	4%	9%	2%

Table 14 shows the distribution of the language arts nature of match ratings by grade. These results show a much higher proportion of “same” ratings for the 9th (28 percent) and 10th (36 percent) expectations. These ratings suggest that the panelists could not see differences in expectations that may have been intended, and so indicate the potential need for further explanation of the expectations at these grade levels.

**Table 14. Nature of Increases in Language Arts Expectations by Grade**

Grades	Number of Ratings	Type of Match				
		% Broadened	% Deepened	% Both	% Same	% New
2 to 3	90	32%	50%	7%	3%	8%
3 to 4	72	50%	22%	14%	14%	0%
4 to 5	106	42%	38%	9%	10%	0%
5 to 6	73	52%	45%	3%	0%	0%
6 to 7	68	38%	47%	15%	13%	0%
7 to 8	92	24%	52%	11%	13%	0%
8 to 9	79	8%	57%	5%	28%	3%
9 to 10	56	38%	27%	0%	36%	0%
Total	356	35%	43%	8%	12%	1%

### Ratings of the Importance of Increases in Expectation

#### Mathematics

Table 15 shows the distribution of the ratings of importance assigned to increases in expectation from one grade to the next for each mathematics content standard. A greater proportion of the grade-to-grade increases in expectation for geometry (58 percent) and number and operations (47 percent) were judged important compared to algebra (38 percent) and data analysis and probability (28 percent). As noted above, there was not a high level of agreement across independent tables with respect to the importance ratings, so some caution is needed in interpreting these results.

**Table 15. Importance of Expectation Increases for Each Mathematics Standard**

Standard	Number of Ratings	Importance		
		% Low	% Medium	% High
1. Number and Operations	199	19%	34%	47%
2. Algebra	120	21%	40%	39%
3. Geometry	178	15%	27%	58%
4. Data and Probability	89	22%	49%	28%
Total	586	19%	35%	46%

Table 16 shows the distribution of the ratings of importance assigned to increases in expectation from one grade to the next for each for each grade. Increased expectations were rated somewhat more important for the 6th grade (57 percent high importance) and notably lower at the 10th grade (only 27 percent high importance). This result suggests the need for further clarification of the 10th grade expectations in mathematics relative to the 9th grade expectations.

**Table 16. Importance of Increases in Mathematics Expectations by Grade**

Grades	Number of Ratings	Importance		
		% Low	% Medium	% High
2 to 3	34	21%	38%	41%
3 to 4	41	29%	29%	41%
4 to 5	53	17%	36%	47%
5 to 6	46	15%	28%	57%
6 to 7	54	17%	35%	48%
7 to 8	50	14%	52%	44%
8 to 9	36	8%	44%	47%
9 to 10	15	33%	40%	27%
Total	356	36%	35%	10%

### English Language Arts

Table 17 shows the distribution of importance ratings for each of the language arts standards. As with mathematics, the majority (57 percent) of the increases in expectation from one grade to the next were rated as high importance. Overall, expectations for the writing standard had higher ratings of increased importance than was the case for two reading standards.

**Table 17. Importance of Expectation Increases for Each Language Arts Standard**

Standard	Number of Ratings	Importance		
		% Low	% Medium	% High
1. Writing	272	4%	32%	64%
2. Comprehension	399	12%	33%	55%
4. Connections	84	11%	40%	49%
Total	755	9%	34%	57%

Table 18 shows the distribution of importance ratings for grade-to-grade increases in expectation, separately for each grade. Expectation increases at the lower grades were slightly less likely to be rated as high importance (47 percent for grades 3 and 4 compared to 63 percent overall).

Table 18. Importance of Increases in Language Arts Expectations by Grade

Grades	Number of Ratings	Importance		
		% Low	% Medium	% High
2 to 3	85	5%	48%	47%
3 to 4	72	0%	53%	47%
4 to 5	106	0%	32%	68%
5 to 6	73	8%	34%	58%
6 to 7	68	12%	13%	75%
7 to 8	92	4%	22%	74%
8 to 9	78	7%	24%	69%
9 to 10	47	4%	28%	68%
Total	621	5%	32%	63%

### Clarity of Alignment of Expectations

#### Mathematics

The final summary of results concerns the ratings of the clarity of increases in related expectations from one grade to the next. Table 19 shows the clarity ratings for each mathematics standards. Overall, the panelists rated the increases as having high clarity (84 percent). Low clarity ratings were generally accompanied by specific comments or suggestions. Low ratings were slightly more frequent for geometry expectations (10 percent compared to 4–5 percent for other mathematics standards).

Table 19. Clarity of Expectation Increases for Each Mathematics Standard

Standard	Number of Ratings	Clarity		
		% Low	% Medium	% High
1. Number and Operations	196	4%	9%	87%
2. Algebra	133	5%	12%	83%
3. Geometry	183	10%	7%	83%
4. Data and Probability	87	5%	16%	79%
Total	599	6%	10%	84%

Table 20 shows the distribution of the clarity ratings for the mathematics expectations at each grade level. There was a higher proportion of “low clarity” ratings for grades 4, 8, and 9 (14 percent, 16 percent, and 14 percent respectively).

Table 20. Clarity of Mathematics Expectation Increases by Grade

Grades	Number of Ratings	Clarity		
		% Low	% Medium	% High
2 to 3	34	3%	24%	73%
3 to 4	42	14%	10%	76%
4 to 5	53	6%	15%	79%
5 to 6	46	2%	13%	85%
6 to 7	54	7%	6%	87%
7 to 8	49	16%	10%	73%
8 to 9	42	14%	14%	71%
9 to 10	15	7%	13%	80%
Total	335	9%	13%	79%

Note: In this table, each expectation is only counted once (335 expectations). In the previous table, multiple ratings of the same expectations were included (599 ratings).

### English Language Arts

Table 21 shows the distribution of clarity ratings for each language arts standard. There were significantly more low clarity ratings for the writing standard (18 percent compared to 8 percent and 5 percent for the two reading standards).

**Table 21. Clarity of Expectation Increases for Each Language Arts Standard**

Standard	Number of Ratings	Clarity		
		% Low	% Medium	% High
1. Writing	276	18%	26%	56%
2. Comprehension	409	8%	29%	63%
4. Connections	84	5%	19%	76%
Total	769	11%	27%	62%

Table 22 shows the distribution of clarity ratings for the language arts expectations at each grade. There was a much higher proportion of low clarity ratings at grades 9 and 10 (30 percent and 38 percent). This is very consistent with the result above that more of the 9th and 10th grade expectations were judged to be the “same” as the corresponding lower grade expectation.

**Table 22. Clarity of Language Arts Expectation Increases by Grade**

Grades	Number of Ratings	Clarity		
		% Low	% Medium	% High
2 to 3	89	25%	28%	47%
3 to 4	75	11%	15%	75%
4 to 5	106	7%	36%	58%
5 to 6	73	11%	37%	52%
6 to 7	69	3%	41%	57%
7 to 8	92	20%	23%	57%
8 to 9	80	30%	12%	58%
9 to 10	47	38%	30%	32%
Total	631	17%	28%	55%

### Detailed Workshop Results

In addition to the summary information described above, the workshop yielded two types of important results about individual expectations. First, we captured panelists’ comments about each specific expectation. Some of these comments applied to the expectation itself and others applied to the relationship of the expectation to the corresponding expectation at the prior grade. These comments were transmitted to the DOE for use in revising or explaining the current grade-level expectations.

The second detailed result was content maps for each of the mathematics standards. These maps show linkages of the expectations across grades. Spreadsheets including these maps have also been sent separately to the department. The maps might be used to reorder the expectations within a standard to increase the correspondence across grades.

The alignment linkages shown in the content maps raised some issues with the way in which mathematics expectations were grouped into “big idea” categories within each standard. Panelists sometimes linked an expectation in one category to a prior grade expectation in a different category. Sometimes the distinction within

a grade level was also not clear. For example, one of the 8th grade number and operations expectations in the first big idea category, number sense is

**8.105 Use proportional reasoning to solve problems.**

Another very similar expectation listed under the second big idea category, operations, is:

**8.117 Apply proportional reasoning strategies to solve real-world problems.**

Panelists did not appear to see a clear distinction between these two expectations and linked them to prior grade expectations in a different “big idea” category.

### **Debriefing Session**

The evaluation survey results show that 62 percent of the panelists in English language arts and 100 percent in Mathematics reported that the orientation and training had prepared them for the alignment workshop adequately or very well. Over 90 percent of the English language arts panelists and nearly 80 percent of the mathematics panelists felt comfortable or very comfortable to match expectations across grades and identify the type of match. More than 90 percent of the mathematics panelists were comfortable or very comfortable rating the importance of increased expectations; however, only 67 percent of the English language arts panelists reported in the same level. Similarly, about 80 percent of the panels in both subject areas were comfortable or very comfortable of rating the clarity of increased expectations across grades. According to the survey, over two thirds of the panelists in both English language arts and mathematics believed that the Grade-Level-Expectations are aligned from grade to grade.

Generally we received very positive feedback about the workshop. The majority of participants reported that the workshop had provided them with an opportunity to review the expectations not just for one grade but also the adjacent grades and discuss these expectations with fellow teachers who work in different grades. The alignment activities were “very helpful to listen to above, middle, and below grades about the concepts” and “very helpful for going back to teaching.” For many teachers, through the alignment process, they created a clear vision of aligning the expectations from one grade to the next. To improve the process and the accuracy of alignment, the definition of each category (e.g., deep, broader, same) should be clearer and content-specific examples should be used.

## Chapter 4: Recommendations

---

In this chapter, we present two types of recommendations. First, we provide recommendations for extending and improving the vertical alignment process. Second, we offer recommendations to the Delaware Department of Education for clarifying their grade level expectations and for using the alignment results in developing performance level descriptions.

### *Improvements to the Vertical Alignment Process*

The opportunity to conduct a full-scale pilot of the vertical alignment process was extremely valuable. The workshop extended the vertical alignment process developed by CCSSO in several important ways. Specific extensions that appear worth replicating are the following:

The addition of importance ratings. The ratings of the importance of specific increases in expectation from one grade to the next will help content experts focus on key areas in developing performance level descriptions. Also, low-importance ratings are another way of identifying expectations were further consideration or clarification may be needed. A more in-depth discussion with panelists of what is meant by the importance of the difference between standards in adjacent grades may improve the process and agreement rates.

Introducing the term “clarity.” The vertical alignment concept paper talks about the quality of the vertical alignment of content standards (grade-level expectations). For most panelists, quality is somewhat ambiguous in this context; clarity is not.

Focus on the importance and clarity more than the nature of linkages in expectations. Panelists continued to struggle to distinguish whether increased expectations represent broadening or deepening. The distinction did not yield practical implications for improving the standards and expectations. While better training might have helped panelists distinguish these characteristics, the more important characteristic is that standards that are broader or deeper are related in a particular way to standards at the lower grade. That is, they are neither “new” nor “the same.”

Logistical innovations. The selection and pre-training of table leaders proved vital to moving the process along. In addition, breaking each table into grade-range subgroups proved very effective in promoting greater participation by all panelists and in covering a large number of standards and expectations in a limited time. The focus on transition grades for the whole-table discussions provided a useful degree of interaction across grade ranges while limiting redundancy in the work of the different grade-range subgroups.

### *Suggestions for Delaware’s Grade-Level Expectations*

Notwithstanding relatively low agreement rates, a few of the specific findings about the nature and clarity of the vertical alignment of the grade-level expectations did stand out. These findings are summarized here by subject. A caveat: this was a pilot of a new procedure; therefore, the results must be interpreted with care. As noted earlier, individual panelist comments are likely to be the most useful source of information for reviewing the grade level expectations.

For mathematics, the large percent of expectations judged to be new at grade 9 (40 percent) suggests a possible disconnect between the nature and wording of the expectations for middle school grades and for high school. In addition, a significant proportion (33 percent) of the linkages from grade 9 to grade 10 mathematics were judged to be of low importance. Many of the linkages (29 percent) between mathematics expectations at the 3rd and 4th grade levels were also judged to be of low importance.

For language arts, a significant number (36 percent) of the expectations for grades 9 and 10 were judged the same, and a significant proportion (38 percent) of the linkages in expectations for these grades were judged to be of low clarity. Expectations for increases in knowledge and skill from the 8th to the 9th grade were also judged to be of lower clarity (28 percent were judged the same and 30 percent received low clarity ratings).

While the findings above have focused on areas where clarification of the grade-level expectations might be fruitful, it is important to keep in mind that most of the increases in grade-level expectations were judged to be clear and important. Overall, the results were fairly positive for the first exposure of these expectations.



Nonetheless, we offer some suggestions for further work in introducing and clarifying Delaware's grade-level expectations for mathematics and language arts.

**Recommendation 1. Continue work to clarify the grade-level expectations, particularly as they are introduced to new teachers.**

One valuable outcome of this workshop was the identification of specific questions that the panelists had about the wording and meaning of some of the expectations. These questions should be helpful in preparing additional explanatory material to clarify questions that teachers new to the grade-level expectations may have.

**Recommendation 2. Some version of this process might also provide effective professional development for introducing the grade-level expectations to the teachers who must help students meet them.**

Several teachers commented that it was very enlightening to review and discuss expectations for students at grade levels below and above the students they teach.

---

## References

- Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education*. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6. Madison: University of Wisconsin, Wisconsin Center for Education Research.
- Wise, L. L and Alt, M. (2005) Assessing vertical alignment. HumRRO Technical Report FR 95-05. Alexandria, VA: Human Resources Research Organization.

## Appendix A: Delaware Content Standards

### English

Standard Code	Description.
E02	Students will construct, examine, and extend the meaning of literary, informative, and technical texts through listening, reading, and viewing.
E04	Students will use literary knowledge accessed through print and visual media to connect self to society and culture.

### Writing

Standard Code	Description.
W01	Students will use written and oral English appropriate for various purposes and audiences.

### Mathematics Reasoning

Standard Code	Description.
M01	Students will develop their ability to SOLVE PROBLEMS by engaging in developmentally appropriate problem-solving opportunities in which there is a need to use various approaches to investigate and understand mathematical concepts; to formulate their own problems; to find solutions to problems from everyday situations; to develop and apply strategies to solve a wide variety of problems; and to integrate mathematical reasoning, communication and connections.
M02	Students will develop their ability to COMMUNICATE MATHEMATICALLY by solving problems in which there is a need to obtain information from the real world through reading, listening and observing; to translate this information into mathematical language and symbols; to process this information mathematically; and to present results in written, oral and visual formats.
M03	Students will develop their ability to REASON MATHEMATICALLY, by solving problems in which there is a need to investigate significant mathematical ideas in all content areas; to justify their thinking; to reinforce and extend their logical reasoning abilities; to reflect on and clarify their own thinking; to ask questions to extend their thinking and to construct their own learning..
M04	Students will develop their ability to MAKE MATHEMATICAL CONNECTIONS by solving problems in which there is a need to view mathematics as an integrated whole and to integrate mathematics with other disciplines, while allowing the flexibility to approach problems, from within and outside mathematics, in a variety of ways.

### Mathematics Understanding

Standard Code	Description.
M05	Students will develop an understanding of ESTIMATION, MEASUREMENT, and COMPUTATION by solving problems in which there is a need to measure to a required degree of accuracy by selecting appropriate tools and units; to develop computing strategies and select appropriate methods of calculation from among mental math, paper and pencil, calculators or computers; to use estimating skills to approximate an answer and to determine the reasonableness of results.
M06	Students will develop NUMBER SENSE by solving problems in which there is a need to represent and model real numbers verbally, physically and symbolically; to use operations with understanding; to explain the relationships between numbers; to apply the concept of a unit; and to determine the relative magnitude of real numbers.
M07	Students will develop an understanding of ALGEBRA by solving problems in which there is a need to progress from the concrete to the abstract using physical models, equations and graphs; to generalize number patterns; and to describe, represent and analyze relationships among variable quantities.

M08	Students will develop SPATIAL SENSE and an understanding of GEOMETRY by solving problems in which there is a need to recognize, construct, transform, analyze properties of, and discover relationships between, geometric figures.
M09	Students will develop an understanding of STATISTICS AND PROBABILITY by solving problems in which there is a need to collect, appropriately represent, and interpret data; to make inferences or predictions; to present convincing arguments; and to model mathematical situations to determine the probability.
M10	Students will develop an understanding of PATTERNS, RELATIONSHIPS AND FUNCTIONS by solving problems in which there is a need to recognize and extend a variety of patterns; and to analyze, represent, model and describe real-world functional relationships.

## Appendix B: Sample Rating Sheet for English Language Arts

### Matching Expectations for Indicator 1.5: Persuasive

Grade	Expectation Number	Expectation
3	3.1.501	Take a position on an easily understood debatable “issue” or “question”
2	2.1.501	Take a position on an easily understood debatable “issue” that can initially be answered as a “yes” or “no”

Nature of Linkage Code(s): \_\_\_\_\_ (**B**roaden, **D**eepen, **S**ame, **N**ew)

Importance: \_\_\_\_\_ (**H**igh, **M**edium, **L**ow)

Quality of the Linkage Rating: \_\_\_\_\_ (**3**-High Quality, **2**-Minor Ambiguity, **1**-Not Clear)

Source(s) of ambiguity:

☐ Different words are used for the same knowledge or skill

Which words? \_\_\_\_\_

☐ The same word(s) is (are) used in different ways

Which words? \_\_\_\_\_

☐ Differences between the two standards are not clear.

(They appear to be essentially the same.)

