# Summary of External Review:
# Idaho Standards Achievement Tests (ISAT)

**May 2005**

**Based upon research conducted by the**
**Human Resources Research Organization**
**Alexandria, VA**

## Overview

In May 2004, the Idaho State Board of Education released a Request for Proposals (RFP) for an external review of the Idaho Standards Achievement Tests (ISAT). The purpose of this review was to ensure that the ISAT was meeting the criteria of the federal law as well as the rigor necessary for implementing a state exit exam.

All states are required to ensure the assessments used to meet the requirements of the "No Child Left Behind" Act are "aligned with the State's challenging academic content and student academic achievement standards" and "be used for purposes for which such assessments are valid and reliable, and be consistent with relevant, nationally recognized professional and technical standards" (Public Law 107-110, §1111.c.3.c.ii-iii).

In September 2004, the Human Resources Research Organization (HumRRO), based in Alexandria, Virginia, was selected through a competitive bid process and began work on the external review of the ISAT. The RFP specified that HumRRO would review content validity, curriculum validity and instructional validity, reliability, and conduct an alignment study. Some of the reports were a review of the existing evidence from the state assessment contractor, Northwest Evaluation Association (NWEA), and the state. Other reports were a collection of original evidence or an independent analysis of the data. HumRRO completed the work in phases and delivered all final reports to the Office of the State Board of Education.

## Contracted Studies

I. ISAT Reliability

Report: Idaho Standards Achievement Test: Independent Calculations of Reliability Estimates, Standard Errors of Measurement, Classification Accuracy, and Classification Consistency

A. Existing NWEA documentation

B. Original calculation of test reliability statistics

II. Content Validity

A. Existing NWEA documentation

Report: Idaho Standards Achievement Test: Review of Northwest Evaluation Association Content Validity Documentation, Task 3.0.1

B. Collection of original evidence

Report: Idaho Standards Achievement Test: Review of Evidence of Content Validity, Task 3.0.2

III. Curriculum Validity and Instructional Validity

A. Existing NWEA documentation

Report: Idaho Standards Achievement Test: Review of Northwest Evaluation Association Curriculum Validity and Instructional Validity Documentation, Task 3.0.4.1

B. Collection of original evidence

1. Self-study of school districts

Report: Idaho Standards Achievement Test: District Self-Study (DSS) Findings Regarding Curricular and Instructional Validity, Task 3.0.4.2

Report: Idaho District Self-Study (DSS): Executive Summaries, Task 3.0.4.2

2. On-site visitations of selected school districts

Report: Idaho Standards Achievement Test: Curriculum Validity and Instructional Validity Study Results from On-Site Visits to Six Idaho Public School Districts, Task 3.0.4.2

IV. ISAT Alignment with Idaho Content Standards

Report: Idaho Standards Achievement Test (ISAT): Test Alignment Study, Task 3.1

A. Existing NWEA documentation

B. Collection of original evidence

V. ISAT Mapping

Report: Idaho Standards Achievement Test: Item Mapping for the 2004 Idaho Standards Achievement Tests

# Terminology

**Content Standards** – In Idaho, Content Standards are referred to as the Idaho Achievement Standards. They can be found in State Board rule, Idaho Administrative Procedures Act (IDAPA), Rules Governing Thoroughness, Section 200 at: http://www2.state.id.us/adm/adminrules/rules/idapa08/0203.pdf. Content Standards define what students in Idaho are expected to know and be able to do by subject and by grade. Standards are required as part of the "No Child Left Behind" Act.

*Example: A student in 8th grade math should be able to "Perform computations accurately." More specifically, the standards are divided into Content, Knowledge and Skills that further delineate the skill: "Consistently and accurately add, subtract, multiply, and divide rational numbers."*

**Blueprint** -- The blueprint is built from the state Content Standards. This document specifies the particular content that will be included on the statewide assessment. In addition, it defines the type of items to be used and the distribution of items across the various instructional strands within the Content Standards. In each grade there are some standards that do not lend to an evaluation in a multiple-choice format; however, these standards are still important and should be taught in the classroom. Classroom instruction for all of the standards is essential for students to achieve desired results on the Idaho Standards Achievement Tests (ISAT), because many of the tested standards assume knowledge of the skills not on the test. The current science blueprint is at: http://www.boardofed.idaho.gov/saa/documents/ISAT_ScienceBlueprint.pdf.

