Executive Summary

This report summarizes key findings and recommendations from a review of the Idaho Reading Inventory (IRI) and the current practices in Idaho related to the assessment of early reading and the identification of children at risk for reading failure. Drs. Kristi Santi and David Francis of the University of Houston under contract from the Idaho State Department of Education conducted the review.

Evaluation of any assessment begins with the purpose(s) for assessment. Review of the IRI found that schools use the assessment for different purposes and that some of these purposes may conflict with one another. For example, some schools use the results to evaluate teacher performance while also using the test to identify children at-risk for reading problems. Even if the IRI is appropriate for each purpose separately, using the IRI for both of these purposes simultaneously poses a problem. The legislative intent behind the IRI was stated to provide teachers with information relevant to student reading skills and to use the results to assist in the identification of students needing early reading intervention. Using the test to evaluate teacher performance conflicts directly with using the test to identify children at-risk for reading problems because the teacher performance objective discourages identifying children with reading problems. The IRI is also used to classify students with regards to reading proficiency and to support decisions regarding the placement of students into reading intervention. However, the IRI is not well-suited to these purposes due to its format, which focuses exclusively oral reading fluency in grades 2-3, oral reading and letter sound fluency in grade 1, and letter name and sound fluency in Kindergarten. A major source of discontent with the IRI stems from its use in the classification of students into reading proficiency groups given the test's narrow focus.

One question posed to reviewers concerned the adequacy of IRI benchmarks and proficiency indicators. The current proficiency indicators were reviewed and compared to three other nationally recognized norms for similar assessments. Like the IRI, those norms were not based on random samples from the population of students, but rather from convenient samples of students who were already taking the assessments. For the most part, the IRI proficiency indicators are consistent with norms from similar assessments. Some discrepancies were noted, such as higher performance expectations for Grade One Letter Sound Fluency in the Winter and Spring for the IRI (63/72) in comparison to *easyCBM* and *AIMSWeb* (40/46). In the Spring of Kindergarten the Letter Sound Fluency benchmark is lower for the IRI as compared to *easyCBM* and *AIMSWeb*. A bigger concern stems from the IRI's use of only a single set of forms at all assessment periods within each grade. Test development did not examine the extent to which performance gains over time within-grade might be biased due to repeated use of the same forms.

ISDE requested recommendations for alternative benchmark assessments. Several assessments were identified for the IRI Steering committee to review, with the recommendation that each assessment be critically evaluated based on the Steering Committee's decision regarding the purposes for which the IRI will be used in the future. Reviewers were also asked to comment on how the IRI would work within the framework of the Common Core State Standards (CCSS). Because the CCSS does not challenge or alter the foundational skills sequence which underlies early reading development, the IRI would match the expectations set forth in the CCSS and could be used in conjunction with CCSS assessments focused on early reading development.

Reviewers were also asked to examine the scoring of the IRI. Generally, IRI scoring procedures are consistent with those used for other CBM measures being used nationally. However, a significant shortcoming of the IRI stems from the test administration guidelines. Specifically, the IRI does not administer comprehension questions and test administration procedures have not been adapted to instruct students to read for comprehension. The validity of fluency-based assessments is weakened when students are allowed to increase reading speed without regard for the impact on comprehension. Consequently, many fluency-based CBM measures will administer comprehension questions or require that students retell what they have read to emphasize to students the need to read as quickly as possible while still reading for comprehension. Concerns were also raised about the reliance on a single raw score taken as the median score from three one minute readings without regard for differences across the three reading probes in a given grade, and the reliance on a single set of forms at each grade. Reviewers questioned whether the accuracy of the IRI as a screener could be strengthened by using scores from all three CBM probes used at each time point, by equating scores across probes in a single grade, and by developing multiple probe sets per grade. Finally, reviewers felt that interpretation of IRI scores in terms of state standards for reading proficiency was not supported, and that the lack of a standard setting process to derive the performance level descriptors was an area of concern given the desire to classify students into proficiency categories.

The reviewers raised significant concerns regarding the limited psychometric data supporting the IRI. Technical documents submitted to the reviewers contained scant data from the development process. Missing entirely were studies equating test forms or demonstrating comparability of the test across protected subgroups of students. Validity data were quite limited as were data on reliability. No systematic study of the IRI and its use to predict performance on the State of Idaho's standards based assessment was undertaken by the test developer, and few studies have been undertaken by the state or individual districts subsequent to test's adoption. Most of what is known about the performance characteristics of the test comes from studies undertaken by ISDE or by individual districts, especially in the critical domain of validity of the test for identifying children at-risk for poor performance on the ISDE standards based assessment.

Reviewers were also asked to comment on best practices for training proctors and on the validity of using test proctors as test administrators. Reviewers discussed how the training of proctors could be standardized and strengthened. Standardized proctor training would utilize videos to highlight proper administration and scoring techniques as well as to highlight the kinds of problems that occur during routine test administrations. Also, standardized training could require proctors to submit video taped test-administrations in order to be certified to administer the IRI as part of the state's assessment program.

It is recommended that continued use of the IRI as a screening device was possible, but consideration should be given to modifying test administration guidelines to focus students' attention on reading for comprehension. Also, ISDE and local districts should engage in systematic data collection to support investigation into the reliability and validity of test based decisions. ISDE and the IRI Steering Committee must also specifically address the question of the key purpose(s) for which the IRI is to be used and consider supplementing the IRI to addresses those purposes for assessment that are not well-matched to the IRI. Specifically, there is a clear need for broader adoption of standardized diagnostic testing for students who fail the IRI screen in order to inform instructional decisions regarding intervention as well as a need for

progress monitoring assessments for students in Tier 2 and 3 intervention. The need for early reading assessments is not diminished by the adoption of the CCSS. The assessments provided through the CCSS can aid in the screening of students, as well as in guiding instructional decisions and monitoring student progress. However, it is incumbent on ISDE and the IRI Steering Committee to integrate the various components into a coherent assessment system and to monitor the performance characteristics of that system, and continuously adapt it to maintain optimal performance.

Charge to the Review Team

The charge to the University of Houston was to look at the Idaho Reading Inventory (IRI) and the current practice in Idaho related to early reading assessment in order to assist the state in advancing its reading initiative. Presently, there is concern that the initiative may not be achieving its stated goals with maximum efficiency and there is concern that the IRI may not reflect best practice in the assessment industry with regards to early reading assessment, identification of risk, and assessing reading outcomes. Our task is to examine the technical specifications of the IRI and how information is reported and acted upon to ensure that the best practices are being followed in order to give the state the maximum opportunity to achieve its goals in reading achievement for the constituents of ISDE. The ISDE has also fielded a number of concerns from the IRI Steering Committee regarding the IRI and would like to ensure that this review examines and addresses the concerns that have been expressed by the IRI Steering Committee.

The Idaho law, 33-1614. Reading assessment, stipulates:

> "The state department of education shall be responsible for administration of all assessment efforts, train assessment personnel and report results.
>
> (1) In continuing recognition of the critical importance of reading skills, and after an appropriate phase-in time as determined by the state board of education, all public school students in kindergarten and grades one (1), two (2) and three (3) shall have their reading skills assessed. For purposes of this assessment, the state board approved and research-based "Idaho Comprehensive Literacy Plan" shall be the reference document. The kindergarten assessment shall include reading readiness and phonological awareness. Grades one (1), two (2) and three (3) shall test for fluency and accuracy of the student's reading. The assessment shall be by a single statewide test specified by the state board of education, and the state department of education shall ensure that testing shall take place not less than two (2) times per year in the relevant grades. Additional assessments may be administered for students in the lowest twenty-five percent (25%) of reading progress. The state K-3 assessment test results shall be reviewed by school personnel for the purpose of providing necessary interventions to sustain or improve the students' reading skills. Results shall be maintained and compiled by the state department of education and shall be reported annually to the state board, legislature and governor and made available to the public in a consistent manner, by school and by district.
>
> (2) The scores of the tests and interventions recommended and implemented shall be maintained in the permanent record of each student.
>
> (3) The administration of the state K-3 assessments is to be done in the local school districts by individuals chosen by the district other than the regular classroom teacher. All those who administer the assessments shall be trained by the state department of education.
>
> (4) It is legislative intent that curricular materials utilized by school districts for kindergarten through grade three (3) shall align with the "Idaho Comprehensive Literacy Plan." (http://www.legislature.idaho.gov/idstat/Title33/T33CH16.htm).

Summary of Previous Work Presented in Binder

Background Information. A number of documents were provided to the review team documenting the development of the IRI, its current form, and supporting research, as well as supporting materials for teachers, students, and parents. In this section, we describe the documents that have been received and the key information. Among the materials received, the IRI Quick Guide, Parent and Teacher Brochures, Student Report Card, and IRI FAQ will not be discussed further in this report. These documents provide useful information about supporting materials that accompany the assessment, but are not seen as central to the charge other than as background information for the review team.

IRI Targets. The state legislature sets the targets for each grade level. Currently a score of 3 indicates a student is at benchmark and has mastery of the skills. A score of 2 identifies the student in the strategic category meaning that there is partial mastery of some or all of the skills. A student identified as intensive has a score of 1 and lacks mastery of some or all of the skills. In kindergarten, the target is based on Letter Naming Fluency (LNF) in the fall and Letter Sound Fluency (LSF) in the spring. In grade one, LSF is the target skill in the fall whereas Oral Reading Fluency as measured on the Reading Curriculum Based measurement (RCBM) is used in the spring. In both second and third grade, RCBM scores are used for the targets in the fall and spring.

Graphs and Reports. Two sets of data were provided for review. One set of data compared scores in 2000 to scores in 2012 in terms of the percentage of overall students reaching benchmark. The second set of data provides a chronology of performance from 2007 to 2011 of students in each category (intensive, strategic, and benchmark). Individual school data were provided as well. None of the data provided for review compared the number of students in each grade level at each testing point, the demographic characteristics of students and how these might have changed over this period, or indicated changes in the curriculum or state standards during the reported time period. Thus, we must assume a stable demographic, curricular, and standards environment in interpreting these data, assumptions which may or may not hold.

The 2011 and 2012 legislative reports showed growth from 2000 up to the present. The gains shown are based on students meeting benchmark on the spring administration of the IRI. The growth for each grade level was as follows: kindergarten 57 to 82 percent; first grade 52 to 72 percent; second grade 53 to 72 percent; and third grade 49 to 76 percent.