☐ The higher grade standard is unclear or imprecise.

☐ The lower grade standard is unclear or imprecise

☐ Other (explain):

\_\_\_\_\_  
\_\_\_\_\_

## Appendix C: Sample Rating Sheet for Mathematics

### Topic: 1a Number Sense

#### Expectation:

#### 3.101 CONNECT SKIP COUNTING TO MULTIPLICATION

Most Similar Grade 2 Expectation (Number from Grade 2, Sheet 1a) \_\_\_\_\_

Importance: \_\_\_\_\_ (High, Medium, Low)

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deepen, Same, New)

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear)

Source(s) of ambiguity:

- ☐ Different words are used for the same knowledge or skill  
Which words? \_\_\_\_\_
- ☐ The same word(s) is (are) used in different ways  
Which words? \_\_\_\_\_
- ☐ Differences between the two standards are not clear.  
(They appear to be essentially the same.)
- ☐ The higher grade standard is unclear or imprecise.
- ☐ The lower grade standard is unclear or imprecise
- ☐ Other (explain):  
\_\_\_\_\_  
\_\_\_\_\_

Alternate Grade 2 Standard (No.): \_\_\_\_\_

Importance: \_\_\_\_\_ (High, Medium, Low)

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deepen, Same, New)

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality to 1-Not Clear)

Source(s) of ambiguity:

- ☐ Different words are used for the same knowledge or skill  
Which words? \_\_\_\_\_
- ☐ The same word(s) is (are) used in different ways  
Which words? \_\_\_\_\_
- ☐ Differences between the two standards are not clear.  
(They appear to be essentially the same.)
- ☐ The higher grade standard is unclear or imprecise.
- ☐ The lower grade standard is unclear or imprecise
- ☐ Other (explain):  
\_\_\_\_\_  
\_\_\_\_\_

## Appendix D: Sample Reference Sheet for English Language Arts

### Standard 1.

Students will use written and oral English appropriate for various purposes and audiences.

### Performance Indicator 1.1

Writers will produce examples that illustrate the following discourse classifications; by the completion of the grade, writers will be able to write **persuasive** (audience-oriented), **informative** (subject-oriented), and **expressive** (author-oriented) texts.

### Grade 2, Reference Sheet 1.1 - Process

Number	Text
2.1.101	Writers use and adapt the <b>writing process</b> (planning/brain storming/organizing, drafting, conferring, revisiting, editing, sharing, and publishing with ongoing revision throughout the process)
2.1.102	Writers engage in a writing process that is recursive (not linear, but spiraling or cyclical) and individual
2.1.103	Writers recognize that not all writing pieces go to the publication stage
2.1.104	Writers engage in self-assessment and reflection on their writing
2.1.105	Writers select their own topics and occasions for writing and also write in "on-demand" (prompted) occasions.
2.1.106	Writers engage in collaborative writing (e.g., "buddy writing," guided writing, interactive writing, shared writing)

### Grade 2, Reference Sheet 1.2 - Purpose

Number	Text
2.1.201	Students understand that <b>persuasive</b> writing is audience-centered; the needs and perspective of the intended audience are the most important consideration. Students understand that persuasive writing involves taking a position on a debatable issue to convince an audience.
2.1.202	Students understand that <b>informative</b> writing is subject-centered; the need to communicate information clearly so that the audience can understand the content/subject is the most important consideration.
2.1.203	Students understand that <b>expressive</b> writing is author-centered; the most important consideration is the writer's intent to tell a story or make meaning of an experience (reflection, self-discovery), to achieve personal goals, or to create literary pieces.

## Appendix E: Sample Reference Sheet for Mathematics

---

### Standard 1: Number and Operations Reference Sheet

#### Grade 2 Sheet 1a: Number Sense

##### No. Text of Grade-Level Expectation

---

*Building upon the K–1 expectations, all students in Grade 2 will be able to:*

#### NUMBER SENSE


- 2.101** Use multiple strategies for counting using groups of 1s, 5s and 10s.
- 2.102** Connect number words for fractions ( $\frac{1}{2}$ 's and  $\frac{1}{4}$ 's) with pictures and numerals.
- 2.103** Compare size of two numbers by counting or counting back.
- 2.104** Use combinations of one and two-digit numbers to build larger (2 digit) numbers

#### Grade 2 Sheet 1b: Operations

#### OPERATIONS

- 2.105** Use number sentences to represent number combinations up to 20.
- 2.106** Use number sentences with missing addends to represent subtraction combinations up to 20.
- 2.107** Use a variety of strategies to model combination and separation problems up to 100.
- 2.108** Show number sentences that demonstrate that addition and subtraction are inverse operations (e.g., join, separate, part-part-whole, compare).
- 2.109** Represent repeated addition using pictures and models.
- 2.110** Understand that addition of whole numbers result in a larger number and subtraction of whole numbers result in a smaller number.

## Appendix F: Table Leader Orientation Slides




# **Vertical Alignment Workshop Table Leader Orientation**

Lauress Wise  
Human Resources Research Organization (HumRRO)

Delaware Vertical Alignment Workshop  
April 20, 2005

Human Resources Research Organization



# **Workshop Goals**

- **Context: States moving to meet NCLB requirements**
  - Define Reading, Mathematics, and Science Objectives for every grade (3-8 and High School), not just ends of grade ranges.
  - Administer assessments to all (at least 95%) students in every grade and set performance standards for proficiency.
- **Goal of the Vertical Alignment Workshop**
  - Take a detailed look at expectations for each grade level
  - Determine what "More" is expected at each grade
  - Begin to describe most important increases in expectation from one grade to the next
    - ✓ Performance level descriptions for each grade
    - ✓ Used in setting performance standards for each grade
    - ✓ Need to achieve consistent expectations of growth

April 20, 2005      Overview of Vertical Alignment      Human Resources Research Organization      2



## Comparing Expectations Across Grades

### What you will be asked to do:

1. Match each expectation to one or more expectations at the next lower grade
  - Language Arts expectations are already matched.
2. Describe how the matching expectations are related
3. Rate the importance of differences in expectation from one grade to the next
4. Rate the clarity of the differences in expectations and identify ambiguities in the grade-level expectations

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

3

## How are Matching Expectations Related?

- Types of Relationships
  - Knowledge or skills **Broadened** to wider range of content
    - ✓ Same skills applied to wider content
  - **Deeper** understanding (cognitive processes) for the same content
    - ✓ Watch the verbs (recognize => explain)
  - **New** (or different) content and/or skills
    - ✓ No matching expectation at next lower grade

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

4

## Importance of Differences

- We want to identify the most important differences in expectations from one grade to the next
  - **High** – differences in matched expectations are central to what “More” we want students to know and be able to do
  - **Medium** – differences are important, but not the most central
  - **Low** – differences are trivial or less important
- Ratings are individual judgments – there are no “Right” answers

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

5

## Quality of Content Alignment

- Content standards are **not** clearly articulated across grades if:
  - Related standards are not clearly differentiated.
    - ✓ What new knowledge or skill is required?
    - ✓ One or both standards may not be described in sufficient detail.
  - Differences in terminology are not explained.
    - ✓ Different words for the same skill?
  - Terminology drifts.
    - ✓ The meaning of terms appears to be expanded.

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

6

## Reference Sheets - Math

### Grade-Level Expectations: Grade 2 Mathematics Standard 1: Number and Operations Reference Sheet Grade 2 Sheet 1a: Number Sense

#### No. Text of Grade-Level Expectation

*Building upon the K-1 expectations, all students in Grade 2 will be able to:*

#### **Number Sense**

- 2.101 Use multiple strategies for counting using groups of 1s, 5s and 10s.
- 2.102 Connect number words for fractions ( $\frac{1}{2}$ 's and  $\frac{1}{4}$ 's) with pictures and numerals.
- 2.103 Compare size of two numbers by counting or counting back.
- 2.104 Use combinations of one and two-digit numbers to build larger (2 digit) numbers

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

7

## Reference Sheets – Language Arts

### Grade-Level Expectations: Grade 2 Reading (Continued)

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

#### **Grade 2, Reference Sheet 2.04 – Retelling or Restating**

**Performance Indicator 2.04 [2.4d (2-8) 2.3d (9-10)]:** Students will be able to demonstrate an overall understanding of printed texts by (d) **retelling** a story or **restating** an informative text through speaking and/or writing.

*By the end of Grade 2, using 2nd grade or higher texts, students know and are able to do everything required in previous grades and*

#### **Expectation**

<u>Number</u>	<u>Text</u>
2.2.0401	Retell the story, identifying the main characters and major events in a simple literary text
2.2.0402	<b>Restate</b> the main idea of a simple informative text with supporting details
2.2.0403	Identify (in sequence) the major events in a story

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

8

## Rater Training

- **Task 1: Identify matching objective from next lower grade (if any)**

Most Similar Grade *n*-1 Objective (No.): Enter Objective Number

- [For Language Arts, matching objective is already identified]

- **Task 2: Code the relationship of the matching objectives**

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deepen, Same, New)

- **Task 3: Rate importance of differences**

Importance: \_\_\_\_\_ (High, Medium, Low)

- **Task 4: Rate quality of linkage and describe any source of ambiguity**

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear), Source(s) of ambiguity:

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

9

## 1. Match Objectives

- Each Rating Sheet has a specific objective for the current target grade and standard (content area).
- Panelists are given a “Reference Sheet” listing **all** of the prior grade objectives for that standard.
- Panelists are asked to identify up to two objectives from the Reference Sheet that best match the target objective on the Rating Sheet.

- Enter the number for the matching prior-grade objective on the rating sheet. For example, if the matching 6<sup>th</sup> grade objective were 6.203, you would enter:

Most Similar Grade 6 Objective (No.): 6.203

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

10

## 2. Code Relationship

### ➤ Code one of four types of relationships:

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deepen, Same, New)

- **Broaden** – content is broadened. For example
  - ✓ Integer operations are extended to whole numbers
  - ✓ Reading skills are applied to more complex texts
- **Deepen** – a deeper level of cognitive skill is required:
  - ✓ **Level 1: Recall**
  - ✓ **Level 2: Perform a simple task** (“classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data”)
  - ✓ **Level 3: Strategic thinking** (reasoning, planning, using evidence, and a higher level of thinking than the previous two levels)
  - ✓ **Level 4: Extended thinking** (complex reasoning, planning, developing, and thinking most likely over an extended period of time.)
  - ✓ **Verbs:** recognize->understand->explain->analyze->evaluate
- **Same** – the same knowledge and skill are required.
- **New** – there is no matching objective for the prior grade.

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

11

## More on Depth of Knowledge

### ➤ Comparison of Webb, Porter, and Bloom taxonomies for cognitive complexity:

Webb	Porter	Bloom
Recall	Memorize	Recall of Data Comprehension
Simple Procedures	Perform Procedures Demonstrate Understanding	Application
Strategic Thinking	Conjecture, Generalize Prove	Analysis
Extended Thinking	Solve non-routine problems	Synthesis Evaluation

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

12

### 3. Rate Importance

Importance: \_\_\_\_\_ (High, Medium, Low)

- High – Very important difference
  - Example: Perform one-digit addition/subtraction is Broadened to Perform one-digit addition/subtraction with three digit numbers
- Medium – Important difference
  - “Begin to identify information in a simple text to develop an opinion” versus “Identify information in a simple text to develop an opinion”
  - Difference is important, but not large (or Large, but less important)
- Low – Not an important difference
- Your own opinions - no “Right” answers
- Try not to rate everything at the same level

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

13

### 4. Rate Linkage Quality

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear)

Source(s) of ambiguity checklist:

- \_\_\_ Different words are used for the same knowledge or skill.  
Which words? \_\_\_\_\_
- \_\_\_ The same word(s) is (are) used in different ways.  
Which words? \_\_\_\_\_
- \_\_\_ Differences between the two standards are not clear.  
(They appear to be essentially the same.)
- \_\_\_ The higher grade standard is unclear or imprecise.
- \_\_\_ The lower grade standard is unclear or imprecise.
- \_\_\_ Other (explain): \_\_\_\_\_

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

14

## Judgment Task Example 1

### Grade 7 Estimation and Measurement Objective:

7.202 Use physical models to find the volume and surface area of cubes, prisms, and cylinders.

### Matches which Grade 6 Objective?

6.201 Estimate, measure, and classification angles.

6.202 Measure and find the ratio of the circumference and diameter of circular objects to estimate pi.

6.203 Use physical models to find the area and perimeter of rectangles and triangles.

6.204 Demonstrate an understanding of when to use a unit, a square unit, and a cubic unit.

6.205 Use equivalent fractions to solve problems.

6.206 Make estimates using benchmark fractions and decimals and determine if the estimate is reasonable.

(+4 other, less related, Grade 6 objectives for Estimation and Measurement)

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

15

## Judgment Task Example 1 (Con't)

**Grade 7 Objective: 7.202 Use physical models to find the volume and surface area of cubes, prisms, and cylinders**

**Most Similar Grade 6 Objective (No.): 6.203**

**Nature of Linkage Code(s):** \_\_\_\_\_ (Broaden, Deepen, Same, New)

**Importance:** \_\_\_\_\_ (High, Medium, Low)

**Quality of the Linkage Rating:** \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear)

**Source(s) of ambiguity:**

\_\_\_ Different words are used for the same knowledge or skill.

Which words? \_\_\_\_\_

\_\_\_ The same word(s) is (are) used in different ways.

Which words? \_\_\_\_\_

\_\_\_ Differences between the two standards are not clear.

(They appear to be essentially the same.)

\_\_\_ The higher grade standard is unclear or imprecise.

\_\_\_ The lower grade standard is unclear or imprecise.

\_\_\_ Other (explain): \_\_\_\_\_

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

16

## Judgment Task Example 2

### Grade 4 Science History/Nature Objective:

- 4.1.1 Contrast changes in scientific knowledge resulting from new discoveries (e.g., new knowledge leads to new questions).

### Matches which Grade 3 Objective?

- 3.1.1 Recognize that scientific explanations may lead to new discoveries (e.g., new knowledge leads to new questions).  
 3.1.2 Study the lives and discoveries of scientists of different cultures and backgrounds.  
 5.1.3 Explore science careers in the community.

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

17

## Judgment Task Example 2 (Con't)

Grade 4 Objective: **4.1.1 Contrast changes in scientific knowledge resulting from new discoveries (e.g., new knowledge leads to new questions).**

Most Similar Grade 3 Objective (No.): **3.1.1**

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deepen, Same, New)

Importance: \_\_\_\_\_ (High, Medium, Low)

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear)

Source(s) of ambiguity:

\_\_\_ Different words are used for the same knowledge or skill.

Which words? \_\_\_\_\_

\_\_\_ The same word(s) is (are) used in different ways.

Which words? \_\_\_\_\_

\_\_\_ Differences between the two standards are not clear.

(They appear to be essentially the same.)

\_\_\_ The higher grade standard is unclear or imprecise.

\_\_\_ The lower grade standard is unclear or imprecise.

\_\_\_ Other (explain): \_\_\_\_\_

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

18



## Judgment Task Example 3

### Grade 6 Reading Objective:

- 6.1.3 Determine theme and locate supporting details in a literary passage and across the curriculum

### Matches which Grade 5 Objective?

- 5.1.1 Identify defining characteristics, build background knowledge and develop reading skills to understand a variety of literary passages and texts (e.g., fiction; nonfiction; myth; poems; fantasies; biographies; science fiction, tall tales; supernatural tales).
- 5.1.2 Increase amount of independent reading.
- 5.1.3 Determine main idea and locate supporting details in a literary passage and across the curriculum.
- 5.1.4 Analyze text to determine time and sequence.
- 5.1.5 Use comprehension skills (e.g., draw conclusions; predict; use context clues; summarize).
- (+ 9 other, less related, objectives for Grade 5 Reading)

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

19

## Judgment Task Example 3 (Con't)

What are the main differences between:

- 5.1.3 Determine main idea and locate supporting details in a literary passage and across the curriculum.

and

- 6.1.3 Determine theme and locate supporting details in a literary passage and across the curriculum.

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

20

## Judgment Task Example 3 (Con't)

Grade 6 Standard: **6.1.3 Determine theme and locate supporting details in a literary passage and across the curriculum**

Most Similar Grade 5 Objective (No.): **5.1.3**

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deepen, Same, New)

Importance: \_\_\_\_\_ (High, Medium, Low)

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear)

Source(s) of ambiguity:

\_\_\_ Different words are used for the same knowledge or skill.

Which words? \_\_\_\_\_

\_\_\_ The same word(s) is (are) used in different ways.

Which words? \_\_\_\_\_

\_\_\_ Differences between the two standards are not clear.

(They appear to be essentially the same.)

\_\_\_ The higher grade standard is unclear or imprecise.

\_\_\_ The lower grade standard is unclear or imprecise.

\_\_\_ Other (explain): \_\_\_\_\_

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

21

## Rating Process

- You will be assigned to one of three or four tables.
- Each table begins with a different standard
- Grade-Span Subgroups (Elementary, Middle, and High School) work separately at first (Round 1)
  - Paper-and-pencil ratings, using Grade-Span Rating sheets
- Tables come together to discuss results across all grade spans (Round 2)
  - Results entered into spreadsheets on laptops
- Continue with 2<sup>nd</sup> Standard (Rounds 3 and 4)

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

22

## Role of the Table Leaders

- Keep discussion moving and on topic
  - Limit criticism of expectations themselves (focus on alignment)
  - Help remove impediments to consensus
- Encourage participation by all panelists
  - Don't let one person dominate; ensure "equal time"
  - Promote balance across grade-span teams
- Clarify concerns and seek help as needed
  - Help articulate questions to content leaders or HumRRO staff
- Participate and enjoy!
  - You are playing an important role in clarifying expectations for all Delaware students.