A test blueprint usually specifies the extent of content within the tests, the range of difficulty for test items, and the structure of the test.

*Example: In 5th grade, the science test will include items that test a student's knowledge "that a system is an organized group of related objects that form a whole."*

**Reporting Goals** – NWEA's test structure includes six to seven "goal areas" for each test. These goal areas group the state Content Standards. Each goal has from 6-8 questions on the content area under the goal.

*Example: A reporting goal for math is "Measurement." A reporting goal for reading is "Word Analysis."*

**Achievement Levels** – Idaho has four achievement levels: advanced, proficient, basic and below basic. Each of these levels defines the level of performance of an individual student based on the test scores. A label of proficient means that a student has mastered skills that allows him/her to function at his/her specific grade level. For a description of all achievement levels see: http://www.sde.state.id.us/admin/docs/isat/proficiency-levels-definitions.htm.

**Validity** – Test validity refers to the sufficiency of evidence that supports proper interpretations of a student's test scores. This is a multi-dimensional concept that cannot be determined in a Yes-No fashion.

- **Content validity** refers to whether or not each test question adequately measures specific parts of the Idaho Content Standards.
- **Curriculum validity** is established when the districts and schools adopt and use texts, workbooks, and other materials that are aligned with the state Content Standards.
- **Instructional validity** means that the classroom teachers are actually teaching the state Content Standards.

**Reliability** – A measure is reliable if it can provide consistent results over time. A measure can be reliable but not valid.

*Example: Persons trained to read and score student essays may produce reliable (i.e., consistent) scores, but, if they were inadequately trained, they may be applying the wrong scoring rubrics, thus producing invalid scores.*

**Alignment** – refers to the relative match in depth and breadth of content between the test questions and the Content Standards.

**Mapping** – A process used to determine the achievement level of each ISAT question. Mapping assists in determining if questions measuring a particular Content Standard are equally represented across the achievement levels.

*Example: Mapping measures whether there are questions about "measurement" in the basic, below basic, proficient and advanced categories and if they are evenly distributed.*

## ISAT Reliability

The Question: Are the ISAT assessment tests reliable?

The Importance of the Matter: Statewide assessment tests must be reliable, producing information that is dependable. NCLB language requires documentation of test reliability, as do professional standards in the measurement field. If test scores are unreliable, valid interpretations about student achievements are unattainable.

Methodology of Research: The ISAT contractor, NWEA, routinely collects information about test reliability. This information was reviewed but, in addition, HumRRO conducted its own calculations using a variety of methods. Reliability coefficients were prepared as well as calculations of the standard error of measurement, classification accuracy, and classification consistency. The latter two were needed because the test results are used to classify students into proficiency levels. Such classifications should be precise and replicable.

Calculations were made on the core tests, the blended tests, and on the subscore-reporting units called "Reporting Goals."

Synopsis of Findings: Reliability is determined on a scale from 0 to 1. A measure is considered more reliable the closer it gets to 1. The HumRRO results are similar to those obtained by NWEA. The overall test reliabilities were all above 0.80 and most were around 0.86. The total test standard error of measurement was about 3 points. For the subtests (e.g., Number Sense or Literal Comprehension) where there are fewer test items, the reliability decreases substantially to about 0.50 and the standard errors of measurement increase to about 6-7 points. In other words, the total test score is the most stable and the subtest scores are less stable. These values are typical for statewide assessment tests.

Standard errors of measurement also were calculated using Rasch Item Response Theory methods wherein a standard error of measurement is calculated for each test item rather than for the overall test. The NWEA and HumRRO results are almost identical.