Idaho Reading Initiative FY03. This report, authored by Stacey Joyner, details the background of the IRI. The test was initially broken down by two categories – Kindergarten, which tested reading readiness and phonological awareness, and Grades One through Three, which tested reading fluency and comprehension. A summary of the 2002 study of the IRI conducted by Dr. Frank Gallant, was included in this report. Appendix C of the report is the complete analysis by Dr. Gallant conducted in 2002-2003.

Extended Reading Intervention Analysis 2005-2006. Dr. Frank Gallant provided a second report with the same primary question from the 2002-2003 report. This report attempted to identify effective instructional materials and strategies in the remediation of students identified as intensive by the IRI. There is incomplete information under the heading Instructional Strategies and Materials to determine how each program was implemented and if there was fidelity to the programs. For example, the report states, "some materials are used very little in any of the programs, and conversely some materials are used extensively in all the programs." Missing from this description and the corresponding charts are key information regarding the

amount of time during each instructional day that the materials were used, whether all components of the materials were used, and any indication of the fidelity of use. We also do not know what else was occurring during the instructional time to know if other materials could explain the increase in test scores. The same questions emerge when evaluating the information available for the instructional strategies. As for school effectiveness, the levels of missing and unmatched data are a concern for properly evaluating the results. However, these concerns are minor in comparison to the failure to address possible maturation and regression effects, as in the prior report. The confounding effects of regression to the mean and maturation effects in designs of this type are well known and must be addressed to obtain an unambiguous interpretation of any changes that occur between the pre-test and post-test.

NWREL Technical Report 2007. In June of 2007 Drs. Gary Nave and Art Burke completed a technical report of the IRI under the direction of the Idaho State Department of Education while the state transitioned from the state-developed IRI to the IRI developed by *AIMSWeb*. In general the report finds that achievement is documented differently between the two types of assessments. However, the report clearly states that the study did not evaluate the relative validity of the two measures. Other questions about the analysis remain unanswered. Two in particular are how the cut points for the new *AIMSWeb* test were chosen for both the subtests and the totals and how the subtests for each grade level were chosen. The report concludes with four specific recommendations. No evidence for a follow up to the recommendations can be found in the current documents on hand.

This report from NWREL is the most-current, primary technical report on the validity and operating characteristics of the current IRI. As such, it represents the primary source of technical information on the test from which to answer questions about best practice with respect to the design, build, and performance of the IRI. Consequently, we will review this report in considerable detail below.

IRI Alternate Assessment. Schools have three options for administering the IRI assessment to students with disabilities. Specifically, they may administer the IRI with accommodations, the Student Based Assessment Measure for nonverbal students, or the CORE Phonics Screener. In order for the alternate assessment plan to be enacted for a student, the IEP team or a designee must establish eligibility. This plan appears on the surface to adhere to federal regulations regarding the testing accommodations for students with an IEP. However, to be fully compliant, the performance standards on the regular and alternative assessment must be linked. Also, the use of accommodations should be supported by research on those accommodations for the assessments in question. We are unable to assess if any work has been carried out to link the performance standards across the IRI and the IRI Alternate Assessment. We are also unable to confirm that the choice of accommodations is based on specific research on the use of accommodations with the IRI, or is based on published research with similar assessments for similar purposes.

Summary of Previous Work Presented in Electronic Folder Review

In addition to the printed binder of materials, the reviewers were given access to a variety of materials electronically through a shared folder. The electronic folder contained many additional documents that provided background information documenting the time and effort put forth by the State of Idaho's Department of Education to improve the early reading indicators. In keeping with the scope of work, only the technical reports are reviewed and summarized below.

AIMSweb – New Tech Report IRI. This report provided a summary of the research validating the new IRI developed by *AIMSWeb*. The report is not dated, but the document properties indicate this is a 2004 creation. The information presented in the report indicates that the authors obtained a copy of student tests for a sample of students from selected districts (489 cases: 157 K, 183 G1, 83 G2, and 75 G3) and upon receipt of the tests, the authors checked the scoring for accuracy. The new Reading-CBM was matched to the *AIMSweb* Reading-CBM but the authors do not provide details as to the definition or procedure for matching probes. Results are reported using the correlation coefficient, mean, standard deviations, and number of students tested.

Review of the current format of the IRI

The present IRI covers four grade levels and is administered twice a year. The directions presented in 2010-2011 Proctor Manual state directions for delivery that are common with the majority of fluency based assessments with the exception of articulation and dialect. In the directions for imperfect pronunciation students are given credit when they pronounce the /s/ as /th/. This alteration would result, for example, in pronunciation of the word 'see' as 'thee.'

The Oral Reading Fluency Probes are all narrative passages as opposed to containing a mix of expository and narrative text. This decision reflects a narrow definition of the construct being measured, and is not generally recommended. All probes were evaluated using popular readability calculators to determine the grade level for each passage. These indicators, while not as stable, precise, or psychometrically sound as the Lexile, should show similarity across passages within grade level. Caution should be used in the interpretation of readability calculations due to the varying nature of how each calculator determines readability level. The New Dale-Chall is more useful when scoring text for grade five and higher but is included in the report due to the familiarity most educators have with the name of the formula.

Accommodations. All approved accommodations are appropriate for the intended purpose stated. There are some accommodations, while not detrimental, are not supported by research. For example, under the heading Approved Accommodation for Testing Materials bullet three states, "To further enhance vision, colored overlays, special lighting, and filters may be used." Unless the student uses it as part of the daily routine in school, it should not be introduced for testing situations.

90% Rule. This rule is specified in statute and stipulates which students are included in the computation of the percentage of students at grade level. The rule excludes from the calculation any student who has not been present at the school for 90% of the academic days in that school year.

Kindergarten. In Kindergarten Fall, students take a Letter Naming Fluency Benchmark Assessment which comprises 10 lines containing 51 easily distinguishable capital letters, 43 easily distinguishable lower case letters, and 6 letters that can either be upper case "I" (i) or lower case 'l' (L). The second benchmark is Letter Sound Fluency and contains 10 rows of 10 lower case letters. In the Kindergarten Spring, students take the Letter Sound Fluency first (different ordering of lower case letters) and then proceed to the Letter Naming Fluency (different ordering of the mixed-case letters).

First Grade. In the fall, students take the Letter Sound Fluency with 100 lower case letters. Students then take the Oral Reading Fluency measure following standard protocol for CBM. They are not asked to respond to any comprehension questions after the reading.

In the Spring the students are administered another Letter Sound Fluency task with 100 lower case letters in a different order from the Fall administration and then read the same three passages that were administered during the Fall administration. The information regarding the readability information for each passage is listed in Table 1.

Table 1
Passages: Grade 1

Story	**Word Count**	**Flesch-Kincaid**	**Dale-Chall**	**SMOG**	**Spache**	**Fry**	**Lexile**
Shade Tree	248	G1 (.06)	G1 (1.15)	G2 (1.8)	G2 (1.8)	Early G1	260L
Sea Shell Castle	254	G1	G1 (1.15)	G2	G3 (2.6)	G1	300L
Warm Milk	243	K (.03)	G1 (1.17)	G2 (2.1)	G2 (2.5)	Early G1	240L

*Index is in parenthesis

Second and Third Grade. Both Fall and Spring assessments are Oral Reading Fluency Passages. In second grade the students read the same three passages at both time points. The titles and readability information are listed in Tables 2 and 3 respectively.

Table 2
Passages: Grade 2

Story	**Word Count**	**Flesch-Kincaid**	**Dale-Chall**	**SMOG**	**Spache**	**Fry**	**Lexile**
Thick Fog	258	G2 (2.2)	G1 (1.12)	G2 (2.1)	G2 (2.2)	Early G3	500L
Bunny Hop	250	G2 (1.6)	G5/6 (1.39)	G3 (2.6)	G3 (3.1)	G2	430L
Purple Rat	253	G1 (1)	G4/5 (.99)	G2 (2.1)	G2 (2.3)	G1	390L

*Index is in parenthesis

Table 3
Passages: Grade 3

Story	**Word Count**	**Flesch-Kincaid**	**Dale-Chall**	**SMOG**	**Spache**	**Fry**	**Lexile**
Running for Mayor	271	G3 (3.4)	G5/6 (1.68)	G4 (3.7)	G4 (3.8)	G4.5 border	620L
Playing Rough	342	G3 (2.8)	G5/6 (1.53)	G4 (4.1)	G4 (4.1)	G3/4 border	630L
Squirrel's Machine	314	G2 (2.9)	G4/5 (1.28)	G3 (3)	G4 (3.6)	G4	630L

*Index is in parenthesis

Data. Growth charts were provided for various years based on grade level. While we can view the grade level ability at each year, the data is not linked to specific students, programs of instruction, or intervention. This hinders the ability to ascertain why there is a movement in numbers and what we can learn from the movement in numbers. For example, the Kindergarten jump from 2010 to 2011 is 12% points. Is there one district that is outperforming the other districts and if so what is the source of that district's improvement? Is the amount of change within the range of normal expectation given the random variation of change from district to district from one year to the next? Is that gain typical for that district, such that they consistently achieve higher than expect change year over year? Is the change regression to the mean because of an abnormally low prior year's performance? Have the demographics changed in the district? Is that district doing something differently with the instruction that might account for the change and that would help other districts who had little to no growth?

General Critique

Introduction. The information provided to the review team was well organized and easy to review. The binders contained a variety of reports on the policy relating to the reading initiative and the IRI as a tool for identifying at risk students and monitoring reading development in K-3 on an at least twice annual basis. The binder included information on the format of the assessment, including a description of subtests and forms at each grade, and several reports from evaluations of the assessment and the reading initiative. The binder also included copies of IRI reports that are made available to schools along with information on school performance on the IRI for several years. Review of this material provided strong background on the history of the IRI and its development as well as concerns that have arisen since the adoption of the current IRI based on the rCBM framework and developed by *AIMSWeb*.