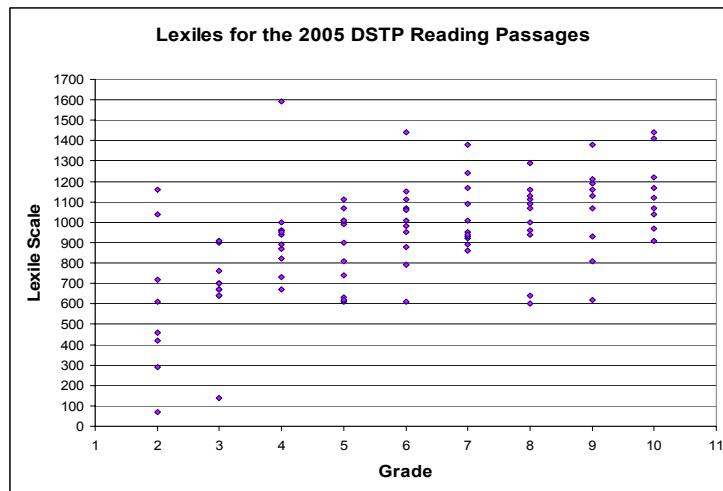
April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

23

## Day 2: Passage Complexity



April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

24

## ***Discussion***

---

### **➤ Questions and suggestions about:**

- Why vertical alignment is needed? How results can be used?
- The proposed process and how it works.
- Other.

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

25

## Appendix G: Vertical Alignment Workshop Training Slides




# Vertical Alignment Workshop

Lauress Wise  
Human Resources Research Organization (HumRRO)

Delaware Vertical Alignment Workshop  
April 20, 2005

Human Resources Research Organization



# Workshop Goals

- **Context: States moving to meet NCLB requirements**
  - Define Reading, Mathematics, and Science Objectives for every grade (3-8 and High School), not just ends of grade ranges.
  - Administer assessments to all (at least 95%) students in every grade and set performance standards for proficiency.
- **Goal of the Vertical Alignment Workshop**
  - Take a detailed look at expectations for each grade level
  - Determine what “More” is expected at each grade
  - Begin to describe most important increases in expectation from one grade to the next
    - ✓ Performance level descriptions for each grade
    - ✓ Used in setting performance standards for each grade
    - ✓ Need to achieve consistent expectations of growth

April 20, 2005      Overview of Vertical Alignment      Human Resources Research Organization      2

## Comparing Expectations Across Grades

### What you will be asked to do:

1. Match each expectation to one or more expectations at the next lower grade
  - Language Arts expectations are already matched.
2. Describe how the matching expectations are related
3. Rate the importance of differences in expectation from one grade to the next
4. Rate the clarity of the differences in expectations and identify ambiguities in the grade-level expectations

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

3

## How are Matching Expectations Related?

### ➤ Types of Relationships

- Knowledge or skills **Broadened** to wider range of content
  - ✓ Same skills applied to wider content
- **Deeper** understanding (cognitive processes) for the same content
  - ✓ Watch the verbs (recognize => explain)
- **New** (or different) content and/or skills
  - ✓ No matching expectation at next lower grade

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

4

## Importance of Differences

- We want to identify the most important differences in expectations from one grade to the next
  - High – differences in matched expectations are central to what “More” we want students to know and be able to do
  - Medium – differences are important, but not the most central
  - Low – differences are trivial or less important
- Ratings are individual judgments – there are no “Right” answers

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

5

## Quality of Content Alignment

- Content standards are **not** clearly articulated across grades if:
  - Related standards are not clearly differentiated.
    - ✓What new knowledge or skill is required?
    - ✓One or both standards may not be described in sufficient detail.
  - Differences in terminology are not explained.
    - ✓Different words for the same skill?
  - Terminology drifts.
    - ✓The meaning of terms appears to be expanded.

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

6

## Reference Sheets - Math

### Grade-Level Expectations: Grade 2 Mathematics Standard 1: Number and Operations Reference Sheet Grade 2 Sheet 1a: Number Sense

#### No. Text of Grade-Level Expectation

*Building upon the K-1 expectations, all students in Grade 2 will be able to:*

#### **Number Sense**

- 2.101 Use multiple strategies for counting using groups of 1s, 5s and 10s.
- 2.102 Connect number words for fractions ( $\frac{1}{2}$ 's and  $\frac{1}{4}$ 's) with pictures and numerals.
- 2.103 Compare size of two numbers by counting or counting back.
- 2.104 Use combinations of one and two-digit numbers to build larger (2 digit) numbers

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

7

## Reference Sheets – Language Arts

### Grade-Level Expectations: Grade 2 Reading (Continued)

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

#### **Grade 2, Reference Sheet 2.04 – Retelling or Restating**

**Performance Indicator 2.04 [2.4d (2-8) 2.3d (9-10)]:** Students will be able to demonstrate an overall understanding of printed texts by (d) **retelling** a story or **restating** an informative text through speaking and/or writing.

*By the end of Grade 2, using 2nd grade or higher texts, students know and are able to do everything required in previous grades and*

#### **Expectation**

<u>Number</u>	<u>Text</u>
2.2.0401	Retell the story, identifying the main characters and major events in a simple literary text
2.2.0402	<b>Restate</b> the main idea of a simple informative text with supporting details
2.2.0403	Identify (in sequence) the major events in a story

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

8



## Rater Training

- **Task 1: Identify matching objective from next lower grade (if any)**

Most Similar Grade *n*-1 Objective (No.): Enter Objective Number

- [For Language Arts, matching objective is already identified]

- **Task 2: Code the relationship of the matching objectives**

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deepen, Same, New)

- **Task 3: Rate importance of differences**

Importance: \_\_\_\_\_ (High, Medium, Low)

- **Task 4: Rate quality of linkage and describe any source of ambiguity**

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear), Source(s) of ambiguity:

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

9

## 1. Match Objectives

- Each Rating Sheet has a specific objective for the current target grade and standard (content area).
- Panelists are given a "Reference Sheet" listing **all** of the prior grade objectives for that standard.
- Panelists are asked to identify up to two objectives from the Reference Sheet that best match the target objective on the Rating Sheet.
  - Enter the number for the matching prior-grade objective on the rating sheet. For example, if the matching 6<sup>th</sup> grade objective were 6.203, you would enter:

Most Similar Grade 6 Objective (No.): 6.203

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

10

## 2. Code Relationship

### ➤ Code one of four types of relationships:

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deepen, Same, New)

- **Broaden** – content is broadened. For example
  - ✓ Integer operations are extended to whole numbers
  - ✓ Reading skills are applied to more complex texts
- **Deepen** – a deeper level of cognitive skill is required:
  - ✓ **Level 1: Recall**
  - ✓ **Level 2: Perform a simple task** (“classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data”)
  - ✓ **Level 3: Strategic thinking** (reasoning, planning, using evidence, and a higher level of thinking than the previous two levels)
  - ✓ **Level 4: Extended thinking** (complex reasoning, planning, developing, and thinking most likely over an extended period of time.)
  - ✓ **Verbs:** recognize->understand->explain->analyze->evaluate
- **Same** – the same knowledge and skill are required.
- **New** – there is no matching objective for the prior grade.

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization 11

## More on Depth of Knowledge

### ➤ Comparison of Webb, Porter, and Bloom taxonomies for cognitive complexity:

Webb	Porter	Bloom
Recall	Memorize	Recall of Data Comprehension
Simple Procedures	Perform Procedures Demonstrate Understanding	Application
Strategic Thinking	Conjecture, Generalize Prove	Analysis
Extended Thinking	Solve non-routine problems	Synthesis Evaluation

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization 12

### 3. Rate Importance

Importance: \_\_\_\_\_ (High, Medium, Low)

- High – Very important difference
  - Example: Perform one-digit addition/subtraction is Broadened to Perform one-digit addition/subtraction with three digit numbers
- Medium – Important difference
  - “Begin to identify information in a simple text to develop an opinion” versus “Identify information in a simple text to develop an opinion”
  - Difference is important, but not large (or Large, but less important)
- Low – Not an important difference
- Your own opinions - no “Right” answers
- Try not to rate everything at the same level

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

13

### 4. Rate Linkage Quality

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear)

Source(s) of ambiguity checklist:

- \_\_\_ Different words are used for the same knowledge or skill.  
Which words? \_\_\_\_\_
- \_\_\_ The same word(s) is (are) used in different ways.  
Which words? \_\_\_\_\_
- \_\_\_ Differences between the two standards are not clear.  
(They appear to be essentially the same.)
- \_\_\_ The higher grade standard is unclear or imprecise.
- \_\_\_ The lower grade standard is unclear or imprecise.
- \_\_\_ Other (explain): \_\_\_\_\_

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

14

## Judgment Task Example 1

### Grade 7 Estimation and Measurement Objective:

7.202 Use physical models to find the volume and surface area of cubes, prisms, and cylinders.

### Matches which Grade 6 Objective?

6.201 Estimate, measure, and classification angles.

6.202 Measure and find the ratio of the circumference and diameter of circular objects to estimate pi.

6.203 Use physical models to find the area and perimeter of rectangles and triangles.

6.204 Demonstrate an understanding of when to use a unit, a square unit, and a cubic unit.

6.205 Use equivalent fractions to solve problems.

6.206 Make estimates using benchmark fractions and decimals and determine if the estimate is reasonable.

(+4 other, less related, Grade 6 objectives for Estimation and Measurement)

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

15

## Judgment Task Example 1 (Con't)

**Grade 7 Objective: 7.202 Use physical models to find the volume and surface area of cubes, prisms, and cylinders**

**Most Similar Grade 6 Objective (No.): 6.203**

**Nature of Linkage Code(s):** \_\_\_\_\_ (Broaden, Deepen, Same, New)

**Importance:** \_\_\_\_\_ (High, Medium, Low)

**Quality of the Linkage Rating:** \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear)

**Source(s) of ambiguity:**

\_\_\_ Different words are used for the same knowledge or skill.

Which words? \_\_\_\_\_

\_\_\_ The same word(s) is (are) used in different ways.

Which words? \_\_\_\_\_

\_\_\_ Differences between the two standards are not clear.

(They appear to be essentially the same.)

\_\_\_ The higher grade standard is unclear or imprecise.

\_\_\_ The lower grade standard is unclear or imprecise.

\_\_\_ Other (explain): \_\_\_\_\_

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

16

## Judgment Task Example 2

### Grade 4 Science History/Nature Objective:

- 4.1.1 Contrast changes in scientific knowledge resulting from new discoveries (e.g., new knowledge leads to new questions).

### Matches which Grade 3 Objective?

- 3.1.1 Recognize that scientific explanations may lead to new discoveries (e.g., new knowledge leads to new questions).  
 3.1.2 Study the lives and discoveries of scientists of different cultures and backgrounds.  
 5.1.3 Explore science careers in the community.

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

17

## Judgment Task Example 2 (Con't)

Grade 4 Objective: **4.1.1 Contrast changes in scientific knowledge resulting from new discoveries (e.g., new knowledge leads to new questions).**

Most Similar Grade 3 Objective (No.): **3.1.1**

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deepen, Same, New)

Importance: \_\_\_\_\_ (High, Medium, Low)

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear)

Source(s) of ambiguity:

\_\_\_ Different words are used for the same knowledge or skill.

Which words? \_\_\_\_\_

\_\_\_ The same word(s) is (are) used in different ways.

Which words? \_\_\_\_\_

\_\_\_ Differences between the two standards are not clear.

(They appear to be essentially the same.)

\_\_\_ The higher grade standard is unclear or imprecise.

\_\_\_ The lower grade standard is unclear or imprecise.

\_\_\_ Other (explain): \_\_\_\_\_

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

18

## Judgment Task Example 3

### Grade 6 Reading Objective:

- 6.1.3 Determine theme and locate supporting details in a literary passage and across the curriculum

### Matches which Grade 5 Objective?

- 5.1.1 Identify defining characteristics, build background knowledge and develop reading skills to understand a variety of literary passages and texts (e.g., fiction; nonfiction; myth; poems; fantasies; biographies; science fiction, tall tales; supernatural tales).
- 5.1.2 Increase amount of independent reading.
- 5.1.3 Determine main idea and locate supporting details in a literary passage and across the curriculum.
- 5.1.4 Analyze text to determine time and sequence.
- 5.1.5 Use comprehension skills (e.g., draw conclusions; predict; use context clues; summarize).
- (+ 9 other, less related, objectives for Grade 5 Reading)

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

19

## Judgment Task Example 3 (Con't)

What are the main differences between:

- 5.1.3 Determine main idea and locate supporting details in a literary passage and across the curriculum.

and

- 6.1.3 Determine theme and locate supporting details in a literary passage and across the curriculum.

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

20

## Judgment Task Example 3 (Con't)

Grade 6 Standard: **6.1.3 Determine theme and locate supporting details in a literary passage and across the curriculum**

Most Similar Grade 5 Objective (No.): **5.1.3**

Nature of Linkage Code(s): \_\_\_\_\_ (Broaden, Deepen, Same, New)

Importance: \_\_\_\_\_ (High, Medium, Low)

Quality of the Linkage Rating: \_\_\_\_\_ (3-High Quality, 2-Minor Ambiguity, 1-Not Clear)

Source(s) of ambiguity:

\_\_\_ Different words are used for the same knowledge or skill.

Which words? \_\_\_\_\_

\_\_\_ The same word(s) is (are) used in different ways.

Which words? \_\_\_\_\_

\_\_\_ Differences between the two standards are not clear.

(They appear to be essentially the same.)

\_\_\_ The higher grade standard is unclear or imprecise.

\_\_\_ The lower grade standard is unclear or imprecise.

\_\_\_ Other (explain): \_\_\_\_\_

Human Resources Research Organization

21

April 20, 2005

Overview of Vertical Alignment

## Rating Process

- You will be assigned to one of three or four tables.
- Each table begins with a different standard
- Grade-Span Subgroups (Elementary, Middle, and High School) work separately at first (Round 1)
  - Paper-and-pencil ratings, using Grade-Span Rating sheets
- Tables come together to discuss results across all grade spans (Round 2)
  - Results entered into spreadsheets on laptops
- Continue with 2<sup>nd</sup> Standard (Rounds 3 and 4)

Human Resources Research Organization

22

April 20, 2005

Overview of Vertical Alignment

## ***Discussion***

---

### **➤ Questions and suggestions about:**

- Why vertical alignment is needed? How results can be used?
- The proposed process and how it works.
- Other.

April 20, 2005

Overview of Vertical Alignment

Human Resources Research Organization

23



## Appendix H: Summary of Responses to Debriefing Questions

---

1. Do the Grade-Level Expectations cover the major content objectives from one grade to the next across grades?

- some gaps (ex., place value in grade 4) [see notes in bookmarks]
- reasonable coverage of material but not enough mention of technology
- difficulty with decimals, place value, fractions (consistency across grades not clear enough for these particular concepts)

2. How well do the Grade-Level Expectations reflect student progress from one grade to the next across grades?

- generally good process
- weak in grade 4 (especially in number sense)
- number sense needs to be looked at (more confusion) gaps → might be simple ambiguity in language

3a. The following question is for English/Language Arts Group:

How well do the sample reading passages and attached items reflect the increasing complexity of the Grade-Level Expectations across grades?

3b. The following question is for Mathematics Group:

Did the activity help you understand the increasing complexity of mathematical reasoning across grades?

- examples and non-examples of GLEs would clarify the language
- importance of glossary
- conceptual development of GLE mentioned by Sally was helpful.

\* Helpful to listen to above, middle, and below grades about concepts

\* Item was helpful when studying reasoning aspects at grade level → more concrete to link to GLEs and reasoning

4. Did this workshop help you understand how expectations for student achievement increase from one grade to the next?

- very good refresher of other grades' expectations
- workshop very helpful for going back to teaching. Unfortunately, doesn't reach everyone
- need to understand how valuable it is for a grade span
- repeat process and workshop at State-wide Professional Development in October

5. Did this workshop help you understand the most important expectations for the grades you work with?

\* Need clear and concise documents so distribution in each district is less of an issue.

6. What were the strengths of the Vertical Alignment process?

- was well done and worked pretty well
- overlap was very valuable for grades 5 and 8 due to different levels

- enlightened some teachers about the foundation and was come next and an awareness/learning of language not familiar in that grade. Especially true in the elementary grades.

7. What are the weaknesses of the Vertical Alignment process?

Rating sheet was a good tool but definitions of “broaden and deeper” need more training

8. Your recommendations for improvement:

- \* Need to do another workshop for other teachers before final draft.

## Appendix I: Vertical Alignment Workshop Evaluation Survey

Technical and Community College, Dover, DE

April 20-21, 2005

This survey is used to evaluate the vertical alignment workshop and collect feedback to improve the alignment process. Thank you for your help.

### Part I: Orientation

- How well do you feel the orientation and training prepared you for the consensus process?  

1	2	3	4
Not Well	Somewhat	Adequately	Very Well
- Please provide us with your recommendations or suggestions to improve the orientation process.

### Part II: Vertical Alignment Ratings

- How comfortable did you feel about matching expectations across grades?  

1	2	3	4
Uncomfortable	Somewhat	Comfortable	Very
	Comfortable		Comfortable
- How comfortable did you feel about identifying the type of match?  