Accuracy addresses the matter of whether the test score accurately classifies a student, just as one might look at a thermometer and consider whether or not the indicated temperature is "accurate." The accuracy indices were in the 0.75 – 0.85 range, with the blended test yielding higher values.

Consistency describes the likelihood that the student would have attained the same proficiency classification on a second administration of a parallel form of the test. The results showed that the consistency indices were in the range of 0.65 – 0.79 with the blended test yielding higher values.

For both accuracy and consistency, if the analyses consider only the Proficiency/Not Proficient dichotomy, the accuracy and consistency values increase to the 0.88 –0.95 range.

Implications for Future Direction: The values obtained in these analyses are typical. The easiest way to increase reliability is to increase the number of items on the tests and subtests. For example, increasing subtests from 6 items to 10 items would increase the reliability of the subscore. Similarly, increasing the overall length of the tests would increase overall reliability.

Reliability should be monitored with each administration of the ISAT and adjustments made in the test structure when necessary.

In its score interpretation guides, the State should reinforce the psychometric principle that the most reliable scores at the individual student level are those derived from the total test, and the subgoal reports are only "advisory." On the other hand, when aggregating data across students into school and district units, the subgoal reports can be considered to be more reliable.

Report: Idaho Standards Achievement Test: Independent Calculations of Reliability Estimates, Standard Errors of Measurement, Classification Accuracy, and Classification Consistency

## ISAT Content Validity

The Question: Do the ISAT assessment tests have content validity?

The Importance of the Matter: Statewide assessment tests must be valid and measure what they are intended to measure. Content validity is not an either-or determination but is, instead, a matter of degree and sufficiency of documentation. Content validity is directly tied to the purposes of the tests.

Methodology of Research: This study had three components: (1) review existing content validity documentation; (2) collect additional documentation as needed; and (3) review mapping of items to Content Standards. Documents were obtained primarily from NWEA but also from the Department of Education and the State Board of Education. All reviews were conducted by HumRRO staff.

In addition, an original item mapping classification study was conducted to determine how well the on-grade level items plus the adaptive items represented the continuum of the four achievement levels (advanced, proficient, basic and below basic) and the distribution of items across the Reporting Goals.

Synopsis of Findings: The available NWEA-provided documentation supports the content validity of ISAT; however, there are gaps in the documentation and inconsistencies were identified. The test blueprint is not in exact concordance with the adopted Content Standards, and the weighting of the items across Standards is unclear. Other issues were identified: there are no item specifications to guide the work of item writers, and it was not clear how many items were attributable to Idaho writers; the content discussions during test item writing sessions were not adequately documented; field testing of new items can occur outside of Idaho and a few items on the test had poor statistical properties.

NWEA initially created the ISAT from the assessment structure that was available at the time. The end result is that the test results are displayed in terms of "Reporting Goals" which are not necessarily synonymous with the Idaho Content Standards. This confuses interpretation of test results.

The mapping study revealed that the items are performing reasonably and measuring students across the four levels of achievement and across the "Reporting Goals." The on-grade level and adaptive portions of the test are functioning as intended by NWEA.

Specifications for quality assurance were not available.

Implications for Future Direction: Content validity can be strengthened through better alignment of the test blueprint and the Content Standards, creation of specific item specifications, and better documentation of future item writing, reviewing, and selection processes. Proper field testing of new items and the consistent use of statistical analyses in evaluating item and test quality should

be required. Items should always be properly distributed across the achievement levels and the Content Standards.

These changes can and should be implemented immediately and reflected in all future item development operations. Quality assurance steps should be specified for all steps in the test development, scoring, and reporting stages.

Additional resources will be needed to address these concerns.

Report: Idaho Standards Achievement Test: Review of Northwest Evaluation Association Content Validity Documentation, Task 3.0.1

Report: Idaho Standards Achievement Test: Review of Evidence of Content Validity, Task 3.0.2

## ISAT Curriculum and Instructional Validity

The Question: Do the ISAT assessment tests have curriculum and instructional validity?