In reviewing the binders, the technical information related to the assessment development and the psychometric properties of the IRI was limited. Missing was information on forms development, text difficulty, scaling of performance across different forms, information on possible confounding of practice effects with growth in performance because of the use of the same form on multiple occasions, validity information on cut-scores, including ROC analysis of diagnostic accuracy or predictive validity of fall benchmarks to outcomes of interest, such as standardized tests of reading achievement or the ISAT. There was one report linking IRI to ITBS, but this report included limited validity information and did not include an analysis of cut-scores at any grade. A second report produced by Dr. Stewart examined the relation between the IRI and Grade 3 ISAT that demonstrated some of the kinds of analyses that should have occurred as part of the test development process. No such reports appear to have been produced by the test developer, or at least no such reports were provided to the state as part of the technical material on the assessment. While development of such information is possible once the test is in operational use, it is typically part of the test development process. Gathering this information during test development is necessary to inform the setting of cut-scores. Also, the information gathered once the test becomes operational is limited because decisions are being made based on the operational test data and these decisions can affect the relationships being studied. For example, once schools begin making decisions to intervene with students based on test performance the relationship between the screening test and student outcomes may be altered. Specifically, if the interventions are effective, then the relationship between screening performance and student outcomes will be altered so as to increase the false positive rate for

screening test decisions. Thus, it is important to estimate these relationships prior to the test becoming operational.

At the same time, there is considerable value in the analysis of operational test data. During the conference calls and site visit, there was discussion regarding the possible existence of other reports, perhaps from the larger districts, relating IRI and other potential reading related outcomes, such as ITBS or SAT 10 at the grades where ISAT is not measured. We also discussed whether analyses had been done relating school level performance improvements on IRI to school level performance changes on ISAT over multiple years. No such reports were available. We learned that there is limited data available at the state level that relates the performance on IRI to performance on ISAT or another standardized reading assessment at the student level on either a cross-sectional or longitudinal basis, and that longitudinal student-level analyses have been hindered by the development of a state-level database for longitudinal tracking of student data. Perhaps because of this limitation on longitudinal data at the student-level, attempts to analyze performance at the school and district level have also been limited. Although reports of performance gains at the school and district level have been produced, there has not been a systematic modeling of school level performance over time relating IRI and ISAT in a manner that might inform the state about the value of the IRI in improving student performance on the ISAT. Multi-level analyses at the school and district level are possible without longitudinal student-level data provided that there is systematic storage of performance data at the school-level.

While there is no doubt that the state would benefit from a longitudinal student-level data system, it also appears that efforts to collect, store, and analyze school-level and district-level information could be strengthened to facilitate state-level decision making related to its student assessment program. This conclusion is based in part on the limitations of the reports that were provided to the reviewers that are based on operational data. For example, interpretation of the changes reported in the 2011 and 2012 reports to the legislature is complicated by several factors that are relevant to the charge. First and foremost, the IRI changed in 2007 so the reported "growth" is based on a comparison of data from two different tests. Thus, the change observed is a function of at least three things: (1) changes in reading ability across the population of Idaho students, (2) possible changes in the difficulty of the test, and (3) changes in the benchmarks tied to different tests. Further complicating this comparison are possible changes in the demographic composition of the student body over this time frame, which could have artificially altered the distribution of reading achievement. To the extent that change is not uniform across all demographic subgroups, the change observed in the total population is a weighted average of the changes within different subgroups of the population. If the percentage of the population that falls into different demographic subgroups has also changed over this time frame, then the observed change in reading performance in the student body as a whole can significantly misrepresent the underlying changes in performance that have occurred within subgroups.

Another report provided to the reviewers based on operational data was the report of Dr. Gallant on the changes experienced by students who had been targeted for intervention. We discuss this report here in some detail because of its relevance to the primary role of the IRI as a screening test. This report endeavored to determine "whether the differences in the pre- and post-remediation average scores are significantly different from zero (p.26)." The report also discussed remediation however, there were no data reported on the specific program or programs that were used for remediation, how long the remediation lasted, or who delivered the remediation. As reported, the only inference one can derive from the report is that growth

occurred during the school year for children receiving intervention. The report failed to address the important question of treatment impact, that is, did the interventions work. This question can only be answered by comparing the change observed for students with and without treatment, adjusted for any selection bias that renders the students receiving intervention different from those not receiving intervention. The analyses presented in the report do not directly address (1) if the IRI is identifying the right students for intervention, or (2) if the interventions are having a positive impact on student performance.

Is the IRI identifying the right students for intervention? All tests that are used for the purpose of screening/identification of students for intervention will identify students for intervention who do not require it and will fail to identify students for intervention who, in fact, require it. These test-based decision errors are called false positives and false negatives, respectively, and are a consequence of any test-based decision rule. They are unavoidable and, most importantly, they are inversely related. That is to say, reducing the risk of one kind of decision error increases the risk of the other kind of decision error. The two kinds of decision error are a direct consequence of the imperfect relationship between the test used to make the decision and the outcome that the test is used to predict. Moreover, for any given test, once the decision rule is set, the error rates are fixed and can only be altered by changing the decision rule. Specifically, we can lower the risk of false positive errors by setting a more stringent criterion for deciding who needs intervention. However, doing so would invariably result in missing more students who need intervention. If we instead lower the risk of missing students who need intervention, we will invariably increase the risk of including students in intervention who do not need it. An important aspect of evaluating a test that is used for such a purpose is to measure the false positive and false negative rate and to develop the test's receiver operating characteristic curve (ROC curve), which shows how the test's performance changes as the criterion for identification is changed across the score distribution on the test. No such analysis has been conducted with the IRI as reflected in the materials provided to the reviewers.

Is the targeting of students for intervention based on the IRI resulting in improved student outcomes on the ISAT or other measures of reading achievement in those grades where ISAT is not measured? This question underscores the second important dimension relating to the primary purpose of the IRI that was not addressed in this report, namely the evaluation of intervention impact resulting from a positive IRI screen. Because students will change over the course of the academic year for a variety of reasons, attributing the change to the instruction/intervention that students have received requires a measure of how much students would have changed in the absence of that instruction/intervention. There are many ways in which to estimate the amount of change that would have occurred in the absence of intervention. Virtually all of these involve some type of comparison to students not receiving the intervention and some attempt to account for any selection bias that might be operating and affecting the estimate of intervention impact. Regression discontinuity designs (RDD) provide a strong approach for impact evaluations in operational settings such as the one covered in the report, where randomization to treatment and control is not possible because it is unethical to withhold treatment from students.

Other concerns about the report are the potential impact of missing data, the extent to which student scores could be matched from pre-test to post-test, and the possible impact of selection bias on measures of change, specifically the impact of regression to the mean. Students targeted for intervention are selected based on low initial performance. Even in the absence of treatment effects or any true change in performance, students selected for extreme scores at the

pre-test would be expected to have higher post-test scores on average due to regression to the mean. Concerns about missing data and unmatched student scores arise because the report does not address them directly. On page 31 the report states, "Most of these students are included in the analysis. Consequently by comparing percentage of changes from pre-test to post-test provides a good, but cursory, indication of the quality of remediation." Without taking into account possible regression effects and possible maturation effects through comparison to children who performed similarly at the pre-test but did not receive intervention, there is no way to unambiguously attribute observed changes to the effects of intervention. When reviewing data to make decisions about allocation of money spent, one would need to drill down further into the data to determine which remedial programs showed the most growth relative to growth in the absence of treatment for similarly performing children at the pre-test. Otherwise, there are several alternative explanations for the changes that are both highly plausible and impossible to rule out.

In what follows, we will provide a more detailed critique of the IRI based on the information provided regarding the development of the test and its operational use. However, it is important to realize from the outset that the amount of technical information available from the test development process and from the operational use of the test are both quite limited.

Test Development. We described above the steps that were taken to establish the level of difficulty for each form of the IRI. The steps used by the developer reflect standard practice. The bigger concerns there are the number of forms developed, i.e., the development of only a single operational form at each grade for use across all waves and years, is a serious short coming, and does not allow for the differentiation of practice effects from true learning, nor does it provide for any alternative assessments should retesting be necessary, or should security be compromised.

Another limitation of the test development concerns the equating of scores across forms. We should stress that no amount of care in equating passages for difficulty/readability can guarantee equivalence in the raw score distributions of performance. Equivalence in score distributions is accomplished through an equating process carried out once the raw score distributions have been estimated on a suitable representative sample. Examples of common equating processes are linear equating (adjusting raw scores for differences in the means and standard deviations of their distributions) and equipercentile equating (adjusting raw scores in a nonlinear fashion to ensure equivalence across percentiles). Test developers follow a process of creating scale scores so that scores from different subtests and different forms of the same test can be expressed on a common scale. A ubiquitous error in the development of rCBM measures is to assume that because test forms have been equated for readability and all scores are expressed as words read correctly per minute that the scores exist on a common scale. However, the raw fluency scores are not equated from different rCBM probes are not equated simply because they express fluency in words read correctly per minute as a significant amount of research on the psychometric properties of rCBM measures has shown (see Francis, et al., 2006). The practice of treating raw fluency estimates as equivalent persists in the use of rCBM measures because of the desire to keep the process of obtaining fluency estimates simple for teachers and other test users. However, the acceptability of this practice depends on the purpose for assessment. When rCBM assessments are used for informal assessment to guide instructional decisions on a periodic basis, the failure to equate raw scores is likely inconsequential to students and teachers in the long run. However, when rCBM assessments are used to identify students for

intervention, or are reported as components of an accountability framework, making the scores consequential for students, teachers, and schools, then it is critical that rCBM scores be formally equated and the quality of the equating across demographic subgroups must be formally evaluated. For example, the errors in equating of scores from forms A and B should not differ significantly for boys and girls, although it is quite possible that raw score distributions on tests will differ for boys and girls, and other demographic groups reflecting true differences in abilities.

It is safe to say that no such equating process was undertaken in the development of the IRI probes for any of the grades or subtests. This conclusion is based on the omission of any description of equating from the technical report, the lack of raw score conversion tables, and the overriding practice in the field to ignore this problem in the development of rCBM assessments. As mentioned, the significance of this omission in the test development process depends in large measure on the purpose for which IRI scores will be used. In so far as the intended use of the scores has evolved over time toward one of accountability and higher stakes decision making for the students, teachers, and schools, the need for equated forms has increased and the omission of this step in the test development process becomes an increasing concern.