1	2	3	4
Uncomfortable	Somewhat	Comfortable	Very
	Comfortable		Comfortable
- How comfortable did you feel about identifying the type of match?  

1	2	3	4
Uncomfortable	Somewhat	Comfortable	Very
	Comfortable		Comfortable
- How comfortable did you feel rating the importance of increased expectations?  

1	2	3	4
Uncomfortable	Somewhat	Comfortable	Very
	Comfortable		Comfortable
- How comfortable did you feel rating the clarity of increased expectations?  

1	2	3	4
Uncomfortable	Somewhat	Comfortable	Very
	Comfortable		Comfortable
- Overall, how well do you think the expectations for different grades were aligned?  

1	2	3	4
Minimally	Somewhat	Adequately	Fully
Aligned	Aligned	Aligned	Aligned

### Part III: Discussion of Complexity

- How helpful was the discussion of complexity during the second morning?  

1	2	3	4
Not Helpful	Somewhat	Moderately	Very Helpful

### Part IV: Discussion of Results

- How helpful was the discussion of results at the end of the workshop?  

1	2	3	4
Not Helpful	Somewhat	Moderately	Very Helpful
- Please provide us with your recommendations or suggestions to improve the vertical alignment process.

Subject: \_\_\_\_\_ Your Grade Level \_\_\_\_\_



## Alignment Report 5

---

### Aligning English Language Proficiency Tests to English Language Learning Standards

H. Gary Cook, Ph.D.  
University of Wisconsin  
Wisconsin Center for Education Research  
1025 West Johnson St  
Madison, WI 53700  
hcook@wisc.edu

December 21, 2005

Prepared for  
Assessing Limited English Proficiency Students  
State Collaborative on Assessment and Student Standards  
Council of Chief State School Officers



## Table of Contents

<b>Aligning English Language Proficiency Tests to English Language Learning Standards</b>	<b>135</b>
Alignment of Assessments to Standards .....	135
<b>English Language Proficiency Test Alignment Process</b>	<b>135</b>
<b>ELP Alignment Statistics</b>	<b>136</b>
Coverage .....	136
<i>Standards Incorporated under Proficiency Levels</i> .....	137
Linguistic Difficulty .....	137
Breadth .....	137
Linkage to Academic Content Standards .....	137
Interpreting Alignment Statistics .....	139
<b>Step-by-Step Procedure for ELL Alignment</b>	<b>139</b>
Step 1: Set up the Alignment .....	139
Step 2 Conduct the Alignment .....	142
Step 3 Analyze Alignment .....	143
Step 4 Interpret and Act on Alignment Results .....	144
<b>References</b>	<b>145</b>
<b>Appendix A: Generic ELL Alignment PowerPoint</b>	<b>147</b>
<b>Appendix B: Linguistic Difficulty Levels by Skill Area</b>	<b>149</b>
<b>Appendix C: ELL Alignment Examples of Assigning Linguistic Difficulty</b>	<b>152</b>
<b>Appendix D: ELL Sample Alignment Coding Sheet</b>	<b>153</b>





# Aligning English Language Proficiency Tests to English Language Learning Standards

---

## Alignment of Assessments to Standards

---

Alignment, the degree to which a test's items cover the content the test is intended to measure, is a critical element in assuring the validity of an assessment. The concept of alignment has always been subsumed under the notion of content validity. Linn and Gronlund (2000) write that content validity, which they often term "content considerations" under the notion of validity, typically is considered in the developmental stages of building assessments. Content considerations, they say, deal with issues of task sampling and test specifications, as they relate to the "domain under consideration" (p.77). The main point of content validation, they contend, is "determining the adequacy of sampling of the content that the assessment results are interpreted to represent" (p.78). Beyond task sampling and test specifications, another element of content validity has emerged and matured--alignment.

In years past, alignment was often evaluated in an ad hoc fashion. Typically, alignment activity was conducted during a test's development process, often during item review (as Linn & Gronlund state). Content experts reviewed assessment items and determined if items matched test specifications, test framework documents, or content standards. Recently researchers have argued that there is more to alignment than just matching (see LaMarca, et al., 2001; Webb 1997, 2002; and Rothman, et al., 2002). Alignment refers not only to matching items to standards but also to ascertaining the breadth and the depth of items relative to the breadth and depth of the standards.

A variety of alignment strategies and methodologies exist (see CCSSO, 2002 & 2005). One of the most prominent methods used today was created by Norman Webb of the Wisconsin Center for Educational Research. The Webb approach to alignment evaluates item match, cognitive complexity and breadth of coverage. One powerful feature of Dr. Webb's approach is its adaptability. The Webb alignment method has been used with a variety of tests and content. It has been adapted for aligning special education assessments to alternate standards (Roach, et. al., 2005) and even for evaluating the alignment of standards to standards (Cook, 2005c). The process described here adapts the Webb alignment methodology to English language proficiency assessments and standards.

## English Language Proficiency Test Alignment Process

---

Alignment research has focused primarily on aligning academic achievement assessments to academic content standards. Until recently, English language proficiency test (ELP) alignment has not been part of investigations. ELP assessments entail test constructs based on second language acquisition (SLA) principles across oral and

written English, which differ from the constructs of academic achievement assessments. Consequently, the process of aligning ELP assessments to the SLA stages reflected in English language development standards is not provided for within existing alignment methodologies.

The goal of ELP alignment is to match an assessment's linguistic skills and acquisition levels to English language development standards. Alignment takes into consideration the match of requisite language skills, i.e., speaking, listening, reading, and writing; linguistic complexity via acquisition levels; and the linkage to a state's academic content standards. It is important to recognize that alignment is a two-way process. It addresses both of these questions: How well is the test aligned to the standards? How well are standards covered on this test? Both ways are important. The process described here takes into account:

1. Degree of coverage of English language development standards,
2. Correspondence between the linguistic complexity of items and standards,
3. Breadth of coverage of English language development standards
  - a. Range of coverage,
  - b. Balance of coverage, and
4. Linkage to state academic content standards.

Raters with expertise in the area assessed provide the information used to determine the degree of alignment between a test and content standards. The raters determine the linguistic complexity of the standards and of each item, and they match items to the standards they assess. The next section describes metrics that are derived from the rater reviews to analyze these elements of ELP alignment.

## ELP Alignment Statistics

---

A variety of statistics are generated as a result of the ELP alignment process. Four areas are examined in this type of alignment: coverage, linguistic difficulty, breadth, and linkage to state academic standards. Each area has associated statistics. As stated earlier, this alignment process is very similar to the Webb alignment model—although not identical—and provides similar types of statistics. ELP alignment differs from academic content based alignments in a few important ways. First, English language development standards cover skill-based content (listening, speaking, reading, and writing) and language proficiency. That is, the standards incorporate content and linguistic complexity. To determine alignment, both elements need to be examined. Thus, statistics associated with ELP alignment provide measures of coverage, linguistic difficulty, and breadth for skill-based content and for language proficiency levels. The following paragraphs describe each area and the statistics used to evaluate that component of alignment.

### Coverage

To evaluate coverage, the statistic **Categorical Concurrence** is used. Categorical Concurrence refers to the average number of items that raters match to specific English language development standards. Ratiers review each item on the test and determine which specific standard(s) match the item. It is important to note that some items can address more than one standard, and raters are allowed to code accordingly. The number of items matched to each standard is averaged across all raters and reported as Categorical Concurrence. Therefore, the Categorical Concurrence rating for a standard is the average number of items raters believe address that standard.

For skill-based content, the suggested acceptable minimum number of items per standard is set at three; i.e., on average raters must identify at least three items per standard for Categorical Concurrence to be minimally acceptable. For example, let us say that a state's set of listening standards at grades 3 – 5 has four standards: following directions, identifying the main idea in a conversation, identifying grade appropriate vocabulary, and making inferences or predictions in a conversation. Using the criterion above, each standard would need to have an average of 3 items identified for the test to be minimally acceptable in terms of Categorical Concurrence.

## Standards Incorporated under Proficiency Levels

In some cases, ELD standards are subsumed under language proficiency levels. For example, a state may have two listening standards for “Beginner,” three additional standard for “Intermediate,” and two more for “Advanced.” Tables 1 and 2 (below) show standards arranged by proficiency level. In this case, the alignment analysis also takes into consideration the number of items corresponding to each proficiency level, calculating Categorical Concurrence for each level. The suggested acceptable minimum number of items per proficiency level is six, although states may choose to have more items per level. Note that the criterion is higher than that for match to content standards. This is because states typically report summative scores indicating students’ proficiency levels. Three items would be far too few to make reliable judgments about a student’s language proficiency. An acceptable criterion for levels depends upon how a state’s standards are structured. Using Webb’s criterion as a starting point, it seems reasonable to require at least six items per proficiency level. However, in many cases a higher number, e.g., 12, would be better. Setting the acceptable criterion for number of items that match each proficiency level is one of the issues to discuss when preparing for an ELP alignment.

## Linguistic Difficulty

To evaluate linguistic difficulty, a metric that represents the percent of items at each **linguistic difficulty level (LDL)** is used. This measure is somewhat akin to Webb’s Depth of Knowledge statistic, but in this case it refers to linguistic complexity. Raters assign each English language development standard a linguistic difficulty level of 1, 2 or 3. Level 1 stands for elementary linguistic features; Level 2 represents standard linguistic constructions, and level 3 refers to complex linguistic formulations. See Appendix B for a matrix of LDLs by skill area.

During the alignment process, raters also assign LDLs to each test item. The notion is to identify the degree of match between the standards’ LDLs and items’ LDLs. The statistic used to describe the match refers to the percent of items coded at the LDL level of the standard. The major purpose for assessing ELL students is to help identify and monitor linguistic progress. If test items do not match the linguistic levels of the standards, the ability to properly evaluate students is limited. If items are coded above or below a standard’s linguistic difficulty level, little information would be available about how that item represents students’ behavior relative to the linguistic difficulty of that standard. This is not to say that all items need be at the LDL of the standard. The suggested acceptability criterion for LDL match is 50% for both skill-based content and language proficiency levels. That is, at least 50% of the test items are at or the LDL of the standards they are intended to measure and the proficiency level they target.

## Breadth

Two statistics are used to evaluate how well an assessment covers the breadth of a state’s English language development standards. The first measure is range. **Range** refers to the how well a test’s items cover a set of standards. Range is linked to Categorical Concurrence. The Range statistic is the percent of standards (and proficiency levels if applicable) that have an acceptable Categorical Concurrence rating. Again, using Webb’s criterion, the suggested acceptable Range criterion is 50%, i.e., 50% of the standards or levels need to meet the minimum Categorical Concurrence criterion.

The next statistic used to evaluate breadth is **Balance**. Balance further expands on the notion of breadth by evaluating whether the distribution of items within a set of standards is consistent. There may be a high proportion of standards and levels covered (i.e., a good range value), but if most items are assessing just one area (sometimes called objective or indicator) within the standards, and the other standards are covered by only one item each, for example, the Balance of coverage is not appropriate. Ideally, if all areas (objectives/indicators) within a standard or level are of equal importance, the test should distribute items over the areas equally. Again following Webb, an acceptable balance criterion is .70 or higher (the formula used to calculate the Breadth statistic is shown in a later section).

## Linkage to Academic Content Standards

Linkage involves the connection of a state’s English language development standards to its academic content standards. During the ELP alignment process, raters not only evaluate alignment to the state’s English language development standards they also identify linkages to the states academic content standards as well. A state’s ELP exam is designed to identify students’ English language ability and their readiness for learning in an environment where English is used. Thus, there is a need to assure that the English discourse features within specific academic content areas are assessed. Raters identify whether items relate, in some fashion, to the language used in grade appropriate content instruction. Ideally, there will be sufficient items covering key content specific discourse features on an ELP examination. An insufficient number of items may reflect

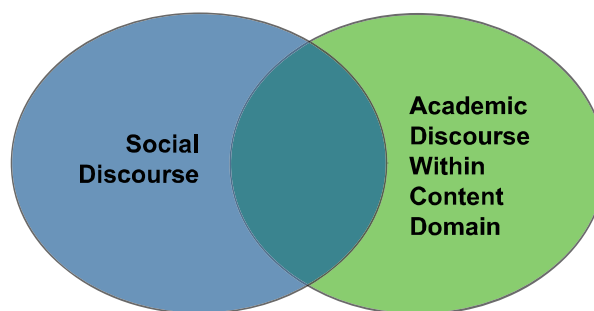
language proficiency more associated with social discourse rather than academically related discourse. Both are needed. Both should be incorporated on the ELP assessment.

A variety of researchers have sought to define and categorize the notion of academic and non-academic language skills as they relate to educating children for whom English is not a native language, e.g., Cummins (1979) Basic Interpersonal Communication Skills (BICS) vs. Cognitive Academic Language Proficiency (CALP) or Brinton, Snow, and Wesche (2003) Content Based Instruction or CBI. In contexts where English is the lingua franca, non-English speaking children must gain the ability to interact with their environment in English. Typically, children find themselves in two contexts: at home and at school. Within the home context, children often speak their native language. When outside the home, children are confronted with and are required to gain the ability to interact in English. Often, the first step in instructing these students is to teach them “survival” English. “Survival” here means common social discourse typically associated with day-to-day interactions. In the past, language instruction for these students has not progressed beyond common social discourse. Certainly bilingual programs have sought to bridge the gap between social and academic development of these students by teaching academic content in students’ native language. Nonetheless, many of these students have lagged behind their age-level peers in the development of their academic skills, even in bilingual programs.

The major thrust of recent federal programs associated with non-native students’ English language development has been to not only focus on the requisite social discourse but also to prepare student to engage in the academic discourse required to learn academic content. This introduces a new paradigm, and a new focus for assessments. English language proficiency assessments must not only evaluate educationally based social discourse features but must also address the unique English discourse associated with learning math, science or English language arts.

An important consideration in developing ELD standards and assessments that measure skills linked to achievement in academic content areas is to clearly separate the language used in a content area from content-specific knowledge and skills. If content knowledge is assessed in combination with content-related discourse features, students’ scores will be confounded. If a student responds incorrectly to an item requiring knowledge of both language features and content, there is no way to tell whether it is because the student lacks the linguistic skills necessary to respond correctly or the content knowledge necessary to respond correctly. In measurement terms, such items add construct-irrelevant variance to a student’s score and threaten the validity of the inferences that can be made from the test. In the context of an alignment study, such items will be flagged as contributing an inappropriate source of challenge.

**Figure 1: Overlap of Social and Academic Discourse**



The figure above illustrates the overlap between the two discourse domains. Language assessments and often language instruction have focused primarily on the social discourse domain. However students also need to be instructed and assessed on the unique academic discourse that they will face in their English-only classrooms. One immediate application of this concept for assessments is to include lexical and syntactic features associated with a state’s mathematics or English language arts standards and curricula. The concept of linkage described in the alignment process here attempts to capture the degree to which academic discourse elements are captured on the ELP examination.

The core academic content domains should be represented on the ELP examination. Determining a criterion for linkage depends upon the way the ELP assessment is linked to a state’s content standards. In some cases, states may choose to develop items that assess general discourse features in the content areas. For example, a proportion of the items would measure discourse in mathematics, another portion in science, and another in English language arts. If this is the case, the criterion for content-linkage alignment might be at least 25% of the items linked to one of the content areas.

In other cases, states may choose to apportion items by standards within each content area. For example, the state might have five math content standards—number sense, algebra, geometry, measurement, and statistics—and include items that measure discourse related to each of these standards. Recall that the issue is to identify items on the ELP exam that have English discourse features related to mathematics—not to identify math content items! In this situation, a rule of thumb is that most standards should have at least one linked item for each standard, perhaps with four out of the five math standards having at least one linked item on the ELL assessment. This criterion establishes a minimal link to key standards within mathematics. It would not be unreasonable to expect higher criteria for English language arts or reading since ELL examinations are designed to evaluate English competency. Instead of one item, it is suggested that three items be linked to English language arts standards. If a state has four reading standards: comprehension, vocabulary, literature, analysis, we would expect at least three of the four standards to have at least three items to meet this criterion. The expected criterion varies depending upon the degree of linkage a state wishes to have with its academic content standards. The criteria set forth here represent minimal expectations.

### ***Interpreting Alignment Statistics***

No one statistic can determine the alignment of a test to a state's standards. All statistics should be used in concert to obtain a complete picture of alignment. In some cases, a state may deliberately target specific standards on its assessments, including more items for some standards than for others. In this case, the criteria for acceptable alignment suggested above will not adequately portray the alignment. However, criteria should be set before an alignment study is conducted and a justification for each criterion should be articulated.

The process described above has been used in three states with a variety of assessments and English language development standards (Cook 2005a, 2005b, 2005d) with success. Experience suggests that careful planning and clear vision of the purpose and intended outcome of the alignment process leads to success. The following section outlines step-by-step how this alignment process is implemented.

---

## **Step-by-Step Procedure for ELL Alignment**

### ***Step 1: Set up the Alignment***

- Meet with stakeholders and develop alignment protocols
- How are the standards to be arrayed for the alignment?
- How many participants and at what grades/subjects are to be aligned?
- How will the alignment be conducted? Paper-and-pencil? On-line?
- Set up the alignment study based on protocols

First, relevant stakeholders groups need to be identified. Suggested stakeholders might be ELL or Bilingual Teachers, District ELL Coordinators, Title III Directors or Coordinators, university ESL or bilingual faculty, and state Title III, Bilingual or ELL Coordinators. The goal is to bring together experts in both ELL students and the state's English language development standards. The initial meeting should be with relatively few people who have this expertise. The goal of the first meeting is to establish alignment protocols.