The Importance of the Matter: Idaho law as well as the "No Child Left Behind" Act of 2001 require states to adopt challenging content expectations and then determine the degree to which the expectations are being met. To achieve this, the school curriculum must match the content expectations, and teachers are expected to teach the desired content. Maintaining the alignment of the curriculum and the instructional programs will provide students with the greatest opportunity to learn. In previous litigation about high school graduation tests, the courts have ruled that curriculum and instructional validity issues must be adequately addressed.

Methodology of Research: This study included three different components. (1) Evidence submitted by NWEA was reviewed. (2) Idaho school districts conducted a self-study. (3) Six districts were visited for an on-site review.

Synopsis of Findings: NWEA information did not directly address issues related to curriculum and instructional validity, as these are ordinarily not the responsibility of the test support contractor. Instead, they are the responsibility of local school districts, the State Department of Education, and the State Board of Education.

All school district superintendents were asked to respond to a 30-item written survey soliciting information about curriculum and instructional activities supporting the Idaho content expectations. Ninety-seven percent of the superintendents responded to the survey, but thirteen percent did not prepare an executive summary as requested. The quality of the responses varied.

Superintendents reported that students and parents had been informed about the assessment program, including the high school graduation requirements. Districts apparently have aligned their curriculum with the content expectations, but it is not clear that individual schools are being monitored in this regard. Superintendents asserted that their teachers were teaching the content expectations and that student remediation opportunities were being provided. The superintendents indicated that the ISAT data were being used.

On-site visits were conducted in 10 schools across six districts. Results were varied, with the larger districts seemingly better prepared to assist schools align the curriculum and instruction to the Idaho Content Standards. Smaller districts lacked the human resources to provide a sufficient level of services. Some districts depend upon the state-adopted texts to address the content expectations whereas larger districts have tailored their own curriculum guides. Some educators were concerned that the Idaho Content Standards were not always being taught. According to the report, more professional development is needed. Remedial activities varied, with more rural locations having the most difficulty providing adequate services.

Implications for Future Direction:

Although based on self-reports, it is reassuring that the survey of superintendents and the on-site visits revealed an awareness of the responsibilities educators have with regard to curriculum and instructional validity. However, there are inconsistencies across the State in the degree to which schools can be said to be effectively providing students with the opportunity to learn the Idaho Content Standards. This is an on-going task because new students are enrolled throughout the year, and they (and their parents) must be provided with information about content expectations. Likewise, new teachers and principals are hired each year, and they must learn how to emphasize providing students the opportunity to learn.

The State should seek ways to continuously inform the various publics about the educational expectations and the manner in which the ISAT can facilitate good instruction. Efforts should be undertaken periodically to verify that schools and districts understand the importance of curriculum and instructional validity and that they are actively strengthening students' opportunity to learn.

Report: Idaho Standards Achievement Test: Review of Northwest Evaluation Association Curriculum Validity and Instructional Validity Documentation, Task 3.0.4.1

Report: Idaho Standards Achievement Test: District Self-Study (DSS) Findings Regarding Curricular and Instructional Validity, Task 3.0.4.2

Report: Idaho District Self-Study (DSS): Executive Summaries, Task 3.0.4.2

Report: Idaho Standards Achievement Test: Curriculum Validity and Instructional Validity Study Results from On-Site Visits to Six Idaho Public School Districts, Task 3.0.4.2

## ISAT Alignment with Idaho Content Standards

The Question: Are the ISAT assessment tests aligned with Idaho Content Standards?

The Importance of the Matter: The "No Child Left Behind" Act of 2001 as well as standard psychometric practice demands that an assessment test measure that which it is intended to measure.

Good alignment simply means that all of the test items match the state Content Standards well. A good match means that the test assesses the depth and breadth of the Content Standards. In other words, the questions have a range of difficulty levels and the test items measure a good range of specific knowledge and skills within the standards.