Data provided in the technical report produced by *AIMSWeb* raise several concerns about the test development process given the intended use of the test as a screening instrument. To begin with, the small sample sizes in each grade, and limited representation at the school/district level, no specificity on how the sample was chosen, and no working definition for matching probes, the data in the technical report are not easily interpreted. The report provides virtually no information regarding the sampling of students, schools, and/or districts for the sake of this report. The sample is described as representing the major demographic subgroups and SES levels of Idaho's then current population of students, but it is unclear how the sample was chosen and how the sample size for the study was determined in advance of collecting the data. Criteria for the probes were stated to be mean values that differed by less than 1 standard deviation from the anchor probes and correlations that were not less than .80. However, these criteria are not linked directly to the choice of sample size, nor are the criteria ever justified psychometrically. A difference of one standard deviation is enormous and far beyond the acceptable range for "parallel" tests. Consider that parallel tests are considered substitutable for one another and produce the same score distribution. If tests A and B are normally distributed and the mean of test A exceeds the mean of test B by one standard deviation, then a randomly sampled score from test A has a probability of .707 of exceeding the value of a randomly chosen score from test B. Similarly, parallel tests should correlate 1.0. Because unreliability attenuates correlations, it is reasonable to express this criterion in terms of the correlation between true scores, that is, the correlation disattenuated for unreliability, which equals the correlation divided by the square root of the product of the reliabilities. Thus, two tests with equal reliability of .8 should have an observed correlation of .8 to have a disattenuated correlation of 1.0. Two tests with reliability of .9 should have a correlation of .9 in order for the disattenuated correlation to equal 1.0. In short, the criterion for the correlation must be tied to the test reliabilities to be meaningful. There is no specification in the report regarding the reliability for the new test probes, only the correlations with the anchor test probes.

Once the criteria for parameter estimates have been set, the necessary sample size to achieve this level of precision should be determined. This sample size should take into account factors that affect the standard errors such as the clustering of students within schools and districts, which tends to inflate standard errors. The sample size for the NWREL technical report

does not appear to have been established through a specific design process, and the sample itself appears to be a convenience sample rather than a purposively selected sample. In Kindergarten and Grade 1, the sample size per estimated relationship is approximately 56 students. With standard deviations ranging from 14 to 18 and a sample size of 56, the standard error of the mean is 1.9 to 2.4, indicating that a 95% confidence interval on the mean would have a width of +/- 3.8 to 4.8, or a total of 7.6 to 9.6 score points. These are generous estimates for the standard error and confidence interval widths, which are likely underestimated due to the inability to take into account the effects of clustering of students within teachers, schools, and districts, which tends to inflate standard errors. While the correlations reported in the tables are uniformly high, the reported numbers ignore the variability in the estimates. For example, a correlation of .90 based on a sample of 56 results in a 95% confidence interval that ranges from .83 to .94 using the r to z' approach to finding the confidence interval. Based on the largest sample of 83 students, the 95% confidence interval would range from .848 to .934. While lower bound estimates for the correlations in the tables would likely still be acceptable, the confidence intervals should be reported to convey the precision in the reported estimates. In addition, it should be noted that clustering affects both point estimates and standard errors of correlations. The only way to avoid the effects of clustering would be to employ a simple random sample in the design, but doing so would highly inefficient because this approach would involve the sampling of individual students from across the state. The more efficient sampling process would involve the sampling of districts, schools within districts, and students within schools. Estimation of all means and correlations and their respective standard errors would then take into account the clustered sampling plan.

Completing missing from the technical report is any discussion of the establishment of the cut-scores on each form of the test. Standard test development for screening tests is to engage in an ROC analysis comparing test sensitivity and test specificity at different potential thresholds and to set the operational threshold based on optimizing the "costs" of the two kinds of test errors (i.e., false-positive errors and false-negative errors). We discuss this issue at some length below in the section on technical standards for screening tests.

Test Use. There is concern that the multiple purposes for which the IRI has been used are at odds with one another. Of the various purposes for which the IRI is currently used, its primary purpose was to serve as a screener to identify students in need of intervention. At the same time, some users have expressed a desire that the IRI also serve diagnostic purposes. As a diagnostic tool, the IRI would not only indicate which students were at-risk, but also the basis for that failure and likely interventions. Some of the criticisms of the test expressed at the pre-review meeting and also in the Steering Committee minutes focus on the diagnostic utility of the test, not its functioning as a screener. However, like screening and accountability, screening and diagnosis are very different purposes that can work against one another in developing an assessment that is optimal for either purpose. Screening is intended to be brief, because it is done with all students, whereas diagnostic assessment is done to refine screening decisions and develop treatment plans. Diagnostic assessment is necessarily broader, more time consuming, and more thorough, but is typically carried out with fewer students, thereby justifying the increased length and expense of a thorough diagnostic assessment. Typically, diagnostic assessment serves the dual purpose of refining screening decisions for identified students, helping to reduce or eliminate false positive errors, and providing a guide for targeting intervention.

Originally established in statute as a screener, the statute governing the IRI was later altered to include accountability provisions, reflecting the desire to hold schools and districts, and in some cases teachers, accountable for performance on the IRI. While the desire to use the same information for multiple purposes is understandable given the cost of testing both financially in dollars and educationally in lost instructional time, the success of these endeavors hinges on the intended uses for a test not being at odds with one another. But screening and accountability work at cross-purposes, which is apparent in some of the policies governing the use of IRI. For example, the 90% rule relating to the reporting of IRI data is a curious rule if we think of the IRI as a tool for identifying students at risk of reading failure because the existence of the rule suggests that non-enrollment/absence from school changes the value of the test for determining the child's risk status. While it is understandable that the school might not be the unit of the educational system that is accountable for the student under these circumstances, some other unit of the educational system is accountable for that student (i.e., the district or state) unless the student is new to the state during this academic year and has not been in the state for the full academic year. If the IRI's sole purpose was to screen students to identify risk-status, the 90% rule would be seen as counterproductive to that purpose as the state would want to know the percentage of its students that start the school year at risk of reading failure.

Using an assessment as both a screener and an outcome assessment in an accountability framework is also problematic because the screener lacks the scope and precision demanded of an accountability assessment. The accountability assessment also requires security and equating across forms, such that it is arbitrary to the student which form has been administered. In contrast, the screener need not be equated across forms provided that appropriate decision rules have been established for each form in use, and those decision rules lead to equitable decisions (i.e., comparable accuracy in decision making), again making it arbitrary to the student which form was administered. Security is also not necessary for a screening test and security around the IRI has resulted in unwarranted criticism of the test as a screening instrument. There also is no requirement that the domains of knowledge and skill assessed by the screener be the targets of instruction, or proxies for the outcome per se, only that they relate statistically to the outcome in a manner that allows for the adoption of decision rules that have the desired level of precision for all students who are subject to screening. Technically, a screener could use different decision rules for different subgroups of students, if the relationship between the screener and the criterion varied across subgroups. This heterogeneity in decision rules is allowable because the goal of the screener is to classify students into risk categories and for these decisions to be comparable in their accuracy for all students, even if the decisions are arrived at by different calculations.

Technical Standards as a Screening Test. As a screener, the IRI is not exceptional. The analysis by Dr. Stewart of Third Grade IRI and ISAT performance demonstrates significant problems with the IRI as a screener, and highlights that the screening cut-points do not operate comparably for different student groups. While these results are based on operational test data, and thus may reflect changes in the test's relationship with the ISAT that result from the decisions that are being made based on student performance, there is cause for concern. As Dr. Stewart's analysis points out, and as we have discussed elsewhere in this report, when a screener is used to predict risk on a future criterion, the screener can yield a correct decision or an incorrect decision. Moreover, there are two kinds of correct decision and two kinds of errors. This process can easily be captured in the form of a 2 X 2 contingency table, where the rows of the table represent decisions based on the screening test and the columns of the table represent

performance on the outcome criterion. The table below has been labeled to show whether a decision is correct or incorrect, the type label for that decision, and a symbol to represent the number of cases in that cell (e.g., a, b, c, and d represent the number of true negatives, false negatives, false positives, and true positives, respectively). These symbols will be helpful in computing some probabilities that are useful in evaluating a screening test.

		ISAT Criterion	
		Passed	Failed
IRI Decision	Not at Risk	Correct Decision (True Negative) n = a	Incorrect Decision (False Negative) n = b
	At Risk	Incorrect Decision (False Positive) n = c	Correct Decision (True Positive) n = d

The process of setting the threshold for making risk decisions on a screener is based on a careful analysis of such 2 X 2 contingency tables and how these tables change as the threshold is altered. Before describing that process, we will consider the results reported on by Dr. Stewart in 2009 from Grade 3 IRI-ISAT performance from 2008-2009. Dr. Stewart's report is correct in describing the two kinds of correct decisions and the two kinds of errors, but fails to compute several important statistics used in the evaluation of screening decisions. Specifically, Dr. Stewart reports false positives as a percentage of cases, i.e., he reports c / (a+b+c+d) as the percentage of false positives. While this percentage is the percentage of cases that are false positives, a decision theoretic analysis of this table would examine the false positive rate, which is the percentage of positive test decisions that are false positives, i.e., c / (c + d). Similarly, the false negatives are not reported as a percent of cases, but as a percentage of negative test decisions, i.e., b / (a + b). Two other critical numbers are the sensitivity of the test and the specificity of the test. Like the false positive rate (FPR) and the false negative rate (FNR), these are conditional probabilities. Specifically, they are probabilities of true risk status conditional on test decisions. Sensitivity is the probability of failing the criterion test given a prediction of being at-risk on the screener, i.e., d / (c + d), which is 1 – FPR. This number tells us the probability that the student will fail the criterion test given that the screening test has concluded that the student is at-risk of failure. The specificity of the test is the probability of passing the criterion test given that the screener has identified the student as not at risk, i.e., a / (a + b), and is equal to 1 – FNR. This number tells us the probability that the student will pass the criterion test given that the screening test concludes that the student is not at risk.

It is instructive to construct these 2 X 2 tables for the overall student cohort and each of the subgroups of students reported on by Dr. Stewart and to examine the Sensitivity and Specificity of the test for each of these tables. We first provide all four of the tables and the associated probabilities along with the FPR, FNR, Sensitivity, and Specificity and then discuss the results for these tables and their implications for the IRI.

