Executive Summary

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)

Norman L. Webb
Senior Research Scientist
Wisconsin Center for Education Research
http://facstaff.wcer.wisc.edu/normw/

In the spring of 2002, the Technical Issues in Large-Scale Assessment (TILSA) group of the State Collaborative on Assessment & Student Standards (SCASS) organized by the Council of Chief State School Officers (CCSSO) undertook a one-year study of a process for analyzing the alignment between state curriculum standards and assessments. The study was guided by two goals:

1. To provide participating states with data on the degree to which their curriculum standards and state assessments were aligned; and,
2. To analyze the process for conducting an alignment analysis, identify critical issues that arise during this process, and produce data that can be used to resolve these issues.

Along with these goals, the study sought to provide answers to practical questions such as:

1. What information should be produced by an alignment study to provide effective feedback to a state relative to achieving full alignment?
2. What are reasonable acceptable levels to expect on each alignment criterion?
3. What alignment criteria are most effective for providing feedback to states that will enable them to strengthen their systems of standards and assessments?
4. How can procedures for studying alignment be adapted that will enable a state to continually monitor alignment and to train item writers to align their items more effectively?
5. What training materials are needed for a state to conduct an alignment study?

States were informed of the study and were given the opportunity to voluntarily agree to participate, including financing some of the costs for the study. Four states agreed to participate in language arts, three in mathematics, two in science, and two in social studies. Two states participated in the study in all four content areas.

At a May 20-24, 2001 institute held in Alexandria, Virginia, reviewers external to any of the participating states and reviewers from the participating states were trained in the coding process and then analyzed both the standards and assessments. Reviewers did not analyze the documents from their own state other than in replication studies. For two states for one grade each, a replication study was conducted in mathematics in which two groups of reviewers independently coded the assessment items. Three to six reviewers

examined each of the standards and assessments. The grade levels in language arts analyzed in this study by state were:

State E	Grades 4, 7, and 10
State F	Grades 5, 8, and 11
State G	Grades 4, 8, and 11
State H	Grades 4, 5, 6, and 9

In mathematics, three grade levels were analyzed for two states and two grade levels for one state:

State E	Grades 4, 7, and 9
State F	Grades 5 and 8
State G	Grades 4, 8, and 11

A second institute was conducted in Madison, Wisconsin, on October 22-24, 2002 to analyze standards and assessments in the content areas of science and social studies. At this institute, two groups of reviewers coded each of the state's standards and assessments. One group included only reviewers external to the state (external group), while the other group included reviewers from the state's department of education (internal group). The number of reviewers in these groups ranged from two to five. The grade levels in science analyzed in this study by state were:

State E	Chemistry (high school) and Biology (high school)
State F	Grade 8

In social studies, history standards and assessments at two grade levels were analyzed for two states:

State E	Grades 10 and 11
State F	Grade 8 and End of Instruction (high school)

Webb developed the analysis process for the National Institute for Science Education in cooperation with CCSSO in 1997, which was further refined in a four state study conducted in 1999 (http://facstaff.wcer.wisc.edu/normw/state%20alignment%20page%20one.htm). The process includes comparing curriculum standards and assessments on four criteria: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation. The main tasks for the reviewers is to code the depth-of-knowledge levels for objectives under each standard using a consensus process, assigning a depth-of-knowledge level to each item, and coding for each item the corresponding primary and secondary objectives. These data then are entered into an Excel file and analyzed. These analyses produce percentages or an index for each of the four criteria that are compared to a priori-specified acceptable levels.

Results from the study indicated that the process performed well to produce consistent findings. The alignment between curriculum standards and assessments varied by state, content area, and criteria. Mathematics had the greatest alignment. Of the eight analyses performed, six (75%) were aligned. An acceptable level on each of the four criteria was met at about the same rate across criteria, around 75%. In language arts, only about one third of the grade level analyses performed indicated that the standards and assessments were fully aligned. There was strong alignment for two states and poor alignment for the other two states. There were similar results in science, where in only one of three grade levels analyzed were standards and assessments found to be aligned. In

social studies, none of the four analyses indicated that the standards and assessments were aligned.

In general, the agreement among reviewers in assigning depth-of-knowledge levels to assessment items was high. The intraclass correlation among reviewers for 13 language arts analyses ranged from .36 to .92, with an average of .73. The average intraclass correlation among reviewers for ten mathematics analyses ranged from .47 to .93 with an average of.79. For science, the intraclass correlation among reviewers for six analyses ranged from .40 to .81, with an average of .68. For social science, the intraclass correlation among reviewers for eight analyses ranged from .53 to .93, with an average of .76.

Two groups independently conducted an analysis on the same assessment and standards for two grade levels. This replication study in one analysis produced exact agreement between the two groups on 15 of the 16 criterion-by-standard judgments in attaining an acceptable level. In the second analysis, the two groups had exact agreement on 19 of 28 criterion-by-standard judgments (68%), near agreement on two judgments, and no agreement on seven judgments.

For science and social studies, two groups of reviewers independently coded each of the assessments and standards. One group contained only reviewers from outside of the state (external group) and one group containing reviewers of the standards and assessments from the state being analyzed. Overall, there were no systematic differences between the coding by an external group of reviewers and the coding by an internal group of reviewers that would suggests that one group coded items differently from the other. The differences that did exist between the two types of reviewers were more associated with a specific group of reviewers than due to the fact one group was from the state being analyzed. This implies that findings from an internal group of reviewers can have validity, but the findings may have more credibility if at least some reviewers are external to the state whose standards are being analyzed

In conclusion, the process does produce information that distinguishes different qualities in the relationship between the assessments and standards in different states. The process does not produce all of the desired information, particularly the *degree to which* the content of an objective or standard is measured by an assessment item. Reviewers were able to code the depth-of-knowledge levels of objectives and items with fairly high consistency. The acceptable levels for each criterion were set with a certain rationale, but in the end they are only based on judgment. What an acceptable level of congruence is between standards and assessments still is an open question. The process described here remains under development. The results from the studies for the CCSSO-TILSA group will be used to make revisions in the process. There are principles of testing implied in the process that would be useful considerations for assessment development, such as attending to the depth-of-knowledge levels. More formal studies of alignment are needed in which different processes are used to analyze the same sets of standards and assessment to help determine more readily what the differences and similarities are among the different methods and how valid each one is.