At least three questions should be discussed when establishing ELL alignment protocols and procedures. First, how are the standards to be arrayed for the alignment? Each state has unique language development standards that are arrayed differently. Tables 1 and 2 show two different states' standards. In both states, the content standards are arranged by proficiency level. That is, each content standard is designed to assess language skills within a single proficiency level. Other states may have separate content standards and proficiency levels, similar to the structure of their academic content standards and achievement levels.

Table 1:

## Sample of South Dakota's English Language Proficiency Standards in Speaking

Grades 3-5

Speaking Standard 2.1		
Standard 2.1: The student will produce spoken English to participate in social (informal) contexts. Student will be able to:		
Pre-Emergent	Emergent	Basic
a. listen to language and observe appropriate cultural and learning behaviors of peers and adults	b. participate in short, limited social exchanges with peers and teachers c. express basic needs and preferences, answer simple questions d. use gestures, single words, phrases and formulaic expressions e. mimic words, speech patterns and phrases f. know and communicate first and last name, address and phone number g. imitate pronunciation of number words and mathematical terms	h. participate in highly predictable conversations on familiar topics with peers and teachers i. express many needs and preferences j. ask and answer simple questions k. produces phrases and simple sentences l. speak with sufficient accuracy that listeners accustomed to language learners comprehend some of the message m. use familiar, general vocabulary n. use math vocabulary related to daily routine

Speaking Standard 2.2		
Standard 2.2: The student will produce spoken English to participate in academic (formal) contexts. Student will be able to:		
Pre-Emergent	Emergent	Basic
a. Listen to language and observe appropriate cultural and learning behaviors of peers and adults	b. respond to routine classroom questions c. express basic needs and preferences d. answer simple questions about familiar material e. use gestures, single words, phrases and formulaic expressions f. use limited target vocabulary when prompted g. know and communicate first and last name h. imitate pronunciation of number words and mathematical terms	i. participate in limited/guided discussions j. give very simple oral reports related to self or topics of high personal interest k. express many needs and preferences l. ask and answer questions m. produce phrases and simple sentences n. recombine learned material o. speak with sufficient accuracy that listeners accustomed to language learners comprehend some of the message p. use general target vocabulary in classroom activities q. use some math vocabulary

Table 2:

## Sample of Michigan's English Language Proficiency Standards in Speaking Grades K-12

Standard	Level 1	Level 2	Level 3
<b>S.1 Use spoken language for daily activities within and beyond the school setting</b>			
	S.1.1.a Use learned phrases to respond to questions and directions	✓	✓
		S.1.2.a Make requests and obtain information from the community	✓
			S.1.3.a Participate in conversations on social topics by asking and requesting information
<b>S.2 Engage in conversations for personal expression and enjoyment</b>			
	S.2.1.a Communicate basic wants and needs in English	✓	✓
	S.2.1.b Use common social greetings and simple repetitive phrases	✓	✓
		S.2.2.a Participate in social conversations with peers and adults on familiar topics by asking and answering questions and requesting information	✓
			S.2.3.a Participate in social conversations with peers and adults on unfamiliar topics by asking and answering questions and restating and requesting information

As can be seen right away, each state arrays its standards in slightly ways. Each table displays the first three proficiency levels of speaking standards. Note that there are two elements that merit alignment: the standards and the proficiency levels. This first issue to settle is to decide how standards should be aligned. Prior to discussing how each state's standards are aligned, it's important to develop common terminology for standards in the ELL context. We can think of standards being arrayed hierarchically into three levels: standards, goals, and objectives. In the examples above, there are only two levels: standards and goals. For example, South Dakota has speaking standard 2.1 which has several goals—in the table goals are represented by letters “a” though “n”. Michigan has several standards (only two are displayed) and several goals represented by S.1.1.a, S.1.2.a., etc. South Dakota has only two speaking standards while Michigan has 8. Next one needs to look at levels. South Dakota's English Language Development Standards have five levels, while Michigan's standards have four. These need to be considered in the alignment as well.

How do you array standards and levels to conduct alignments? This is an issue decided by stakeholder groups. In the case of South Dakota, it was decided to create “standards/levels.” That is, standards were combined with proficiency levels. Thus, the alignment was conducted on Standard 2.1 Pre-Emergent/Emergent, 2.1 Basic, 2.2 Pre-Emergent/Emergent, 2.2 Basic, and so on. In this protocol standards and levels are combined.

Michigan chose not to cross standards with levels. In the Michigan alignment, each speaking standard, S.1 to S.8, was aligned. In both states, an additional alignment was done on the proficiency levels, across all skill areas. For South Dakota, the alignment was on standards/levels and proficiency levels. For Michigan, it was on standards and levels. In both states, the standards/levels or standards alignment identified how well assessments matched or covered the language proficiency standards within skill areas. The proficiency level alignment identified how well proficiency levels were matched and covered across skill areas (i.e., speaking, listening, reading and writing). Since decisions are made regarding students' language progress based on the



level they are assigned across all skills areas, this is a particularly relevant analysis. As can be seen, stakeholders need to clarify the protocols used to conduct the alignment and this protocol and process will be unique to each state.

Once the alignment protocol is decided, the next decision to be made is who should participate in the alignment review. As a rule of thumb, six to eight participants review tests in each grade band covered by the assessments. Commonly used grade bands are kindergarten to second grade, third to fifth grade, sixth to eighth grade, and ninth to twelfth grade. Using these bands, one would expect between 24 and 32 raters participating in an alignment review. Alignment participants should have experience with English language learners and the assessments used to evaluate those learners. Like the stakeholders who planned the study, ELP or Bilingual Teachers, District ELL Coordinators, Title III Directors or Coordinators, and university ESL or bilingual faculty are ideal candidates.

Once the protocol and participants have been determined, the next step is to decide how to collect alignment information. Two methods are available to collect alignment information: paper-and-pencil and the Webb Alignment Tool ([www.wcer.wisc.edu/wat](http://www.wcer.wisc.edu/wat)). The paper-and-pencil method requires the handling of rating forms as well as test materials and standards. The Webb Alignment Tool (WAT) has all forms on-line. However, test materials and copies of standards are also needed, while not necessary. The most beneficial feature of the WAT is the speed of results. In the paper-and-pencil recording method, results have to be tallied and then analyzed. The WAT recording method conducts the alignment analysis as raters are coding. It is much quicker and more efficient. However, the WAT method requires that each participant have access to a computer that has internet access.

After protocol, participants, and collection method have been determined, all materials (e.g., copies of standards, assessments, ancillary materials like Directions for Administration, etc.) needed for the alignment are created, collected and staged for the study.

### **Step 2 Conduct the Alignment**

- Convene an independent alignment committee for pre-assigned ELD standards and grade spans
- Train the alignment committee on the alignment process
- Have alignment committee assign linguistic difficulty levels to ELD standards and proficiency levels, if these are separate. This is a consensus process.
- Have the alignment committee independently assign, to each test item, a linguistic difficulty level and a match to an ELD standard and a proficiency level
- Have the alignment committee identify links to the state's academic content standards (e.g., English language arts and/or mathematics)

Participants are placed into grade-level groups and are convened for the alignment study. The actual alignment process has four phases:

**Phase one** involves training alignment participants in the process. Appendix A displays a generic PowerPoint presentation used to train participants. The key to good alignment is good training. It is critical that participants are familiar with the alignment process as well as with the English language development standards and the assessments being aligned. It is always good practice to have participants review all assessment and standards materials thoroughly at the meeting, regardless of their prior familiarity with the materials. Appendices B & C provide copies of the LDLs by skill area and example materials. The materials in Appendices B & C are akin to rubrics used for rating student work. In this case, they are guidelines for assigning linguistic difficulty and aligning items to standards. Appendix D displays the paper-and-pencil alignment coding sheet, for use if computers are not being used in the review.

**Phase two** is to assign LDLs to the English language development content and proficiency standards. Typically, the objective-level is where LDLs are assigned. However, this is one of the decisions that needs to be made at the initial planning of the alignment. Using Michigan's speaking standards as an example (Table 2), let us look at standard S.1 (Use spoken language for daily activities within and beyond the school setting). Under this standard there are several objectives. The first objective is S.1.1.a "Use learned phrases to respond to questions and directions." Raters would look at this standard and assign an LDL level to this objective. Using Appendix B, one might assign an LDL of 1 to this objective. Each objective is rated accordingly. The assignment of LDLs to English language development standards is a consensus process. That is, all raters must agree on



the assigned LDL is for all objectives. Rich discussion is often the outcome of this second phase, and it is critical that group members agree.

In **phase three**, participants have two tasks. Unlike assigning LDLs to standards, these tasks are done independently, i.e., each participant rates items individually. First, they assign LDL levels to each assessment item. Next, they identify which standard or standards an assessment item addresses. Note, that an assessment item can align to more than one standard, but there must be one “primary” standard to which the item aligns. Participants are also given the chance to identify any “sources of challenge” that items might display. “Source of challenge” means that a participant believes that an item has something in its stem, item foils or format that mislead students, i.e., it can or will lead students to the wrong or right answer for the wrong reason. The item might be biased or poorly constructed. Raters should be given the opportunity to comment on any items that they believe represents a “source of challenge.”

**Phase four** of the ELP alignment is to identify how ELP assessment items link to selected state academic content standards. As stated earlier, how assessments incorporate academic discourse within content areas is a decision made in the planning stages of an alignment. Here raters independently identify which content area(s) an ELP assessment item might address. A broad degree of flexibility is suggested here. That is, perfect linkage (or match) is not the goal. The goal is to identify which items relate to the academic discourse of the content standards. After all participants complete their ratings, this part of the alignment is complete, and participants can be dismissed. The next step is to analyze results.

### Step 3 Analyze Alignment

- Identify assessment's coverage of ELD standards
- Identify assessment's linguistic appropriateness relative to ELD standards
- Identify assessment's breadth of ELD standards coverage
- Identify assessment's linkage to state's content standards

This section briefly highlights how each statistic used to evaluate the alignment of a test to standards is calculated. While the statistics are straightforward, the actual analysis is somewhat complex. It can be done, with some work, in Microsoft Excel. However, it is more efficient to analyze results using the on-line alignment tool.

Coverage (categorical concurrence) averages the total number of “hits” or matches of items to standards or objectives. If Rater #1 assigns 2 items to Standard 1; Rater #2 assigned 3 items to Standard 1, and Rater #3 assigned 4 items to Standards 1, there is a total number of 9 “hits” on Standard 1. There are three raters; thus nine divided by three is three. The Categorical Concurrence for Standard 1 is 3.

LDL Consistency expresses the proportion of LDLs assigned to the items within a standard that is below, at, or above the standard's LDL. Using the example above, if three items were coded at an LDL of 1, three items at a 2, and three items at a 3, and the standard was coded at an LDL of 2, then 33% of aligned items are below the standard's LDL, 33% are at the standard's LDL, and 33% are above the standard's LDL. Said differently 66.7% (or 2/3) of the items are at or above the standard's LDL.

Range is calculated by determining if 3 or more items are associated with a standard. If so, that standard is coded with a 1. Standards with less than 3 items are coded with a 0. In the Michigan example above, there are 8 speaking standards. If five speaking standards have 3 or more aligned items, then the range is 5/8 or 62.5%.

The balance statistic is slightly more complex. The equation below displays the equation used to calculate balance index.

$$Balance = 1 - \frac{\left( \sum \left| \frac{1}{O} - \frac{I_k}{H} \right| \right)}{2}$$

Where O is the total number of objectives identified, I<sub>k</sub> is the number of items identified that are associated with objective k, and H is the total number of items hit. Balance identifies emphasis of standards. It is plausible that a set of standards receives acceptable range, but if one or two standards are greatly emphasized on the test, it will not be balanced. The balance index assumes equivalence in emphasis of standards. An index value of 0.70 or greater is acceptable. Note that if specific standards are purposely emphasized the balance statistics may not be as meaningful.

### ***Step 4 Interpret and Act on Alignment Results***

The purposes behind conducting an ELP alignment study may be many. Content validation of an assessment and an evaluation of its relationship to standards is certainly in the forefront of purposes. Communicating the strengths of a test's alignment for policy purposes might be another reason for conducting such a process. The eventual outcome, however, should be to better understand assessments and standards and to adapt either or both to better evaluate students' progress in learning English.

As stated earlier, reliance on one statistic--be it match, complexity or breadth--to make statements about the quality of an alignment is very misleading. We have often heard statements like, "What percentage of your test is aligned?" This statement presumes that one number, say 95%, is sufficient to explain the relationship between what is assessed and what is expected. Unfortunately, it is not that simple. ELP alignment is a two way process (test to standards and standards to test) which incorporates not only match and breadth but linguistic complexity. One metric cannot say it all. When interpreting results, you must use all statistics to get a full picture of how well a test aligns to standards or how well standards align to a test. Further, if a test does not meet the criteria described here that does not mean it is a bad test. It could be that the standards need revision. It could be that those creating the test wanted specific standards to be highlighted and other to be stressed less. That is why it is critical to make decisions, before the alignment process on what criteria for "acceptable" should be. Thinking about and setting up these decisions a priori leads to more meaningful results and to a more actionable process.

Categorical Concurrence provides information on how well the test and test items match standards. Low Categorical Concurrence values reflect misalignment in the match between assessments and standards. One outcome of this misalignment would be to provide more items aligned to inadequately covered standards. This would affect test specifications or item maps for particular forms of the test. It may be that the degree of misalignment is acceptable -- that is, those standards with low Categorical Concurrence are deliberately stressed less than other standards because they are not as critical. It could also be that there are too many standards. For example, if a test of 30 items covered 5 standards equally, it is easy to see that the Categorical Concurrence for that test would be higher than for a 30 item test covering 10 standards equally. In the latter case, states may want to consider reducing the number of standards covered, perhaps by collapsing similar skills into a single standard, or increasing the number of items in the test. The number of standards to be aligned will affect the alignment for a test of a given length.

Linguistic Difficulty deals with the language acquisition expectations of students. At the beginning of the alignment process, LDLs are assigned to each objective or standard. This process places linguistic complexity expectations on standards, and the assessments should adequately cover those expectations. LDL misalignment reflects the need to examine the linguistic processes sampled on the assessment. If the linguistic expectations of the assessment are too low, student test performance may be high but actual acquisitional development may be over reported. Often teachers comment that a student has passed a specific language test but cannot function in class in English. This may be a reflection of the limited linguistic difficulty expected by that test. Low LDL consistency should be dealt with by examining test specification and test maps. Further, during item review, examining an item's LDL prior to acceptability may increase this aspect of alignment.

Breadth is evaluated with two statistics: Range and Balance. These metrics address adequate coverage of standards. The overarching presumption of breadth is the equivalence of item coverage across all standards. If equivalence is not desired across standards, breadth statistics are less meaningful. Low Range estimates indicate that specific standards or levels have inadequate numbers of items and hence inferences about those standards or levels are limited. Low Balance typically indicates that specific standards or levels are emphasized while others are not. The resolution for low Range and/or Balance, given that equivalence is desired, is to examine test specifications and test maps and adequately "fill in the gaps."

Notice that almost all actions above deal with examining test specifications and/or test maps. Misalignment can typically be easily dealt with by adding or redistributing items. However, this may not be possible. Test length may be an issue. It may not be feasible to add more items. In this case, other alternatives should be examined, e.g., looking at the standards. It may be that the standards are too expansive to be covered by a particular assessment. If this is the case, other assessment formats could be explored or a paring down of standards may be in order. Unfortunately, revisiting standards is not typically an easy process. Nonetheless, misalignment should result in some sort of action, be it revision of test specifications, reexamining and reapportioning items on test maps, reviewing and revising standards, or adding procedures to the item and test review process. Taking appropriate action from alignment results will lead to better assessments and a better understanding of what students know and are able to do.

## References

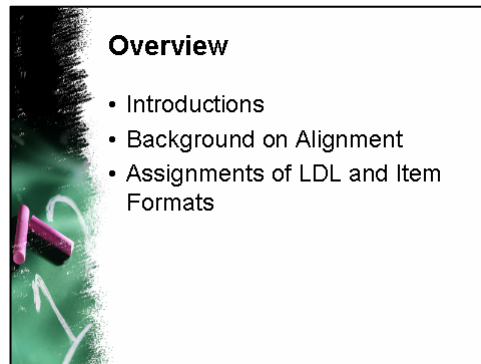
- Brinton, D.M., Snow, M.A., and Wesch, M. (2003). *Content-Based Second Language Instruction*. Ann Arbor, MI: University of Michigan Press.
- Cook, H.G. (2005a). *Alignment Study--South Dakota's English Language Proficiency Standards for English Language Learners K-12 to Stanford English Language Proficiency Examination*. Pierre, SD: South Dakota Department of Education.
- Cook, H.G. (2005b). *Alignment Study: Michigan's K-12 English Language Proficiency Standards to Mountain West's ELL Assessment and Accountability/Works' ELL Assessment*. Lansing, MI: Michigan Department of Education.
- Cook, H.G. (2005c). *Milwaukee Public Schools Alignment Study of Milwaukee Public Schools' Learning Targets In Reading and Math To Wisconsin Student Assessment System Criterion-Referenced Test Frameworks in Reading and Math, Research Report #0504*. Milwaukee, WI: Milwaukee Public Schools.
- Cook, H.G. (2005d). *Alignment Study: Wyoming's K-12 English Language Development Standards for English Language Arts to Stanford English Language Proficiency Test*. Cheyenne, WY: Wyoming Department of Education.
- Council of Chief State School Officers. Alignment Analysis. Retrieved July 27, 2005. Author's website: [www.ccsso.org/projects/Alignment\\_Analysis/](http://www.ccsso.org/projects/Alignment_Analysis/).
- Council of Chief State School Officers. (September 2002). *Models for Alignment Analysis and Assistance to States*. Washington, D.C.: Author.
- Cummins, J. (1979) Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, No. 19, 121-129.
- LaMarca, P.M., Redfield, D., Winter, P.C., Bailey, A., and Hansche Despriet, L. (2001). *State Standards and State Assessment Systems: A Guide to Alignment. A study of the State Collaborative on Assessment & Student Standards (SCASS) Comprehensive Assessment Systems for ESEA Title I (CAS)*. Washington, D.C.: Council of Chief State School Officers.
- Linn, R.L. and Gronlund, N.E. (2000). *Measurement and Assessment in Teaching, Eighth Edition*. Prentice Hall, Inc.: Upper Saddle River, NJ.
- Roach, A. T., Elliott, S. N., & Webb, N. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin alternate assessment. *The Journal of Special Education*, 38, 218–321.
- Rothman, R., Slattery, J.B., Vranek, J.L., Resnick, L.B. (2002). *Benchmarking and Alignment of Standards and Testing, CSE Technical Report 566*. Los Angeles, CA: Center for Student and Evaluation, National Center for Research on Evaluation, Standards and Student Testing.
- Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education. Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6*. Madison: University of Wisconsin, Wisconsin Center for Education Research.
- 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)*. Washington, D. C.: Council of Chief State School Officers.