All Students

	ISAT Performance		
	No Risk	Risk	Total
No Risk	12,595	487	13,082
Risk	2,201	1,717	3,918
Total	14,796	2,204	17,000

False Positive %	12.9%
True Positive %	10.1%
True Negative %	74.1%
False Negative %	2.9%
Specificity	96.3%
Sensitivity	43.8%
FPR	56.2%
FNR	3.7%

Hispanic Students

		ISAT		
		No Risk	Risk	Total
IRI	No Risk	1509	154	1663
	Risk	387	489	876
	Total	1896	643	2539

False Positive %	15.2%
True Positive %	19.3%
True Negative %	59.4%
False Negative %	6.1%
Specificity	90.7%
Sensitivity	55.8%
FPR	44.2%
FNR	9.3%

Title I Students

		ISAT		
		No Risk	Risk	Total
IRI	No Risk	5064	261	5325
	Risk	1307	1050	2357
	Total	6371	1311	7682
False Positive %		17.0%		
True Positive %		13.7%		
True Negative %		65.9%		
False Negative %		3.4%		
Specificity		95.1%		
Sensitivity		44.5%		
FPR		55.5%		
FNR		4.9%		

Special Education Students

		ISAT		
		No Risk	Risk	Total
IRI	No Risk	482	67	549
	Risk	355	715	1070
	Total	837	782	1619
False Positive %		21.9%		
True Positive %		44.2%		
True Negative %		29.8%		
False Negative %		4.1%		
Specificity		87.8%		
Sensitivity		66.8%		
FPR		33.2%		
FNR		12.2%		

What is clear from these tables is that the false positive rate is unacceptably high for most, if not all, subgroups. We can think of the FPR as providing an indication of the percentage of identified students who are not actually at risk. This percentage ranges from a low of 33% for Special Education Students to a high of 56% of students, indicating that over half of the identified students are not actually at-risk. The false negative rate is acceptable over all, but is not acceptable for specific subgroups, particularly for Hispanic students and students in Special Education, where the FNR exceeds 5%.

While these numbers are a concern, there is an even greater concern from the standpoint of best practices in testing. Specifically, tables such as these should have been constructed for each and every subtest and form and for each and every time point at which screening decisions

were to be based. This kind of contingency table analysis is fundamental to the evaluation of screening instruments. Even more fundamental is that an analysis of Sensitivity and Specificity in the form of receiver operating curve (ROC) examination should be used to determine where to place the decision criterion on the screening test. If one can establish costs associated with the different kinds of errors, more complex approaches to setting the cut point on the screener are possible. Based on the analyses reported in the technical reports and supplemental information provided to the reviewers, no such ROC analysis was conducted. It appears that the screening cut-point was set based on some other, unspecified criteria, or were set arbitrarily, which can never be justified. Whether the tests performance could be improved substantially through an adjustment of the cut-point is doubtful for grade 3 given the weakness of the relationship between the IRI and the ISAT as reflected in the correlations, although it is possible that better use of the information from the multiple probes collected per assessment period could improve the performance of the screener. However, adherence to the tradition in rCBM of taking the median score from three probes likely precluded consideration of this possibility for improving the performance of the screener.

The other factor limiting the utility of the current IRI or whatever instrument is substituted for it in the future is the lack of a longitudinal student data system with strong data standards and a requirement that all student assessment data be managed through this system. This limitation prevents the ongoing monitoring of the utility of all aspects of the assessment system, and makes it difficult to conduct the necessary research and evaluation of educational decisions based on test results. Hopefully, significant progress has been made on this front since the current IRI was first put into place, facilitating evaluation of future assessments used in the state.

Response to Specific Questions Posed in the Charge to Reviewers

Recommend best practices for early reading assessments

Developmental, Systems Approach to Reading Assessment: Reading Development. Best practices for early reading assessment demand a systems approach to assessment and recognize the developmental nature of reading, by which it is understood that reading is measured differently at different points along the continuum from beginning literacy to skilled reading. Early reading assessments should also explicitly allow for their systematic and repeated use over the grade span from the beginning or middle of kindergarten through the end of grade 3. There is also a general recognition that early literacy skills unfold in a generally fixed sequence from the ability to do rhyming and engage in simple sound based activities not directly tied to print, to knowledge of letter names and sounds, to the blending of sounds in the forms of onsets and rimes, followed by the blending of phonemes, and on to the segmentation of words and non-words into phonemes and the elision of phonemes. During the emergence of these last skills – from blending of phonemes to the elision of phonemes, students begin reading words, including sight words and words read through the application of letter-sound correspondence rules. As students practice the decoding of words in isolation and in connected text, these skills become more automated and students become more fluent in their reading of both isolated words and connected text. The ability to read words and to apply phonemic decoding rules to the decoding of non-words precedes the ability to read connected text with fluency and understanding, which themselves precede students' ability to read strategically. As students become more fluent and automatic in decoding words and gain practice with the reading of connected text, they also become more automatic in their comprehension of text through improved

efficiency in access to word meanings and the ability to connect information automatically across larger and larger spans of text.

Reading assessments measure these different skills to varying degrees depending on their comprehensiveness and their developmental focus. For example, reading assessments targeting very young readers focus on skills tied to the development of decoding (i.e., letter names and sounds, phonemic awareness, word reading accuracy) and to the assessment of comprehension as measured through the understanding of spoken rather than written language. Assessments attempting to measure reading throughout this developmental process will include a number of subtests that range from assessment of discrete skills (e.g., phonemic awareness, decoding accuracy, vocabulary), and application of those skills in reading comprehension (e.g., measuring the fluent reading of connected text and answering questions requiring text recall and inference making). Other assessments focus only on the measurement of skilled reading as reflected in the assessment of fluency, comprehension, and vocabulary. These assessments are most appropriate for readers toward the end of grade 2 and beyond.

Developmental, Systems Approach to Reading Assessment: Systems Approach. To say that best practice in reading assessment takes a systems approach implies that assessment is always undertaken with a specific set of purposes in mind. No assessment is optimal for all purposes. If an assessment program fails to adopt a systems approach, some components of the program will not be as effective as they could be in so far as assessment decisions will not be optimal (i.e., error rates will not be as low as possible) and the total cost of the program will consequently be greater because of the costs associated with a higher than necessary rate of incorrect assessment decisions. We have delineated above the two kinds of decision errors, but have not discussed the costs associated with them. Obviously, there are costs to the child and the school associated with failing to identify children who are at-risk for poor outcomes. There are also obvious costs associated with delivering intervention to students who are wrongly identified as being at-risk. While these students may not be harmed by participation in an intervention that is not needed, the resources used to deliver these interventions could have been used to deliver intervention to students who needed it.

There is general agreement in the early reading arena that a comprehensive assessment program distinguishes between four purposes of assessment and five domains of assessment. The four purposes are Screening, Diagnosis, Progress Monitoring, and Outcomes Assessment. We have discussed the differences among these purposes previously in this report. The five domains of assessment in early reading are phonemic awareness, phonics, fluency, vocabulary, and comprehension. Depending on the purpose and the developmental stage of reading, one or more of these domains will be the focus of assessment. Specific assessments may target as few as one of these domains, or as many as all five domains, and may address the development of reading from pre-reading skill to skilled reading. Best practice does not dictate that any given assessment target a particular purpose, developmental range, or set of domains, only that the assessment program cover the purposes, domains, and stages of reading.

Current proficiency indicators

Currently the state of Idaho has set performance indicators for the IRI. Based on performance on the IRI, students are designated as a) on benchmark, mastery of skills, b) strategic, partial mastery of some or all skills, or c) intensive, lack of mastery of some or all skills. Depending on the grade, oral reading fluency or IRI subtest scores are used to determine the overall performance on the IRI. According to the documents provided to the review team, the IRI proficiency levels were in alignment with the *AIMSWeb* proficiency levels (Technical Report, 2011). *AIMSWeb* norms are derived from their national database of users. The data was not systematically collected to be a

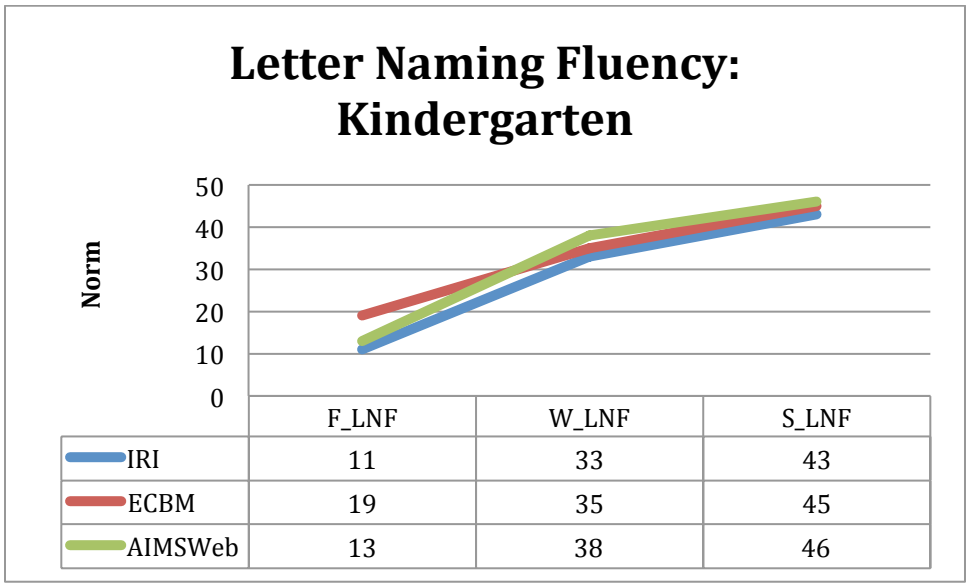
random sample of the population representing differing levels and types of students. The technical manual does provide the descriptive statistics for all data collected via the users input. Given that the original proficiency levels were used in the development of the IRI, the current *AIMSWeb* (2012) cut scores are included in the report. The table below provides information on nationally recognized benchmarks (i.e., norms) for oral reading fluency and shows how those norms relate to the established IRI performance indicators. The first row is the reported performance indicator currently used for the IRI assessment. The second row provides norms from the *easyCBM* (ECBM: 2012), which is a test similar to IRI and DIBELS. At the lower grades *easyCBM* also has norms for Letter Naming Fluency and Letter Sound Fluency measures. The values provided in the table reflect performance at the 50th percentile and are deemed 'average' performance in the *easyCBM* score interpretation manual for 2012-2013. The norms provided by *easyCBM* were determined based on the available data from 1200 students who took *easyCBM* at all three time periods. This is not a random selection of students representing a full range of reading abilities. The final column, AIMSWeb, provides the publishers cut scores developed for the 2012-2013 academic year. The cut scores are based on a success probability whereby students scoring at or above these identified scores were 80% likely to also do well on other standardized measures.