Methodology of Research: Some information about alignment was obtained from NWEA and has been reported in the discussion about content validity. Because of the importance of the matter, HumRRO was required to conduct an independent analysis of test alignment. HumRRO utilized the Webb alignment methodology and retained the services of Idaho educators to conduct the analyses. Other alignment methods exist, and they may have produced somewhat different results. However, these methods assume that the test is derived from the blueprint and the blueprint is derived from the standards. Since there is an issue with this relationship between the ISAT and the standards, it is highly likely that these other methods would have similar findings.

The alignment process used to analyze the ISAT was based on four measurements: categorical concurrence, depth of knowledge, range of knowledge, and balance of representation.

*Categorical Concurrence* is a basic measure of alignment between the Content Standards and test questions. Essentially, the percent reported in this category illustrates how many of the Content Standards were assessed by at least six test questions each.

*Depth of Knowledge* (DOK) measures the amount of cognitive complexity required by the test questions and by the standards separately. For example, if a part of the standard specifies that student should be able to simply recall basic knowledge, then the corresponding test item should be written to assess basic recall only. There are four levels of complexity ranging from recall to complex reasoning and thinking. Each specific content objective (or substandard) within a standard is assigned a DOK level and each test item is assigned a DOK level separately. Then, once the test item is matched to the particular standard, it is evaluated to ensure that the test item DOK level is either at or above the specific content objective. The methodology used (Webb, 1999) suggested that at least 50% of the items be at or above the matched knowledge, skill or ability's DOK.

*Range of Knowledge* can also be referred to as the "breadth" of how many of the objectives within the standards are covered by at least one test question. Another way of referring to range is how completely the test items cover the content in the Content Standards.

*Balance of Representation* takes the range of knowledge test one step further. This test analyzes the content objectives. For those objectives that are assessed, an approximately equal number of questions should be included in the test. The balance measure indicates whether the test questions really are distributed evenly across those objectives.

Synopsis of Findings:
As noted above, alignment using Webb's methodology is determined by four different measurements and therefore a single statement of "aligned" or "not aligned" is not possible. In other words, alignment is a matter of degree and does not provide a single yes or no answer. A test could be well aligned in some areas but not in other areas. The report highlighted several areas where the ISAT process can be improved and should be revised. For example, the reading blueprint has a category for vocabulary but there is not a specific Content Standard for vocabulary.

The study revealed that there are alignment issues that primarily stem from the lack of agreement between the Idaho Content Standards, the items obtained from NWEA, the "Reporting Goals" created by NWEA, and the test blueprints. These matters were reported in the section on "content validity."

The Idaho Math Content Standards have a large list of standards for this one subject area, which makes it difficult to fully assess the range of knowledge within those standards. The Language Arts/Communication Content Standards include five standards: reading, writing, speaking, viewing and listening. Thus, the ISAT for reading is built on just one standard, reading. The language usage test is built on just the writing standard. This small number of standards for a particular test means that it is more likely to have six items per standard (the categorical concurrence test), but that the balance of both the reading and the language usage test is not sufficient. If the reading and writing standards are viewed at the next level of granularity, it becomes evident that the tests are so focused on one or two of the sub-standards that the test does not do an adequate job of covering all the aspects in the standards.

In addition, the educators who participated in the study expressed concern about the wording of some of the Idaho Content Standards and that there were questions included in the test that were not directly related to a particular standard. For example, there were questions about the vocabulary word test items.

Implications for Future Direction: The alignment study confirmed weaknesses seen in the separate analysis of content validity. The problems stem from the use of the existing NWEA test structure, or blueprint, rather than the creation of an assessment made to specifically measure the Idaho Content Standards.

It is recommended that several actions be taken.

1. The Idaho Content Standards should be reviewed and possibly reorganized to solve the problems and inconsistencies that have been revealed. The review should seek balance across the content areas and creation of standards and skills that allow measurement of students across a wide range of proficiency.