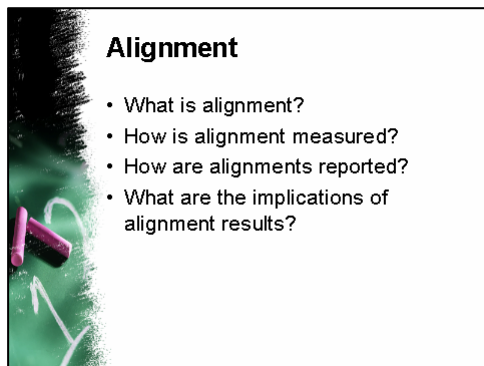
## Appendix A: Generic ELL Alignment PowerPoint



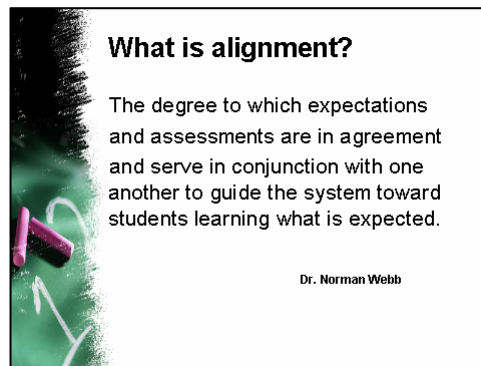
1



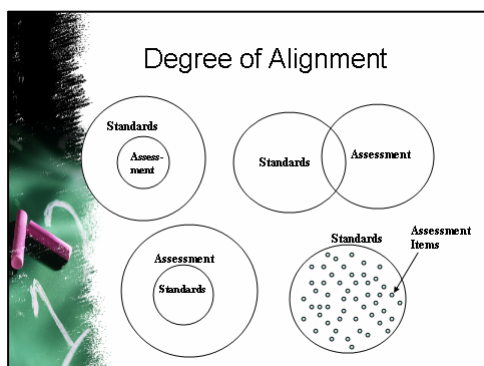
2



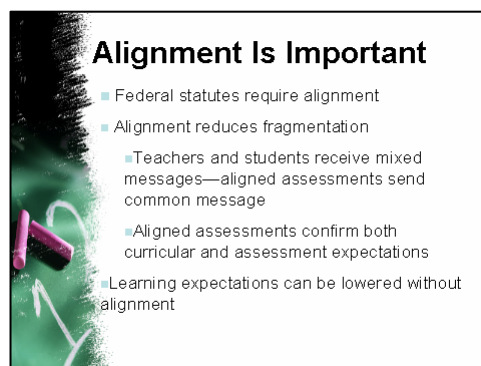
3



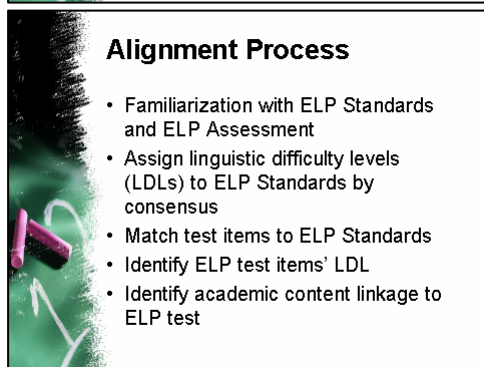
4



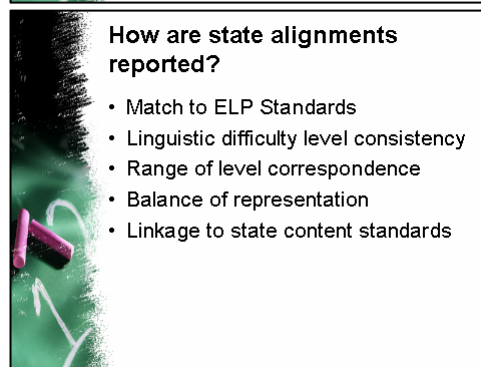
5



6



7



8

### Linguistic Difficulty Levels

Skill Area	Linguistic Difficulty Levels			
	Level 1: Elementary Features	Level 2: Standard Constructions	Level 3: Complex Formulations	
Oral/Aural	Listening	Understands and attends to simple formulaic classroom and social discourse, attending to the day-to-day and common language interactions.	Basic understanding of and attending to more complex and specialized classroom and social discourse, common formulaic expressions, both active classroom and in social situations and contexts.	Understanding of and attending to more complex or specialized grade appropriate discourse and interaction, comprehending contextualized and decontextualized communication, e.g., often complex, lengthy, and extended.
	Speaking	Understands and attends to simple formulaic classroom and social discourse, attending to the day-to-day and common language interactions.	Basic understanding of and attending to more complex and specialized classroom and social discourse, attending to the day-to-day and common language interactions.	Understanding of and attending to more complex or specialized grade appropriate discourse and interaction, comprehending contextualized and decontextualized communication, e.g., often complex, lengthy, and extended.
Written	Reading	Understands and attends to simple formulaic classroom and social discourse, attending to the day-to-day and common language interactions.	Basic understanding of and attending to more complex and specialized classroom and social discourse, attending to the day-to-day and common language interactions.	Understanding of and attending to more complex or specialized grade appropriate discourse and interaction, comprehending contextualized and decontextualized communication, e.g., often complex, lengthy, and extended.
	Writing	Understands and attends to simple formulaic classroom and social discourse, attending to the day-to-day and common language interactions.	Basic understanding of and attending to more complex and specialized classroom and social discourse, attending to the day-to-day and common language interactions.	Understanding of and attending to more complex or specialized grade appropriate discourse and interaction, comprehending contextualized and decontextualized communication, e.g., often complex, lengthy, and extended.

9

### Linguistic Difficulty Level Descriptors

- Level 1 Elementary Features
  - A limited to basic ability to process formulaic English linguistic features.
- Level 2 Standard Constructions
  - A basic to moderate ability and facility to process English linguistic features
- Level 3 Complex Formulations
  - A moderate to sophisticated ability and facility to process English linguistic features

10

### ELP Alignment Training Handout

11

### Reference Documents

12

### ELL Assessment

13

### What are the implications of the results?

- Confirm quality of assessment items and standards
- Attend to linguistic difficulty levels
- Content validation of ELP assessment
- Identify acceptable levels for alignment criteria

14

### Alignment Process

- Break up into grade span groups
  - Early Elementary, Elementary
  - Middle School, High School
- <Log on to Web Alignment Tool (WAT)—On-Line only>
- Individually assign LDLs to ELP Standards
- Discuss and assign LDLs' to ELP Standards by Consensus

15

### Alignment Process (continued)

- Assign LDLs to ELP test items
- Match ELP test items to the State's ELP standards
- Identify linkage of ELP test items to state's content standards

16

## Appendix B: Linguistic Difficulty Levels by Skill Area

Skill Areas		Linguistic Difficulty Levels		
		Level 1: Elementary Features	Level 2: Standard Constructions	Level 3: Complex Formulations
Oral/Aural	Listening	Limited to a basic understanding of simple formulaic, classroom and social discourse; attending to day-to-day brief and common language interactions	Basic understanding of and attending to every day classroom and social discourse, common idiomatic expressions both in the classroom and in social situations and contexts;	Understanding of and attending to more complex or specialized grade appropriate discourse and interactions, comprehending contextualized and acculturated communications (e.g., ellipsis, comedy, parody)
	Speaking	Limited to a basic ability to produce formulaic expressions in standardized classroom and social situations	Facility to produce standard classroom and social discourse interactions using extended formulaic expressions as well as common idiomatic expressions	Facility to produce and interact within complex classroom and social discourse interactions utilizing more contextualized and acculturated forms and constructions.
Textual	Reading	Little to a basic facility to process and attend to English phonemic and alphabetic constructions; little to a basic ability to comprehend high frequency grade appropriate classroom and survival vocabulary items	Basic understanding of and ability to attend to standard everyday grade appropriate texts which include vocabulary and passages most commonly encountered classroom and every day social situations	Understanding of and attending to grade appropriate vocabulary and texts; grade-level ability to comprehend classroom and socially appropriate texts.
	Writing	Limited to a basic ability to copy and/or produce simple text constructions (e.g., letters, basic vocabulary items, name)	Basic ability to produce simple, grade relevant classroom based and/or social text utilizing standard vocabulary and grammatical features and constructions	Facility to produce grade appropriate text constructions using appropriate vocabulary and grammatical features and constructions; ability to produce and express grade appropriate ideas and concepts

## Appendix C: ELL Alignment Examples of Assigning Linguistic Difficulty

### Level 1 Elementary Features

#### Example 1:1 Listening

**Objective:** Follow simple two-step oral directions to complete a task in English

*This objective is an example of level 1. Students most commonly meet this expectation by using elementary linguistic features—an elementary understanding of direction would be all the student would need to know to meet this objective.*

#### Example 1:2 Speaking

**Objective:** Use common social greetings and simple repetitive phrases.

*This objective is a level 1 since only simple constructions and formulaics are needed to meet his objective.*

#### Example 1:3 Reading

**Objective:** Recognize some common English morphemes in simple phrases or sentences.

*This objective requires students to identify elementary text-based features and represents a linguistic difficulty level of one.*

#### Example 1:4 Writing

**Objective:** Student will be able to write basic personal information (name, address, phone number)

*This objective highlights basic writing expectations and is a level 1 Linguistic Difficulty.*

#### Example 1:5 Writing

Item: Prompt  
Choose the correct word.

Triangles always \_\_\_\_\_ three sides

- A      have
- B      half
- C      has

This item is a level 1 since it requires a student to use elementary vocabulary and syntactic features.



## Level 2 Standard Constructions

### Example 2:1 Listening

**Objective:** Participate in routine classroom discussions.

*This objective requires students to move beyond simple directions or responding to common formulaic expressions. It requires students to listen to more sophisticated, albeit routine classroom interactions and is thus a level 2.*

### Example 2:2 Speaking

**Objective:** Deliver simple narrative and informative presentations and express with simple, detailed sentences.

*Here students must go beyond simple formulaics. This objective requires students to engage in simple by dynamic interactions representing standard classroom interactions and discourse and hence is a level 2.*

### Example 2:3 Reading

**Objective:** The student will read, comprehend, and analyze fiction and nonfiction. Answer simple, factual questions about what is read.

*Students are required to process more sophisticated everyday texts and understand and respond to simple factual questions; hence this is a level 2.*

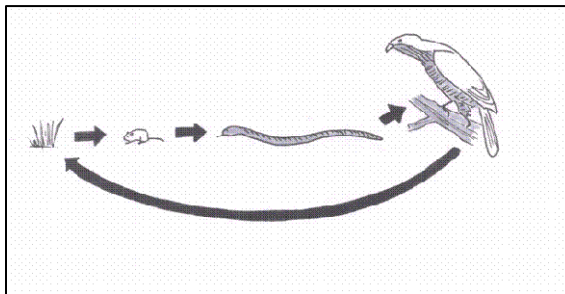
### Example 2:4 Writing

**Objective:** Write a short narrative story that includes the basic elements of setting and characters and that follows a visually supported outline provided by the teacher.

*This objective asks students to write short, basic, grade-relevant narratives making this a level 2.*

### Example 2:5 Listening

Item: Prompt



#### SCRIPT

**Narrator:** Listen to a teacher talking to her science class

**Teacher:** When we study the food chain, we might want to think of a circle. There is not a real beginning or end—it just continues to go round and round. It is a system or circle of soil, plants, animals, and then all of those things dying and becoming “soil” again. So, think about it in concrete terms. You have soil, which provides food for plants to grow. Then you have plants. They are the food of many animals. The animals eat them, and eventually the animals die. The organic material in the animals’ bodies then becomes part of the soil, which feeds new plants, which feed new animals. The circle continues around and around. These are the workings of an ecosystem – plants and animals feed each other.

Now take this model of an ecosystem that I’ve just described for you and work with a partner to describe specific plants and animals that live together exactly the way I’ve described.

**Narrator:** What does the teacher compare the food chain to?

- A A circle
- B Death
- C A house
- D Plants

*This item requires students to process the passage and then determine how “circle” fits into the discussion. The passage is relatively long and involved, but the question above taps into standard listening expectations.*

## Level 3 Complex Formulations

### Example 3:1 Listening

**Objective:** Evaluate use of materials or resources needed to complete tasks based on oral discourse

*To meet this objective, students must be involved in complex academic and social interactions and negotiations. Task specific vocabulary and discourse strategies are needed and hence this is a level 3.*

### Example 3:2 Speaking

**Objective:** In a variety of academic and social contexts, ask for or provide specific information that confirms or denies beliefs

*Here students must have complex mastery of a variety of discourse features. Successful mastery of this object exhibits a level 3 linguistic difficulty.*

### Example 3:3 Reading

**Objective:** The student will use strategies to read a variety of materials, fiction and nonfiction. Make connections between previous knowledge and/or experiences and what is read.

*Students are required to not only process a variety of reading materials, but they must also connect currently gained knowledge with previous experience or knowledge; thus, this is a level 3.*

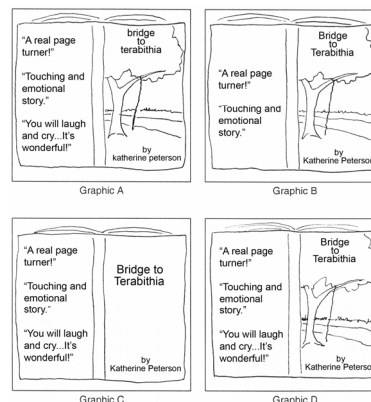
### Example 3:4 Writing

**Objective:** Write clear and coherent grade appropriate paragraphs with effective transitions and sentence structures.

*For students to meaningfully exhibit this objective, they must have grade appropriate fluency in writing. This characteristic exhibits a level 3.*

### Example 3:4 Reading

**Item: Prompt**



*Read the instructions for making a book cover that the teacher gave to her students:*

We have finished reading *Bridge to Terabithia* by Katherine Paterson. Now, you are going to prepare a book jacket for the book. Follow these instructions for designing your book jacket:

1. The title and author are placed on the front cover. Don't forget to follow the rules for capitalization of both title and author.
2. Put an illustration on the front cover. The illustration should reflect the characters, plot, or setting of the book.
3. On the back cover, include at least three positive reviews.

Which chart is made correctly?

- Graphic A
- Graphic B
- Graphic C
- Graphic D

*This task is a level 3 since students have several contextualized reading tasks, e.g., understand the notion of book jackets, process the three requested tasks and interpret each instruction relative to the tasks context.*

## Appendix D: ELL Sample Alignment Coding Sheet

Coding Form State \_\_\_\_\_ Reviewer \_\_\_\_\_ Date \_\_\_\_\_

Content Area \_\_\_\_\_ Grade \_\_\_\_\_ Test Form \_\_\_\_\_

Item #	Linguistic Difficulty	Primary Objective	Secondary Objective	Secondary Objective	Source of Challenge
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					



## Alignment Report 6

---

### Developing Aligned Performance Level Descriptors for the English Language Development Assessment K-2 Inventories

Prepared by  
María H. Malagón  
Marjorie B. Rosenberg  
Phoebe C. Winter

under contract with the Council of Chief State School Officers

December 19, 2005



## Table of Contents

<b>Developing Aligned Performance Level Descriptors for the English Language Development Assessment K-2 Inventories</b>	<b>159</b>
Procedures .....	160
Results: Revised performance Level Descriptors for ELDA k-2.....	161
<i>Listening</i> .....	161
<i>Speaking</i> .....	152
<i>Reading</i> .....	163
<i>Writing</i> .....	164
<b>Appendix A: Original Performance Level Descriptors</b>	<b>165</b>
<i>Listening</i> .....	165
<i>Speaking</i> .....	166
<i>Reading</i> .....	167
<i>Writing</i> .....	168
<b>Appendix B: Teachers Review Results, Kindergarten Reading Section</b>	<b>171</b>
Importance Ratings .....	171
PLD Assignments .....	172





## Developing Aligned Performance Level Descriptors for the English Language Development Assessment K-2 Inventories

This report describes a process used to revise performance level descriptors to better reflect the knowledge and skills covered by an observational instrument used in K-2. The assessment instrument used in the study is the English Language Development Assessment (ELDA) K-2 inventories. The process involved a review of the inventory items by a panel of experts in educating English language learners in grades K-2, who determined the level of proficiency reflected by each of three score point for each item. The panel used the knowledge and skills required to earn the group of score points assigned to each proficiency level to revise the performance level descriptors.