There is one inconsistency on the *easyCBM* data chart that is not addressed in the *easyCBM* manual, namely, the third grade Spring score is reported as two points lower than the winter score. This is not consistent with other norm charts (i.e., DIBELS, TPRI, etc.), and is not consistent with expectations for reading development. In all likelihood, this drop in expected performance reflects a failure on the part of the test developers to equate test forms used at different benchmark periods, but other factors such as missing data, or variation in sampling of students across time could contribute to such an effect. Importantly, the failure to equate forms influences performance even if the norms appear to be consistent with developmental expectations. That is to say, without a consistent scale across time, which can only be achieved through the equating of test forms and the creation of a common scale across forms, changes in oral reading fluency scores over time are difficult to interpret.

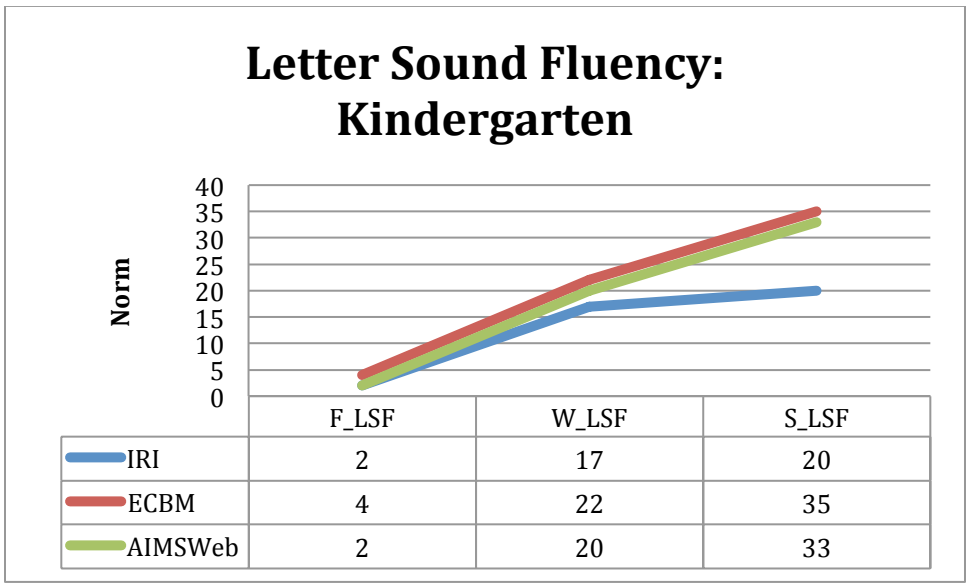
The third row contains the norms from nationally collected data on words correct per minute (wcpm) based on timed readings with the reported number representing a score for students at low risk of reading difficulties (the equivalent to the IRI benchmark). The Hasbrouck and Tindal (HT: 2006) Oral Reading Fluency Norms are commonly referred to when evaluating student fluency rates. The authors report WCPM scores for Fall, Winter, and Spring in percentiles (i.e., 10th, 25th, 50th, 75th, and 90th). Students are expected to be at or above the 50th percentile to be considered within an average reading range (plus/minus 10 WCPM at the 50th percentile) and not at-risk for reading difficulties. As mentioned in this report, caution should be used when interpreting raw scores on fluency measures. While the Hasbrouck and Tindal norms are still commonly referenced, the norms have not been updated since 2006.

In addition to the tabled information, we have plotted the IRI proficiency cut-score and the ECBM*, AIMSWeb,* and HT norms for each grade in separate figures below.

Kindergarten

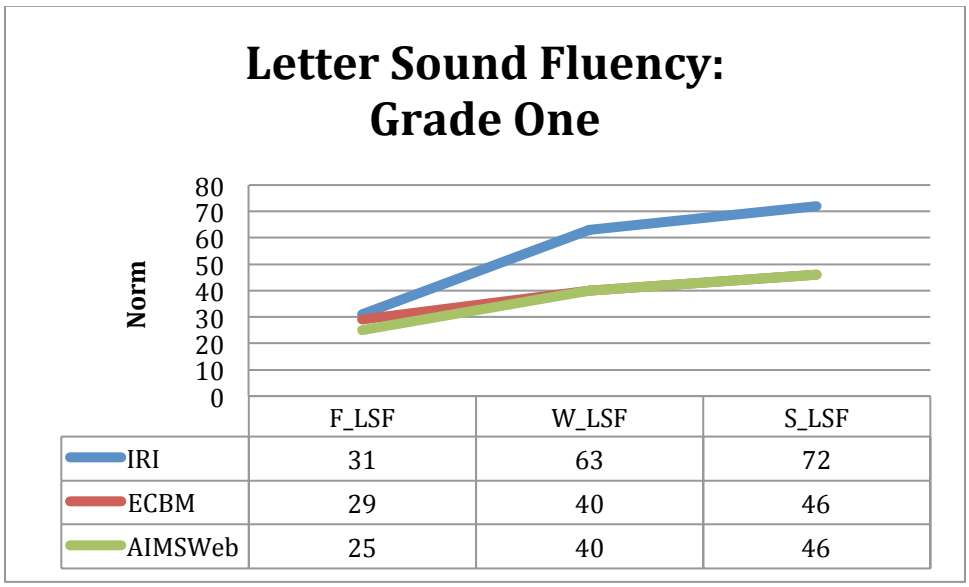


	F_LNF	W_LNF	S_LNF
IRI	11	33	43
ECBM	19	35	45
AIMSWeb	13	38	46

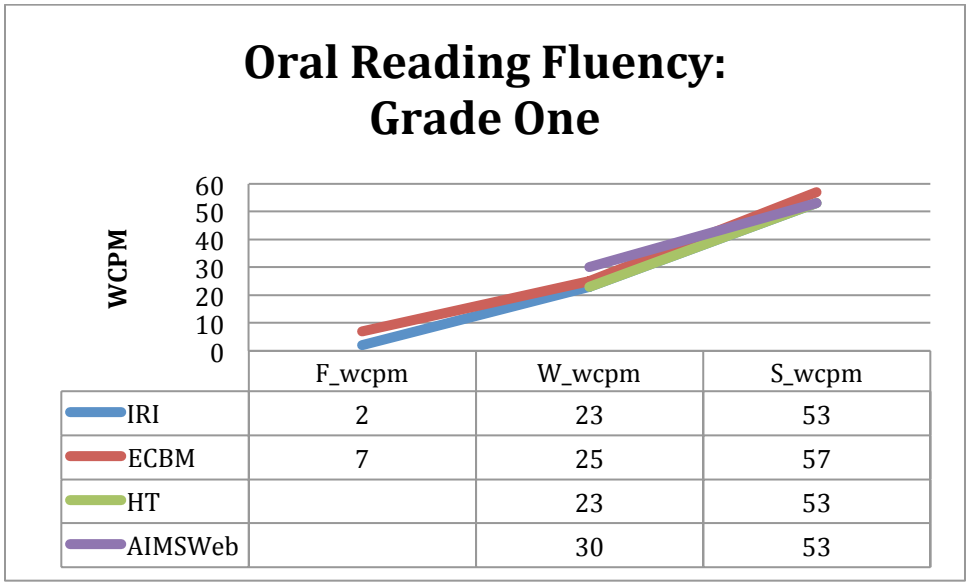


	F_LSF	W_LSF	S_LSF
IRI	2	17	20
ECBM	4	22	35
AIMSWeb	2	20	33

First Grade

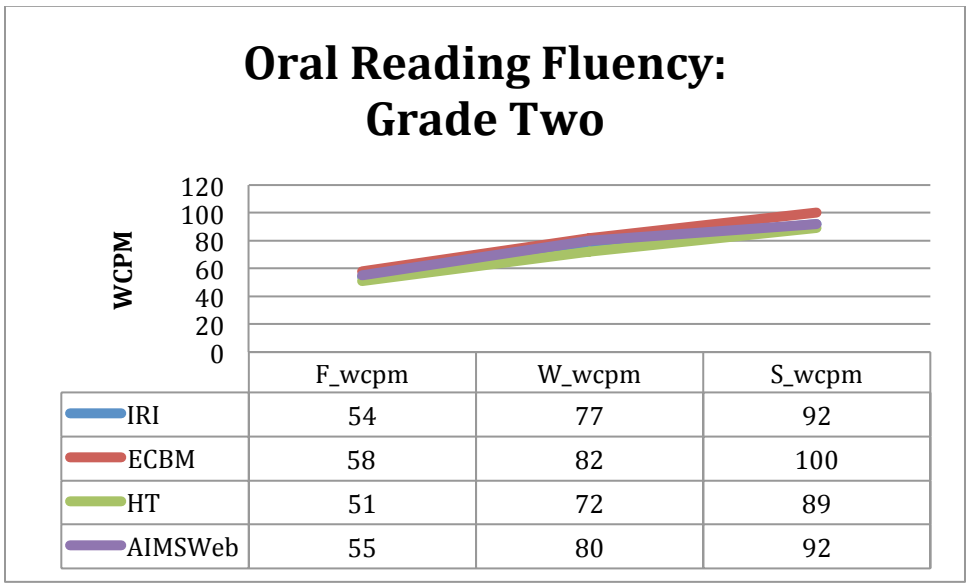


	F_LSF	W_LSF	S_LSF
IRI	31	63	72
ECBM	29	40	46
AIMSWeb	25	40	46



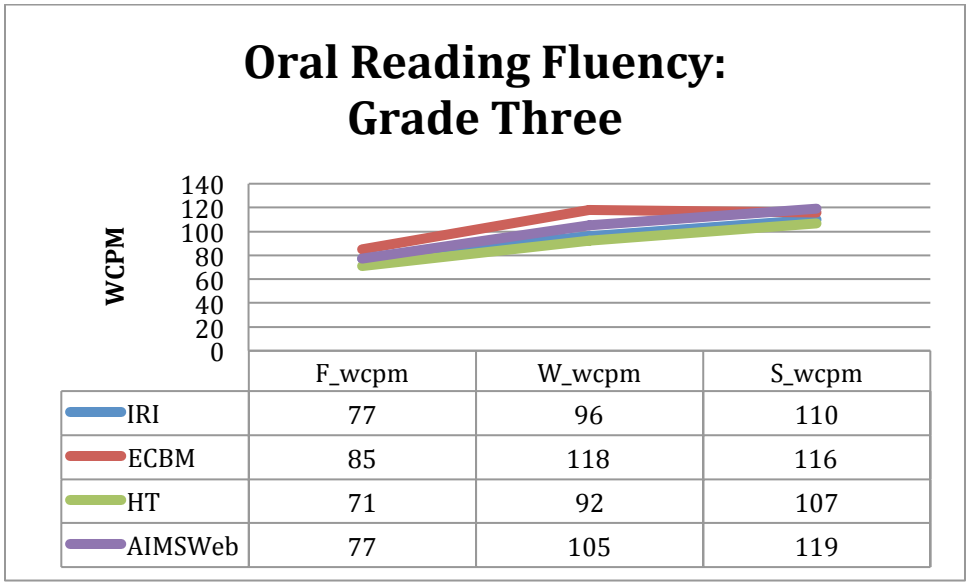
	F_wcpm	W_wcpm	S_wcpm
IRI	2	23	53
ECBM	7	25	57
HT		23	53
AIMSWeb		30	53