2. The test blueprint should be reviewed and revised to reflect measurement of the reorganized Content Standards. If there is to be any content that is not routinely measured by the ISAT, this should be made explicit.

3. Test item specifications should be prepared to define how the Content Standards will be measured across the grade levels to address the lack of breadth in the ISAT.

4. Available items should be classified according to the content standard and grade level skill that they measure. Gaps in coverage should be noted and new items developed.

5. The use of "Reporting Goals" should cease and all reporting should be done in terms of the Idaho Content Standards.

6. The adaptive test structure and the use of "blended tests" should be reviewed in terms of the results of the HumRRO studies and the anticipated changes in items 1-5 above. Decisions must be made about the assessment structure and its capability to provide information needed by the State, by NCLB, and by classroom teachers.

Report: Idaho Standards Achievement Test (ISAT): Test Alignment Study, Task 3.1

## ISAT Item Mapping

The Question: Are the ISAT test items appropriately distributed over the content domains and range of difficulty?

The Importance of the Matter: The purpose of ISAT is to measure the content expectations adopted by the State Board of Education. Therefore, the test must be designed to report student knowledge and skills within various subscore categories representing the important groupings within the expectations. Moreover, the difficulty of the items must be appropriate to measure the full range of students who naturally vary considerably in knowledge and skill.

Methodology of Research:

Item mapping is accomplished by inspecting each test question and matching it to Content Standards. The ISAT, however, is organized by NWEA Reporting Goals that are not necessarily consistent with the Idaho content expectations, thus complicating the analysis.

The item mapping study also matched each test item to the four achievement levels used to report students' skills (e.g., "Proficient"). The goal is to have a sufficient number of items distributed over each of the levels so student performance can be accurately classified.

The results of the mapping study are displayed as bar graphs showing the reporting categories each item measures and the achievement level at which the item measures student performance.

Because the ISAT consists of "on-grade level items" and "adaptive items," analyses were completed for both.

Synopsis of Findings:

The study revealed that the test items are not always uniformly distributed across the four achievement levels (advanced, proficient, basic and below basic). Some of the tests included more items measuring in the lower two achievement levels while others concentrated items within the highest two levels.

In some cases, there were insufficient numbers of items measuring student performance at the "Advanced" level. The adaptive portion of the tests appear to be properly selecting items for students based on the estimation of their proficiency level as determined by the core items.

There appear to be some inconsistencies in the mapping of items to the Content Standards, as opposed to the NWEA "Reporting Goals." For example, a fourth grade goal of "reading technical information" was not measured at all. Two other goals at grade ten were not measured. There is no guiding principle for ISAT test blueprints that specifies the number of items per Idaho content standard that will guide formation of each test.

Implications for Future Direction:

The ISAT item mapping can be improved by two basic changes: (1) Create improved test blueprints and item specifications and (2) Re-examine use of NWEA "Reporting Goals" to better concentrate on the measurement and reporting of Idaho content expectations.

Items should be properly matched to specific Content Standards and spread across the four achievement levels. Because determination of which students are "Proficient" is important for "No Child Left Behind" accountability, it may be preferable to have somewhat more items concentrated around the cut-score for this level for greater accuracy of classification. Steps should be undertaken at the test assembly stages to guarantee that each test item is selected to measure the specified content across the range of student performance.

Report: Idaho Standards Achievement Test: Item Mapping for the 2004 Idaho Standards Achievement Tests

## Action Plan

Based on the findings and recommendations from HumRRO, the Office of the State Board of Education developed an action plan and will present this plan to the State Board of Education at the June meeting.

The plan includes reorganizing the Content Standards in Language Arts/Communications and Math to provide a better outline for how the state assessment should be structured. Once the Content Standards are restructured, the test blueprint will be rewritten. The blueprint will be structured to include reporting areas directly related to the Content Standards. The spring 2006 tests will be built on the new blueprint.

In addition, the Board will discuss how to include more writing components in the language arts test.