The ELDA is a series of tests designed to measure English language learner (ELL) students' acquisition of English language proficiency skills in kindergarten through 12<sup>th</sup> grade, developed by the Council of Chief State Schools Officers (CCSSO) and 18 state education departments through an Enhanced Assessment Grant awarded to the Nevada Department of Education. The ELDA consists of separate tests for each of the four skill domains of listening, speaking, reading, and writing, at each of four grade clusters: K-2, 3-5, 6-8, and 9-12.

The ELDA K-2 assessment includes an inventory for kindergarten and one for grades 1-2. There is some item overlap in the two inventories. The inventories contain observational items, each with four possible score points, 0 to 3. The Reading section has 14 items, the Listening section has 7 items, the Writing section has 9 items, and the Speaking section has 8 items. ELDA K-2 is intended to be used by teachers over a period of time, through in-class observation and inventory-specific tasks. An administration manual provides an explanation of each item on the inventory and sample tasks that can be used to elicit student responses relevant to the items. Figure 1 illustrates the structure of the items.

Figure 1. Sample ELDA K Item

W09 3.5 Edit writing for complete sentences.		
Score Point 1	Score Point 2	Score Point 3
Student attempts to edit writing for words or phrases with teacher support.	Student edits writing for complete sentences with teacher support.	Student edits writing for complete sentences without teacher support.

The ELDA system has Performance Level Descriptors (PLDs) for each domain, intended to be inclusive of grades K-12 (see Appendix A). The PLDs describe the knowledge and skills typical of students entering each of five levels of language acquisition: pre-functional, beginning, intermediate, advanced, and fully English proficient. While the PLDs are appropriate for grades 3-12, in some areas they do not describe the nature and sequence of language acquisition in young English Language Learners (ELLs). For example, the Writing PLD for Entry into Intermediate states:

### Entry into 3 (Intermediate)

Students at this level demonstrate some use of discourse features such as transition words and sentence order. They begin to revise for content, organization, and vocabulary. They demonstrate comprehensible use of basic sentence structures, with errors and can begin to edit for sentence-level structure. They use everyday vocabulary but know very few content-specific words. There is some variation in their register, voice, and tone. They may make frequent mechanical errors, particularly when expressing complex thoughts or technical ideas. Finally, students can compose narrative and some descriptive texts and can begin to write expository and persuasive texts.

The writing section of the ELDA inventory is targeted to the skills acquired by young ELLs, who are developing pre-writing skills, using invented spelling, and communicating their ideas orally and through pictures as well through written text. As can be seen in the example, the existing PLD does not completely match the constructs assessed by the inventory.

This study was conducted to develop PLDs for the K-2 ELDA that more closely correspond to the skills assessed by the K-2 inventory. The existing PLDs were revised, rather than rewritten, to reflect the continuity of language acquisition across the grades.

### **Procedures**

The method used to revise the PLDs consisted of several steps, conducted separately for each section of the inventories:

1. Following Kopriva, Wiley, Chen, Levy, Winter, and Corliss (2004), assign each of the three score point descriptions for each item to the current PLD it most closely measures. For example, score point 1 of item x might describe skills for an ELL functioning at Level 2 (Beginner), score point 2 may describe Level 4 (Advanced), and score point 3 may describe Level 5 (Fully English Proficient).
2. Review the skills represented by the score point descriptions in each Level and make adjustments as needed.
3. Revise the PLD for entry into each level based on the knowledge and skills assessed by the inventory, using descriptive language appropriate to the developmental nature of language acquisition in young ELLs.
4. Review the PLDs and items and make adjustments as needed.

Except in the case of Reading (see below), Steps 1 and 2 were completed by four reviewers with expertise in teaching ELLs in kindergarten through 3<sup>rd</sup> grade, in linguistics, and in educational measurement. Two of the experts, who had expertise in both teaching young ELLs and in linguistics, completed Step 3. The revised PLDs were reviewed by a subcommittee of members from states that had been involved in developing and piloting the K-2 inventories.

### **Reading Inventories**

The expert panel found that the K-2 Reading inventory score points did not fit as well into the existing PLDs as did the score points from items in the other three domains. To illustrate, the Reading PLD for Entry into Beginning is shown below:

#### **Entry into 2 (Beginning)**

Students at this level understand short and simple authentic texts for informative or social purposes (e.g., general public statements, environmental texts, formulaic messages). They have some understanding of short narrative texts or trade books, mostly when below grade level. They begin to understand some straightforward written directions. They understand main ideas and can identify a few explicit supporting ideas of simple authentic informative and narrative materials when they contain simple language structures or rely heavily on visual cues or some prior experience with topic. They have some limited understanding of text purpose. They are unable to extrapolate from text unless related to very basic ideas. They understand simple basic grammatical structures of written English in the school–social environment. They understand simple, basic, everyday vocabulary of the school environment and common, everyday activities.

Figure 2 contains two items from the Kindergarten inventory. These items illustrate the difficulty of assigning many of the Reading item score points to an existing PLD, given the lack of consonance between the skills tested and the structure of the descriptions.

Figure 2. Sample ELDA K Reading Items

<b>R03 1.1 Phonemic awareness: Recognize sounds, combinations of sounds, and meaningful differences between sounds in context.</b>		
<b>Score Point 1</b>	<b>Score Point 2</b>	<b>Score Point 3</b>
Using visuals and with teacher prompting, student demonstrates phonemic awareness by recognizing and identifying sounds, combinations of sounds, and meaningful differences between sounds in written and spoken context.	Using visuals and with teacher prompting, student demonstrates phonemic awareness by recognizing and identifying sounds, combinations of sounds, and meaningful differences between sounds in written and spoken context.	Using visuals and with teacher prompting, student demonstrates phonemic awareness by recognizing and identifying sounds, combinations of sounds, and meaningful differences between sounds in written and spoken context.
- initial sounds	- initial sounds	- initial sounds
- ending sounds	- ending sounds - short vowels	- ending sounds - short vowels
	- long vowels ( ai, ay, ee, ea, igh, oa ,oe, ow, ue, ui, ew)	- long vowels - word families
<b>R12 2.2 Demonstrate understanding of directionality of print.</b>		
<b>Score Point 1</b>	<b>Score Point 2</b>	<b>Score Point 3</b>
With teacher prompting, student shows an initial awareness of print as a form of meaningful communication by consistently demonstrating the following skills:	With teacher prompting, student shows an increased awareness of print as a form of meaningful communication, but is NOT consistent in demonstrating all of the following skills:	With teacher prompting, student shows an awareness
<i>Holds book correctly Turns pages sequentially</i>	<i>Holds book correctly Turns pages sequentially Follows print from top to bottom</i>	of print as a form of meaningful communication by CONSISTENTLY demonstrating all of the following skills:
	<i>Follows print from left to right</i>	<i>Holds book correctly Turns pages sequentially Follows print from top to bottom</i>
		<i>Follows print from left to right</i>

Because the purpose of this study was to revise, rather than rewrite, the PLDs, the researchers consulted with additional experts who were familiar with the K-2 inventories. Eleven current and former K-2 ESL teachers reviewed the Kindergarten Reading items and (1) independently assigned to each item a level of importance for assessing the reading acquisition of ELL students and (2) in groups, determined which of the existing PLDs most closely matched each score point on the items (see Appendix B). In addition, the researchers worked with a panel of educators from states that had developed and piloted the K-2 inventories to revise the Reading PLDs.

## **Results: Revised Performance Level Descriptors for ELDA K-2**

### **LISTENING**

#### **Entry into 5 (FEP)**

Students at this level understand most grade-level appropriate content-area and school/social speech. They understand the main ideas and relevant details of extended discussions or oral presentations on a range of familiar and unfamiliar topics comparable to a native English speaker at the same grade level. They are capable of making interpretations of what they hear. They understand most of the complex structures of spoken English relative to their grade level. They have a broad range of vocabulary, including idiomatic language, relating to both content areas and school/social environments.

**Entry into 4 (Advanced)**

Students at this level understand conversations in most school/social settings. They understand main ideas and significant relevant details of extended discussions or presentations on familiar and relevant academic topics. They are able to comprehend conversations and orally-delivered texts involving description and narration in different time frames or conditions. They understand most of the basic language forms of spoken English including timeless conditionals and sentences using clauses and phrases. They are able to understand cohesive devices to follow the sequence in an oral presentation or text. They comprehend most grade-level vocabulary and idioms, especially school/social environments, and are beginning to develop a wide range of academic vocabulary related to content areas, with limited supports such as visuals and rephrasing. They understand multiple meanings of words and can use context clues to understand messages.

**Entry into 3 (Intermediate)**

Students at this level understand sentence-length statements and questions that include recombinations of learned language structures and on a variety of social and academic topics. They understand simple and compound sentences. They understand time through the use of simple tenses that may not be supported by adverbials of time. They are able to understand multi-step directions. They also understand the difference between statements and questions by intonation, word order, and interrogative words. They understand and are able to identify main ideas and some details from conversations and simple/age appropriate orally-delivered text, usually with visual supports in familiar communicative situations and in academic content areas. They begin to interpret meaning from conversations and orally-delivered text, making predictions and drawing conclusions. They understand some idioms, mostly related to school/social environments, and have key vocabulary from content areas. They are aware of cohesive devices but may not be able to use them to follow the sequence of thought in an oral text.

**Entry into 2 (Beginning)**

Students at this level understand simple, short statements and questions on a well-known topic within a familiar context. Tense is understood through the use of adverbials or situation rather than inflectional endings. They are able to follow simple multi-step directions. They identify the main idea and some details of short conversations or simple orally-delivered text on a familiar topic. They understand basic grammatical structures and vocabulary in the school and social environment. Students at this level still need frequent repetition and rephrasing. They understand what they have heard but not variations or recombinations of what they have heard.

**0→1 (Pre-functional)**

Students at this level may understand some isolated words (particularly school and social environment vocabulary), some high frequency social conventions, and simple (single word or short phrase) directions, commands and questions. They rely on non-verbal cues such as gestures and facial expressions and require frequent repetition and rephrasing to understand spoken language. They need strong situational support to understand most oral language.

**SPEAKING****Entry into 5 (FEP)**

Students who are ready to enter Level 5, Fully English Proficient, can supply coherent, unified and appropriately sequenced responses to an interlocutor. They use a variety of devices to connect ideas logically. They understand and can use a range of complex and simple grammatical structures, as appropriate for topic and type of discourse. Their grammar and vocabulary is comparable to that of a minimally proficient native English speaker—grammar errors very seldom impede communication and their range of school-social and academic vocabulary allows a precision of speech comparable to a native English speaker. They can effectively engage in non-interactive speech. They can use language effectively to connect, tell, expand, and reason. They show flexibility, creativity and spontaneity in speech in a variety of contexts.

**Entry into 4 (Advanced)**

Students entering proficiency Level 4, the Advanced level, are able to restructure the language they know to meet the creative demands of most social and academic situations. They can supply mostly coherent, unified and appropriately sequenced responses to an interlocutor. They use some devices to connect ideas logically and they use a range of grammatical structures. They make some errors in modality, tense, agreement, pronoun use, and inflections. Students have sufficient vocabulary to communicate in non-academic situations and most academic ones. They can engage in extended discussions. They can often use language to connect, tell and expand on a topic; and can begin to use it to reason. They are fluent but may still hesitate in spontaneous in communicative situations.

**Entry into 3 (Intermediate)**

Students entering proficiency level 3, the Intermediate level, are no longer wholly dependent on practiced, memorized, or formulaic language. They restructure learned language to communicate on a range of subjects. Their speech is still marked by errors in modality, tense, agreement, pronoun use, and inflections. These errors seldom interfere with communication in simple sentences, but do interfere in complex constructions. Intermediate level students are limited in vocabulary, especially academic vocabulary. They can retell, describe, narrate, question, and give instructions, although they lack fluidity and fluency when not using practiced or formulaic language. They often use language to connect, tell and sometimes to expand on a known topic.

**Entry into 2 (Beginning)**

Students who are just entering proficiency level 2, the beginning level, predominantly use formulaic patterns and memorized phrases. When they deviate from formulaic language, their speech imitates telegraphic language due to the omission of some meaningful linguistic components. Their language is also marked by the lack of tense, number, and agreement. They may use some very simple transitional markers, usually “and” to link ideas. They rely on schemata in L1. Their school-social vocabulary is limited to key words and they have little or no academic vocabulary. They respond to questions usually with one or two-word answers. They can connect and tell on a known topic.

**0→1 (Pre-functional)**

Students at this level may say or repeat common phrases, words and formulaic language. They may be able to provide some basic information in response to requests and questions. They can ask one or two-word questions without regard to structure and intonation.

## **Reading**

**Entry into 5 (FEP)**

Students at this level participate in reading activities with little teacher support at a level comparable to their English-speaking peers. They read for different purposes across a variety of text types. They have an increasing range of receptive nonacademic and academic vocabulary that allows them to read with greater fluency. They understand multiple word meanings. They have greater comprehension as a result of their increasing control of the structures of English. They can make connections between what they read and other experiences and tasks.

**Entry into 4 (Advanced)**

Students at this level can read familiar text with little teacher or visual support. However, they still need those supports when reading to comprehend unfamiliar text. They can apply their phonemic awareness skills to read more complicated text. They have oral fluency and use self-monitoring and self-correction strategies when necessary. They use pre, during and post reading strategies but still need teacher prompting to use these skills. They can identify all story elements and can recognize cause and effect relationships in the texts they read. They make connections between the texts they read and themselves, the world, and other texts. They comprehend text in read aloud and can participate in the majority of read aloud activities. They are beginning to read across text types and apply what they read to other activities.

**Entry into 3 (Intermediate)**

Students at this level are developing phonemic awareness skills that allow them to read single words and simple text with comprehension. Reading is aided by visual and teacher supports. At this stage oral reading is hesitant and difficult to understand due to a lack of oral language proficiency. These students have a small repertoire of high frequency words. They are beginning to use simple reading strategies and to make self, world, and other text connections to the text they are reading. They comprehend simple sentence structure and sentences with simple compounding. They recognize that words serve different functions, have multiple meanings, and have both synonyms and antonyms. In read aloud, with teacher support, they can identify some story elements and retell the majority of the story.

**Entry into 2 (Beginning)**

Students at this level begin to identify the names of both upper and lower case letters of the alphabet. They use juncture to identify where words begin and end. They begin to recognize that words serve different functions (e.g. nouns, verbs). They can follow multi-step directions depicted graphically. During read aloud they get meaning primarily from pictures and the teacher's tone of voice and gestures.

**0→1 (Pre-functional)**

Students at this level demonstrate an understanding of concepts of print (e.g., front-to-back, top-to-bottom, left-to-right) and begin to track print. They can distinguish letters from other symbolic representations. They can follow one-step directions depicted graphically. They can imitate the act of reading (e.g. holding a book and turning pages); however, they get meaning only through pictures.

## **WRITING**

**Entry into 5 (FEP)**

Students at this level participate in writing activities with no teacher support. They write across all text types. They edit for sentence-level structure, spelling, and mechanics and revise for content, organization and vocabulary. They can use complex sentence structures, with some errors, and can edit for syntax and grammar. They have a range of nonacademic and academic vocabulary that allows for precision, and they begin to use nuanced and alternative word meanings. They employ subtleties for different audiences and purposes. They can use appropriate writing conventions with some errors that do not affect comprehensibility.

**Entry into 4 (Advanced)**

Students at this level participate in writing activities with minimal teacher support. They are able to restructure in writing the language they know to meet the creative demands of most social and academic situations. They can write mostly coherent, unified, and appropriately sequenced sentences. They use devices to connect ideas logically. They use a range of grammatical structures and can switch appropriately from one tense to another as required by the time frame of their text. They make some errors in modality, tense, agreement, pronoun use, and inflections. Students have a strong BICS vocabulary and a functional academic vocabulary that allows them to participate meaningfully in content classes. They write using all text types, at a developmentally appropriate level. They edit for sentence-level structure, spelling, and mechanics and revise for content, organization and vocabulary.

**Entry into 3 (Intermediate)**

Students at this level participate in writing activities with some teacher support. They can write simple and compound sentences and are beginning to write with phrases. They use simple tenses, number, and agreement with random errors. They use transition words to link sentences and order these in a developmentally appropriate manner. They begin to edit for sentence-level structure, spelling and mechanics and revise for content, organization and vocabulary, usually with the support of the teacher. They have a good range of BICS vocabulary and are beginning to use more academic content-specific words. They write mostly descriptive, expository, procedural, and narrative text. Their writing is less dependent on visual supports, shared experiences, and scaffolding.

### **Entry into 2 (Beginning)**

Students at this level participate in writing activities by drawing pictures or dictating words. They are able to write connected words and short telegraphic sentences. They are able to revise or edit their writing with teacher support. Their writing is marked by the lack of tense, number, and agreement. They may use some simple transitional markers, usually “and” to link ideas. Their vocabulary reflects what they can say orally. They make frequent errors in mechanics such as punctuation and capitalization. They write mostly descriptive, expository, and procedural text. Their writing is most effective when supported by a visual, a shared experience, or scaffolding.