Second Grade



	F_wcpm	W_wcpm	S_wcpm
IRI	54	77	92
ECBM	58	82	100
HT	51	72	89
AIMSWeb	55	80	92

Third Grade



	F_wcpm	W_wcpm	S_wcpm
IRI	77	96	110
ECBM	85	118	116
HT	71	92	107
AIMSWeb	77	105	119

The tables and figures provide evidence that the performance indicators, for the most part, align with the 50th percentile value for other rCBM assessments in use nationwide. The two indicators that do not align closely are the Letter Sound Fluency measures for Kindergarten and First Grade. The Letter Sound Fluency expectation for kindergarten students in the Spring evaluation, with the IRI indicator 15 points lower than *easyCBM*. Also, the expected growth is only three points for IRI versus 13 for *easyCBM* for that same time period. The opposite is true for First Grade, where the IRI indicator is 26 points higher than both the *easyCBM* and the *AIMSWeb*. The expected growth is 9 points higher between the Winter and Spring administration on the IRI versus a 6 points expected growth for the same time period on the other assessments. The more important indicator for grade one is the WCPM on Oral Reading Fluency on the Spring evaluation, which is aligned with the other national norms provided.

Identify and Recommend a Benchmark Assessment and an Alternative Assessment (SPED and EL) that aligns with CCSS.

As discussed in sections titled, Test Use (p.11) and Recommend Best Practices (p.16), it is first important to establish the purpose for assessment before a specific assessment can be recommended. According to the statute, the IRI is intended first and foremost as a screening measure used to determine which students may be at-risk of failure with skills that are prerequisite for being successful readers. It is with respect to the purpose of screening that the identification and recommendations for alternatives to the IRI are put forth. A screener, as discussed in the Test Use section of the report, is one way to identify students who may be at-risk for reading difficulties. The screener does not provide diagnostic information, such as how to intervene with a student who may be at-risk. A diagnostic assessment would be typically be utilized for that purpose and the diagnostic could also be linked to specific interventions and progress monitoring tools. Moreover, the diagnostic assessment could be used to reduce false positive errors, thereby allowing the use of a higher initial cut-point on the screening assessment. The use of a higher cut-score would reduce the number of false negative errors on the screener, but increase the number of false positive errors. By administering the diagnostic assessment to all positive screening decisions, some positive decisions would be identified as false positives. Thus, by combining screening and diagnostic information, the impact of false positive screening errors on intervention decisions could be minimized, lowering the cost of intervention by ensuring that intervention is only provided to those students who will not succeed with only quality Tier 1 instruction.

The CCSS do not have a separate set of standards for students with disabilities (SPED) or students who are English Language Learners (EL). There is a document addressing these two specific areas and can be found on the Idaho website: http://www.sde.idaho.gov/site/common/. The area of SPED is typically broken down into two categories, high incidence disabilities (e.g., specific learning disabilities) and low incidence disabilities (e.g., visual impairment) with those two categories having distinctly different recommendations for expectations and accommodations. The Individual Education Plan (IEP) is based on the student's individual needs and this document would take priority for any test accommodations and/or modifications. The recommendations provided below for SPED will involve those students who are considered to fit within the category of high incidence disabilities. As noted under the review of the current format (p.4) the directions to the current IRI include a set of accommodations that align, for the most part, with the research on accommodations for students with disabilities and for students who speak a language other than English. For students who struggle with articulation, the student's speech and language support personnel should provide guidance on what would be acceptable performance on certain speech sounds based on the student's needs as identified via the IEP. Having stated this, it would be redundant to provide a screener for a student who has already been identified as at-risk and is currently being served under an IEP. It would appropriate to provide a diagnostic assessment to determine learning goals for the new academic year. Students whose first language is not English may produce sounds that reflect native language influence. These responses should be credited on assessments for letter names, letter sounds, phonemic awareness and phonics, and oral reading. At the same time, student responses in the native language may or may not be creditable on an English language assessment. Without specific research showing the validity of native language responses for a particular assessment, they would not generally be credited, but students would be given an option to reply in the language of the assessment. The guidelines for test administration, scoring, and interpretation for students with disabilities and students who speak a language other than English must be developed for each specific assessment and established through research. While extant

research on assessments with these populations provides guidance for possible accommodations and modifications, their impact on the validity of test inferences cannot be assumed to be negligible in the absence of research.

The Common Core outlines grade-level expectations through the use of Anchor Standards (National Governors Association Center for Best Practices, 2010) and Foundation Skills for the K-3 population. While the Foundational Skills do not include "new" knowledge, the standards do emphasize an integration of literacy skills so that students are able to read expository text by the fourth grade and that approximately 50% of the materials read are expository. Given that the Foundational Skills are not a new set of standards, or skills, and rely on the research of how students learn to read (e.g., National Reading Panel), the following list provides researched based assessments with benchmarks that are in alignment with the CCSS. It is best for the IRI committee to examine the Technical Manual for each assessment to determine which assessment most closely aligns with the state objective for a screener, and any of the other purposes of assessment of importance to the committee in developing the assessment program. If the intent is to use one assessment that may serve as both a screen and a diagnostic measure, and possibly as a benchmark assessment at multiple time points, the technical report should clearly indicate that the assessment was developed for those purposes. Moreover, the report should provide reliability and validity information for decisions related to each purpose and for each population of students for which the assessment is intended for that purpose. Other factors to consider are the costs involved in administering, scoring and processing the assessment, including the timeliness of reporting and the tracking of student performance, which may differ between assessments delivered through technology and those delivered through paper and pencil. At the same time, regardless of the quality of the assessment and the information provided in the technical manual, it is important to validate cut-points and test decisions for use of the assessment in Idaho. With these caveats in mind, we provide below a list of assessments in use throughout the country to assess early reading.

The list is in alphabetical order.

a. *Basic Early Assessment of Reading (BEAR)* – www.riverpub.com/products/bear
b. *Diagnostic Indicators of Basic Early Literacy Skills (DIBELS)* - https://dibels.uoregon.edu/
c. *Early Reading Diagnostic Assessment (ERDA)* - http://www.pearsonclinical.com/education/products/100000458/early-reading-diagnostic-assessment-second-edition-erdasecondedition.html
d. iStation - http://www.istation.com/
e. *Texas Primary Reading Inventory (TPRI)* - http://www.tpri.org/index.html
f. *Tejas LEE* - http://www.tejaslee.org/default.html

Recommend Scoring Process

There are several dimensions to this question including the assignment of raw scores to student performance, the establishment of cut-points for identifying students at-risk, and the establishment of proficiency standards. Given the current IRI format, the assessment is scored according to typical rules for the assignment of raw scores on fluency-based CBM assessments. We have provided a more comprehensive critique of the IRI administration and scoring guidelines in the first section of the report titled Review of the Current Format of the IRI (page 4). By and large, there is little to be said about the assignment of raw scores to individual test forms. These rules are consistent with standard practice. One concern about the test administration guidelines is that the

failure to advise students to read at a speed to support understanding. Specifically, if the IRI were to remain in tact, the assessment directions for rCBM subtests should be modified to include a sentence stating, "After you read the story, I will ask you a few questions. Begin."

Not supported by specific research with the IRI is the decision to interpret the median score from three forms rather than average the three raw scores, or use all three scores to render screening decisions and decisions about reading proficiency. This decision is based on common practice with CBM assessments, but no research was reported in the IRI technical documents to support this practice with the IRI. It is quite possible that screening decisions could be strengthened by using information from all three test forms in identifying students at-risk. One possible approach would be to follow the current practice using the median score for the initial screening, and to then examine performance on the other two forms to reduce false positive errors. By setting a slightly higher cut-score for the first screening decision, it is also possible that the false negative rate could be reduced using such a "two-step" screening process. However, it is impossible to provide specific cut-points, or specific rules for making such decisions at any particular time point without analysis of test data relating performance on the IRI to performance on criterion reading tests. These cut-points must be established based on an examination of the data and should be established so as to minimize the costs associated with the different types of decision errors.

Recommend state cut-scores

The creation of proficiency categories, their labels, and the assignment of test scores to these categories refer to the process of standard setting. Setting test performance standards involves complex human judgments, is evidence-based, and essentially concerns the process of establishing suitable translations/interpretations for test scores in terms of performance expectations, and should only be undertaken through a formal standard setting process. There are several approaches to standard setting that are common in education and that are based in research (Cizek, Bunch, & Koons, 2004). All standard setting involves human judgments and thus reflects a process of establishing agreement across a group of experts, both on the number of categories, the labels to attach to the proficiency categories (i.e., the Performance Level Descriptors or PLDs), and the placement of cuts on the score distribution to identify the score boundaries that differentiate the proficiency categories. In a test such as the IRI, where there are multiple subtests at some time points, this process may also involve examining and interpreting patterns of test performance across subtests. Because the process ultimately rests on human judgment from content area experts, the process is usually directed by an expert in standard setting who both directs the efforts of the human judges and collects data on the process to demonstrate the reliability and validity of the final score interpretations.

There appears to have been no such standard setting process to derive the PLDs and the associated test scores for the IRI. Dissatisfaction with the proficiency scores is discussed in a single page report by Jenny Fiske, former Reading Coordinator for the State of Idaho, in which she details discussions with Dr. Hulett and Dr. Shin on the possibility of revising the proficiency scores. Nowhere in the available documentation is there any description of a formal standard setting process. Moreover, it is clear from what is described in the documentation that IRI scores were most likely referenced to *AIMSWeb* test scores. Thus, IRI proficiency scores were not established through a standard setting process that related IRI performance to state reading proficiency standards. A formal standard setting process could have been used to relate IRI performance to state reading standards, to performance on the ISAT, or to performance on another standards based reading assessment, any one of which would have allowed IRI proficiency scores to be indexed to reading standards relevant to the students and teachers of the State of Idaho. Without such a formal standard

setting process, the interpretation of IRI proficiency categories in terms of Idaho's reading standards is speculative, at best.

At the same time, the establishment of PLDs for the IRI is antithetical to its use as a screening assessment. The creation of PLDs is not essential to the valid use of the IRI as a screening instrument. The establishment of cut-scores on a screening instrument is essentially a statistical process that is tied to decisions about risk and a desire to minimize decision errors related to risk. On the other hand, standard setting to establish PLDs and proficiency cut-scores is never a purely statistical enterprise. This latter use of cut-scores is typically reserved for outcome assessments, where the establishment of PLDs is an important part in developing the interpretations that will be attached to different levels of performance on the outcome assessment and the decisions about proficiency that will be attached to those scores. It is unusual to attach PLDs to a screening assessment, because the interpretation of cut-scores on a screening test derives entirely from decisions about risk associated with those scores. These interpretations are based on the empirical link between the screening test and the criterion of interest. The importance of cut-scores on a screening test is rooted in their empirical relationship to risk on a criterion measure of some importance, not their interpretation with respect to performance standards. Thus, it would be quite unusual to engage in a formal standard setting process to develop PLD's for the IRI cut-scores because the IRI was developed as a screening assessment, not as a standards based outcome assessment.

Current growth targets: Are new targets needed?

In the section on ***Current Proficiency Indicators*** we provided tables and figures linking current IRI growth targets to targets on other widely used assessments. These tables and figures show that the IRI targets are relatively consistent with those based on other assessments. However, what is not established in those tables and figures is any link between IRI growth targets and risk as reflected in less than proficient performance on standards based reading assessments at the end of each grade. Ultimately, it would be preferable to establish growth targets for the IRI that linked to performance expectations on reading outcome measures of importance to Idaho educators, students, and parents. Without access to data linking IRI performance and changes in IRI performance to such outcome assessments, it is impossible to provide an informed answer to this important question.

Recommend comprehensive training protocols for coordinators and proctors

To train coordinators and proctors on the protocols for test administration, several key elements need to be in place. The first item is to develop a comprehensive technical manual for the assessment. This technical manual provides the basis for the development of training materials that are grounded in the psychometric research behind the assessment. The second step is to develop a standardized training manual that aligns with the testing procedures, addresses common questions and concerns, and provides the wording from the state statute, as well as an overview of the psychometric research supporting the administration, scoring, and interpretation of the assessment. After the training manual is developed, a power point and training notes should highlight the key elements of the technical manual, the training manual, and the assessment. The notes help bring standardization to the training process as they provide all trainers the key points to be covered in training and help to ensure that those key points are covered with all trainees. Trainers may use their own language instead of reading a script provided that the key points of training are clearly articulated. Finally, it is helpful to have a video of successful administrations of the assessment at different grade levels and with all relevant subtests. These videos should showcase key elements to

ensuring valid test administrations. Using video to demonstrate for proctors what a successful administration looks like is helpful as the trainer can stop the video and show how materials are positioned, the words that the proctor should use with the students, appropriate examiner behavior (e.g., no feedback, no facial approvals or disapprovals), and, in general, how long the assessment should take to administer. It is critical to provide proctors opportunities to practice administration of the assessment with trainers present and available to answer questions and provide feedback on administration, scoring, and interpretation.

Appendix A
Glossary of Terms

Types of Assessments

- **Screening** – A screening measure is administered to all students to determine risk status for academic difficulties. Screeners are typically brief assessments where students identified as at-risk are administered a more comprehensive assessment to reduce false-positive errors. In this case the screening measure is the IRI and assesses reading skills.

- **Diagnostic** – A diagnostic measure is administered to students who were identified as at-risk for academic difficulties on the screening measure. This assessment provides teachers instructionally relevant information about student's skills and instructional needs. The diagnostic measure takes longer to administer, as the skill set is more comprehensive so that results can inform instruction.

- **Progress Monitoring** – A progress monitoring measure is a brief but frequent measure of student progress. It is typically given to all students throughout the academic year. Students identified as at-risk by the screening measure and confirmed by the diagnostic measure, should be monitored for academic growth approximately once every two weeks. Students who were not identified as at-risk may be given a monitoring measure once every six weeks or three times a year. The progress monitoring measure would be a standardized measure and would not take the place of formative assessments used daily by the classroom teacher or summative measures.

- **Outcome** – An outcome measure provides information in relation to the level of proficiency each student has made toward the specified year-end goals (i.e., state standards). The information is evaluative in nature and therefore is less instructionally relevant than that obtained from the other three types of assessments.

Standards

- **Benchmark** – Benchmarks are typically specific scores on specific tests that students are expected to attain at particular points in time during the academic year.

- **Standard** – A standard is a 'general statement that represents the information, skills, or both, that students should understand or be able to do– they are broad yet measureable statements.' (Bodrova, E., Leong, D.J., Paynter, D.E., & Semenov, D., 2000, p.9)

- **Standards Setting** –
    - A process by which a group of experts establishes a) performance standards, b) specific scores on tests that map to those performance standards, and c) the performance level descriptor assigned to each performance standard.
        - Content Standards: refers to statements that describe specific knowledge or skills over which examinees are expected to have mastery for a given age, grade level, or field of study.
        - Performance Standards: define 'how much' or 'how well' examinees are expected to perform in order to be described as falling in a giving

category. (both definition are direct quotes from NCME Instructional Module by Cizek, Bunch, & Koons)

Measurement Terms

- **Equipercentile Equating** – one of several processes fro equating raw scores from different forms of the same test.

- **False Positives Error** – The identification of students for intervention who do not required the intervention.

- **False Negatives Error** – The failure to identify students for intervention who, in fact, require the intervention.

- **ROC Curve** – Receiver operating Characteristic Curve. A geographical display showing the relationship between false positive and false negative errors associated with different binary decision rules (i.e., different cut scores on a test).

- **PLD** – Performance Level Descriptors. PLDs are more complete and detailed descriptions of what constitutes performance within a particular score category on a standards based test. For example, NAEP uses three labels, basic, proficient, and advanced.

Reading Terms

- **Phonemic Awareness** - the ability to hear and manipulate the sounds in spoken words and the understanding that spoken words and syllables are made up of sequences of speech sounds (Yopp, 1992)
- **Phonics** - Phonics instruction is a way of teaching reading that stresses the acquisition of letter-sound correspondences and their use in reading and spelling. The primary focus of phonics instruction is to help beginning readers understand how letters are linked to sounds (phonemes) to form letter-sound correspondences and spelling patterns and to help them learn how to apply this knowledge in their reading. (NRP, 2000)
- **Fluency** - Fluent readers are able to read orally with speed, accuracy, and proper expression. (NRP, 2000)
- **Vocabulary** –knowledge of words and word meanings. It is not a concept that is fully mastered but one that continually develops over time. (Honig, Diamond, Gutlohn, 2008)
- **Comprehension** – What one understands about what one has read. Comprehension is critically important to the development of children's reading skills and therefore to the ability to obtain an education. Indeed, reading comprehension has come to be the "essence of reading" (Durkin, 1993), essential not only to academic learning in all subject areas but to lifelong learning as well. (NRP, 2000)
- **Close Reading** – careful and purposeful reading that is in essence the ability to reread with author purpose in mind.

Special Education

- **IDEA** – Individuals with Disabilities Education Improvement Act (2004).
    - **IEP** – Individualized Education Program. This document includes appropriate accommodations necessary to measure student growth and achievement.
    - **RtI** – Response to Intervention – a comprehensive system for delivering quality education to all students within the general education framework. Under IDEA, 15% of federal funding can be allocated for providing services to students considered at-risk for reading failure who have not yet been identified for special education.

References

AIMSWeb: National Norms Technical Documentation 2011. Pearson. Retrieved from: http://www.aimsweb.com/wp-content/uploads/AIMSweb-National-Norms-Technical-Documentation.pdf

AIMSWeb Target Scores 2012-2013. Retrieved from: http://www.usd379.org/DocumentCenter/Home/View/4338.

Bodrova, E., Leong, D.J., Paynter, D.E., & Semenov, D. (2000). A framework for early literacy instruction: Aligning standards to developmental accomplishments and student behaviors. ERIC ED 465-183.

Cizek, G.J., Bunch, M.B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice*. NCME Instructional Module, p. 31-50.

EasyCBM: Interpreting the EasyCBM Progress Monitoring Test Results, 2012-2013. Riverside. Retrieved from http://www.easycbm.com/static/files/pdfs/info/ProgMonScoreInterpretation.pdf

Fisher, 2013

Francis, D.F., Santi, K.L., Barr, C., Varisco, A., Fletcher, J., & Foorman, B.F. (2008). Form Effects on the Estimation of Students' Oral Reading Fluency using DIBELS. *The Journal of School Psychology, 46*(3), 315-342.

Hasbrouck, J. & Tindal, G.A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *The Reading Teacher, 59*(7), p. 636-644.

Honig, B., Diamond, L., & Gutlohn, L. (2008). Teaching Reading Sourcebook: Sourcebook for Kindergarten Through Eight Grade. Novato, CA: Academic Therapy Publishing.

Moats, L. C. (1999). *Teaching reading is rocket science: What expert teachers of reading should know and be able to do*. Washington, D. C.: American Federation of Teachers.

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common core state standards*. Washington DC: Author.

National Reading Panel (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction [on-line]. Available: http://www.nichd.nih.gov/publications/nrp/smallbook.htm.

National Research Council (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Shinn, M. R. (Ed.). (1989). *Curriculum-based measurement: Assessing special children*. New York: Guildford Press.

Yopp, H.K. (1992). Developing phonemic awareness in young children. *The Reading Teacher, 45*, p. 696-703.