### **0→1 (Pre-functional)**

Students at this level participate in writing activities by drawing pictures. They may be able to copy letters or form them from memory and may be able to copy some words. They can imitate the act of writing (e.g. scribbling); however, their text does not transmit a message. They may attempt to apply some writing conventions but do so inappropriately or do so correctly only when copying.





## APPENDIX A: Original Performance Level Descriptors

### LISTENING

#### Entry into 5 (FEP)

Students at this level understand a significant amount of grade-level appropriate content-area and school-social speech. They understand the main ideas as well as relevant details and often subtle nuances of meaning of extended discussions or presentations on a range of familiar and unfamiliar topics comparable to a minimally proficient native English speaker at the same grade level. They are capable of making interpretations of what they listen to on the basis of understanding the speaker's purpose. They understand most of the complex structures of spoken English relative to their grade level. They have a broad range of vocabulary, including idiomatic language, relating to both content areas and school-social environments.

#### Entry into 4 (Advanced)

Students at this level understand speech in most school-social settings and understand main ideas and some key supporting ideas in content-area settings. They understand multi-step directions. They understand main ideas and significant relevant details of extended discussions or presentations on familiar and relevant academic topics. They can interpret text on the basis of understanding the purpose of text when it is on a familiar topic. They understand and are able to make subtle extrapolations from sophisticated speaker perspectives. They understand most of the basic language forms of spoken English and are beginning to develop understanding of more complex structures. They understand a wide range of vocabulary and idioms, especially of school-social environments, and are beginning to develop a wide range of technical vocabulary related to content areas.

#### Entry into 3 (Intermediate)

Students at this level understand main ideas in short conversations on general school-social topics and frequently demonstrate general understanding of short messages or texts as well as longer conversations in familiar communicative situations and in academic content areas. They frequently demonstrate detailed understanding of short discrete expressions but not of longer conversations and messages. They understand single-step and some multi-step directions. They can begin to interpret text on the basis of understanding its purpose. They understand some explicitly expressed points of view and can draw simple conclusions. They understand frequently used verb tenses and word-order patterns in simple sentences. They understand a range of vocabulary and some idioms, mostly related to school-social environments, and have some key vocabulary from content areas.

#### Entry into 2 (Beginning)

Students at this level understand simple and short statements, questions, and messages on familiar topics in school-social settings, and usually understand the main idea of simple messages and conversations. They can understand most common or critical information in the classroom but may identify and understand only key words, phrases, and cognates in content-area settings. They begin to understand straightforward, single-step directions and speaker's purpose. They have limited understanding of details and only of those that are explicitly stated and that support simple, straightforward messages or presentations. They are unable to extrapolate from text unless related to very basic ideas. They understand simple, basic grammatical structures and simple, basic, everyday vocabulary of spoken English in the school environment and common, everyday activities.

#### 0–1 (Pre-functional)

Students at this level may understand some common words or key phrases, especially when highly contextualized or when cognates. They may understand some high-frequency single-word or single-phrase directions, again, when highly contextualized. They generally are unable to use their limited knowledge of simple structural patterns to identify the communicative intent of the speaker.

## **SPEAKING**

### **Entry into 5 (FEP)**

Students who are ready to enter Level 5, Full English Proficient, can supply coherent, unified, and appropriately sequenced responses to an interlocutor. They use a variety of devices to connect ideas logically. They understand and can use a range of complex and simple grammatical structures, as appropriate for topic and type of discourse. Their grammar and vocabulary are comparable to that of a minimally proficient native English speaker—grammar errors very seldom impede communication and their range of school-social and technical vocabulary allows a precision of speech comparable to a minimally proficient native English speaker. They infrequently but effectively use circumlocution. They can understand and use a variety of idiomatic phrases. They can effectively engage in non-interactive speech. They can use language effectively to connect, tell, expand, and reason. They show flexibility, creativity and spontaneity in speech in a variety of contexts. Their pronunciation patterns (including stress and intonation) may be influenced by L1 but seldom interfere with communication.

### **Entry into 4 (Advanced)**

Students entering proficiency Level 4, the Advanced level, can supply mostly coherent, unified, and appropriately sequenced responses to an interlocutor. They use some devices to connect ideas logically and they use a range of grammatical structures. They make errors in modality, tense, agreement, pronoun use, and inflections, but these errors usually do not interfere with communication. Students have sufficient vocabulary to communicate in nonacademic situations and some academic and technical vocabulary. They use circumlocutions and can appropriately use some idiomatic phrases. They can engage in extended discussions. They can often use language to connect, tell, and expand; and can begin to use it to reason. Their flexibility, creativity, and spontaneity are sometimes adequate for the communicative situation. Their pronunciation occasionally interferes with communication.

### **Entry into 3 (Intermediate)**

Students entering proficiency level 3, the Intermediate level, display some use of discourse features but mainly rely on familiar, discrete utterances. They rely on simple transitional markers and use common, straightforward grammatical structures. They make errors in modality, tense, agreement, pronoun use, and inflections. These errors seldom interfere with communication in simple sentences, but do interfere in complex constructions or when talking about academic issues. Intermediate level students are limited in vocabulary, especially academic and technical vocabulary. They use repetition; everyday, imprecise words; and code-switching to sustain conversations. They begin to use idiomatic expressions. They can retell, describe, narrate, question, and give simple, concrete instructions. They can often use language to connect and tell and sometimes to expand. They have some creativity and flexibility but often repeat themselves and hesitate. Their pronunciation patterns frequently interfere with communication.

### **Entry into 2 (Beginning)**

Students who are just entering proficiency level 2, the beginning level, use predominantly formulaic patterns in speech without regard to their connectivity. They may use some very simple transitional markers. They predominantly use formulaic patterns and memorized phrases, relying on schemata in L1. Their word order is frequently inappropriate, and frequent grammatical mistakes impede communication. Their school-social vocabulary is limited to key words; they have little or no technical vocabulary. They rely on survival vocabulary (needs and wants) and vocabulary provided by interlocutors. They may be able to name or list and can sometimes use language to connect or tell. Their limited vocabulary and knowledge of English structures impedes flexibility.

### **0–1 (Pre-functional)**

Students in proficiency level 1 are not yet at a functional level in English. They may repeat common phrases with very simple structures; be able to say a few, common, everyday words; and may be able to provide some basic information in response to requests.

## READING

### Entry into 5 (FEP)

Students at this level understand the range of texts available to minimally proficient native English speakers, including literary and academic genres and texts from school-social settings. They understand main ideas and can extract precise and detailed information from a range of texts on familiar and unfamiliar topics in a number of genres comparable to a minimally proficient native English reader at the same grade level. They often successfully interpret text on the basis of understanding its purpose. They often successfully understand and can evaluate multiple perspectives of meaning. They understand complex structures of written English and have a broad range of vocabulary and idioms relating to both content areas and school-social environments.

### Entry into 4 (Advanced)

Students at this level understand most nonacademic and non-technical texts appropriate for grade level. They understand many content area texts, mostly on familiar topics and approaching grade level. They understand excerpts from literature. They understand most written directions. They understand main ideas of a broad range of texts, especially when below grade level but also when approaching grade level. They can begin to interpret text on the basis of understanding its purpose. They understand significant relevant details and can make subtle extrapolations of extended narratives or presentations on familiar academic topics. They understand sophisticated writer perspectives. They understand most of the basic language forms of written English and are beginning to develop understanding of more complex structures. They understand a wide range of vocabulary and idioms, especially of school-social environments, and are beginning to develop a wide range of technical vocabulary related to content areas.

### Entry into 3 (Intermediate)

Students at this level understand many authentic narrative and descriptive texts, especially when below grade level but with less complete comprehension for such texts on grade level. They understand content-area texts with familiar content, mostly when below grade level. They understand excerpts from literature especially when below grade level. They understand simple written directions as well as some more complexly expressed directions. They understand main ideas of narrative and descriptive texts and some of the main points of expository and persuasive texts when they deal with areas of personal interest or topic familiarity. They begin to understand text purpose. They can understand some supporting ideas of expository and persuasive texts when dealing with areas of special interest. They understand some explicitly expressed points of view of writer and are able to draw simple conclusions. They understand frequently used verb tenses and word-order patterns in simple sentences. They understand a range of vocabulary and some idioms, mostly related to school-social environments, and have some key vocabulary from content areas.

### Entry into 2 (Beginning)

Students at this level understand short and simple authentic texts for informative or social purposes (e.g., general public statements, environmental texts, formulaic messages). They have some understanding of short narrative texts or trade books, mostly when below grade level. They begin to understand some straightforward written directions. They understand main ideas and can identify a few explicit supporting ideas of simple authentic informative and narrative materials when they contain simple language structures or rely heavily on visual cues or some prior experience with topic. They have some limited understanding of text purpose. They are unable to extrapolate from text unless related to very basic ideas. They understand simple basic grammatical structures of written English in the school-social environment. They understand simple, basic, everyday vocabulary of the school environment and common, everyday activities.

### 0–1 (Pre-functional)

Students at this level may identify isolated words and key phrases and cognates, especially when highly contextualized. They may understand some high-frequency, simple written directions, especially when highly contextualized. They are unable to identify any ideas intended by writer of text or to use limited knowledge of vocabulary and structural patterns to identify communicative intent of text or part of text. They do not understand how words, morphemes, and word order convey meaning in English.

## WRITING

### Entry into 5 (FEP)

Students at this level demonstrate almost completely appropriate use of discourse features such as transition phrases and word order. They can revise for content, organization, and vocabulary. They can use complex sentence structures, with some errors, and can edit for syntax and grammar. They have a range of technical and nonacademic vocabulary that allows for precision and they begin to use nuanced and alternative word meanings. They employ subtleties for different audiences and purposes. They can use appropriate writing conventions with some circumlocutions and errors that do not affect comprehensibility. Finally, they can successfully compose narrative, descriptive, expository, and persuasive texts.

### Entry into 4 (Advanced)

Students at this level demonstrate mostly successful use of discourse features such as transition words and sentence order. They can revise for content, organization, and vocabulary and show good control of the most frequently used grammatical structures, with errors. They can edit for sentence-level structure. They have sufficient vocabulary to express themselves with some circumlocutions, which are more frequent in academic contexts. Their tone indicates some awareness of audience. They can use appropriate writing conventions, with circumlocutions and errors that infrequently affect comprehensibility. Finally, they can successfully compose narrative and descriptive texts and they may be successful writing expository and persuasive texts.

### Entry into 3 (Intermediate)

Students at this level demonstrate some use of discourse features such as transition words and sentence order. They begin to revise for content, organization, and vocabulary. They demonstrate comprehensible use of basic sentence structures, with errors and can begin to edit for sentence-level structure. They use everyday vocabulary but know very few content-specific words. There is some variation in their register, voice, and tone. They may make frequent mechanical errors, particularly when expressing complex thoughts or technical ideas. Finally, students can compose narrative and some descriptive texts and can begin to write expository and persuasive texts.

### Entry into 2 (Beginning)

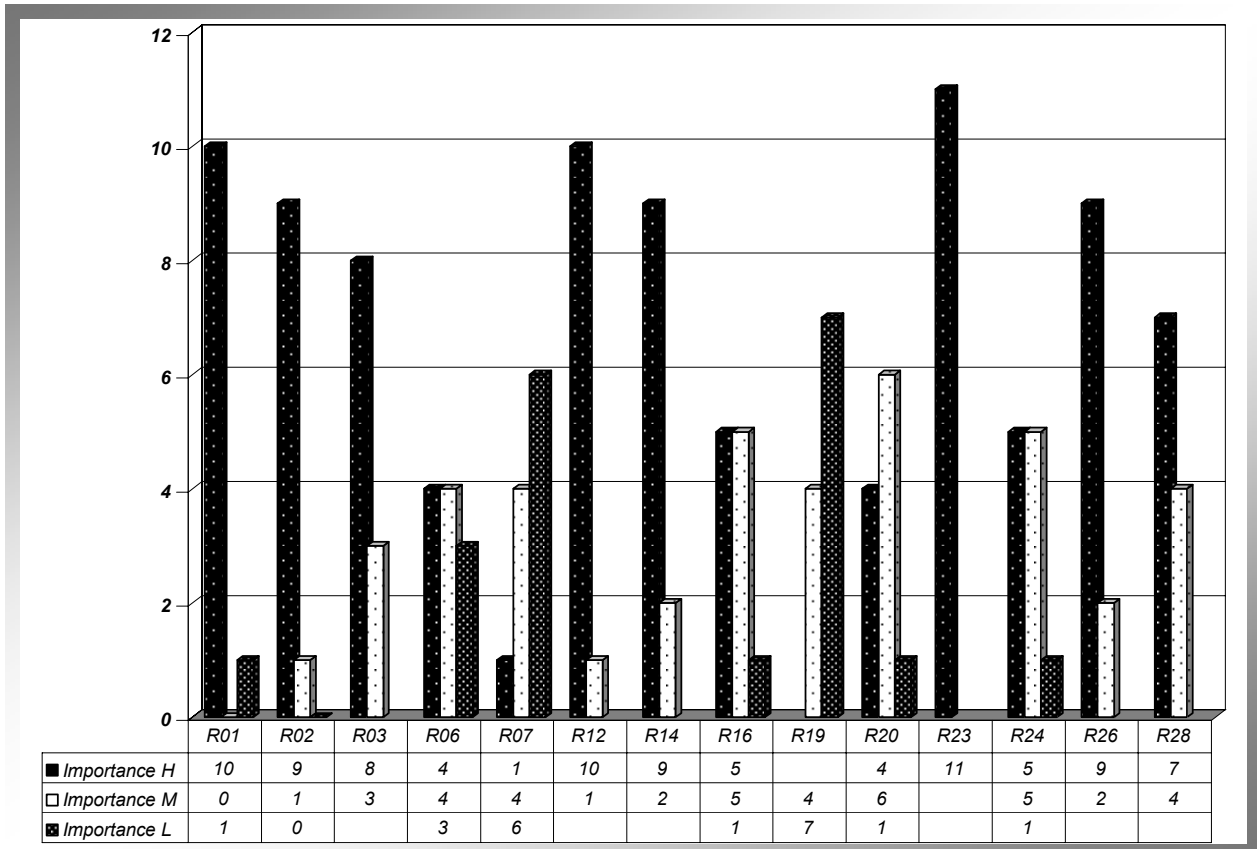
Students at this level may or may not use some basic rhetorical features such as ordering sentences appropriately and using simple cohesive devices. They are unlikely to revise their writing spontaneously. Their writing is limited to typical, present-tense, subject-verb-object sentences or phrases and is likely to be repetitive. They edit only with explicit support and direction and have a limited vocabulary. They make frequent errors in mechanics, which is characteristic and expected. Their text range is limited to narrative or simple descriptive.

### 0–1 (Pre-functional)

Students at this level are not yet functional in English. They might be able to copy letters or form them from memory and might be able to write words; however, their text does not transmit a coherent message. They do not use discourse features in their writing. There is no evidence of appropriate text structure and sentence-level structure is predominantly inappropriate. They may attempt to apply some writing conventions but do so inappropriately or do so correctly only when copying.

## APPENDIX B: Teacher Review Results, Kindergarten Reading Section

### Importance Ratings



**PLD Assignments\***

Group 1						Group 2					
Item and Score Point	Level needed to earn score point					Item and Score Point	Level needed to earn score point				
	Pre-Func.	Beg.	Inter.	Adv.	FEP		Pre-Func.	Beg.	Inter.	Adv.	FEP
<b>R01</b>						<b>R01</b>					
1		X				1		X	X		
2				X		2			X	X	
3					X	3					X
<b>R02</b>						<b>R02</b>					
1		X	X			1		X	X		
2			X	X		2			X	X	
3			X	X		3					X
<b>R03</b>						<b>R03</b>					
1		X				1		X	X		
2				X		2			X		X
3					X	3					X
<b>R06</b>						<b>R06</b>					
1		X				1			X		
2			X			2			X		
3			X			3			X		
<b>R07</b>						<b>R07</b>					
1			X			1			X		
2				X		2				X	
3				X		3					X
<b>R12</b>						<b>R12</b>					
1		X				1	X				
2		X				2		X			
3		X				3		X			
<b>R14</b>						<b>R14</b>					
1	X					1	X				
2		X				2	X				
3		X				3		X			

\* If group members disagreed, there are multiple assignments per score point.

Group 3						Group 4					
Item and Score Point	Level needed to earn score point					Item and Score Point	Level needed to earn score point				
	Pre-Func.	Beg.	Inter.	Adv.	FEP		Pre-Func.	Beg.	Inter.	Adv.	FEP
<b>R16</b>						<b>R16</b>					
1			X			1			X		
2				X		2				X	
3					X	3					X
<b>R19</b>						<b>R19</b>					
1		X				1			X		
2			X			2				X	
3				X		3					X
<b>R20</b>						<b>R20</b>					
1		X				1	X				
2		X				2		X			
3			X			3			X		
<b>R23</b>						<b>R23</b>					
1		X				1		X			
2		X				2		X			
3			X			3			X		
<b>R24</b>						<b>R24</b>					
1		X				1		X			
2			X			2			X		
3				X		3				X	
<b>R26</b>						<b>R26</b>					
1			X			1			X		
2				X		2				X	
3					X	3					X
<b>R28</b>						<b>R28</b>					
1			X			1		X			
2				X		2			X		
3					X	3					X