Achieve, Inc.

Setting the Bar

An Evaluation of ISTEP+ Assessments for

INDIANA

ACHIEVE'S BENCHMARKING INITIATIVE

Achieve, Inc.

Achieve, Inc., is a bipartisan, nonprofit organization created by the nation's governors and corporate leaders to help states raise academic standards, improve assessments and strengthen accountability to prepare all young people for postsecondary education, work and citizenship. To do this, we:

- help states benchmark their standards, assessments and accountability systems against the best in the country and the world;
- provide sustained public leadership and advocacy for the movement to raise standards and improve student performance;
- build partnerships that allow states to work together to improve teaching and learning and raise student achievement; and
- serve as a national clearinghouse on standards and school reform.

Michael Cohen, President

Matthew Gandal, Executive Vice President

Jean Slattery, Director of Benchmarking

SETTING THE BAR

AN EVALUATION OF ISTEP+ ASSESSMENTS FOR

INDIANA

Prepared by Achieve, Inc.
August 2003

TABLE OF CONTENTS

ACHIEVE'S WORK IN INDIANA

The nation's governors and business leaders created Achieve, Inc., in 1996 to help states raise their academic standards, measure results against those standards, and hold schools and students accountable for their performance. One way Achieve fulfills its mission is to benchmark the quality and rigor of a state's standards against the best examples from the United States and abroad and to examine the alignment between a state's standards and the tests it uses to measure student performance.

In 1999, Indiana was among the first states to ask Achieve to appraise the quality of its standards and tests. In its report to the Education Roundtable, *Measuring Up: A Report on Education Standards and Assessments for Indiana,* Achieve noted that while the state's standards were detailed and specific, they were not as rigorous or as focused as those from other leading states and nations. The report also indicated that Indiana's tests were less challenging than they could be, most likely as a result of the low level of demand in the state standards. Achieve recommended that the state strengthen the standards, making them clearer and more challenging, and we recommended that new assessments be developed to measure the new standards.

Indiana has worked diligently to respond to that 1999 report. The state has a new generation of standards and tests. Its expectations are clearer, more explicit and more challenging. We now consider the Indiana standards to be among the best in the nation. At the same time, the state has taken important steps to hold schools and students accountable for meeting higher standards. In some areas, including its effort to align high school assessments with the expectations of college coursework, Indiana is leading the nation.

INDIANA'S CHARGE TO ACHIEVE

Having completed development, field-testing and administration of its new ISTEP+, Indiana asked Achieve to examine its assessments in English/language arts and mathematics to make recommendations for improving them. Specifically, the state asked Achieve to:

- Analyze the quality and rigor of its English/language arts and mathematics tests in grades 3, 6 and 8 and their alignment to the state's new standards;
- Analyze the proposed scores for passing to determine if they represent "solid academic performance";
- Analyze proposed pass+ scores to determine whether they represent "exemplary performance"; and
- Propose alternative, more appropriate scores, if necessary, for passing and pass+.

Achieve has had extensive experience evaluating the quality and alignment of state standards and assessments, having performed such analyses for over 15 states since 1999. This is the first time, however, that Achieve has evaluated cut scores on state tests. To respond to that part of Indiana's

charge, Achieve partnered with the National Center for the Improvement of Educational Assessment (NCIEA), an organization with considerable technical expertise in assessment.

It is important to note that no state has ever asked Achieve or NCIEA for this sort of in-depth, independent evaluation of its cut scores. To the knowledge of the experts involved in this analysis, it is altogether unprecedented for a state to seek such a review. States and publishers typically have set passing scores using procedures that are highly technical and have kept the details secure. As a result, it has been very difficult for educators, parents or policymakers to get a firm handle on what it takes to pass the tests.

Under the new federal No Child Left Behind law, state assessments must be aligned to grade-level standards/benchmarks, and every school in Indiana and other states will soon be judged based on its ability to raise the percentage of students achieving proficiency on these tests. This will no doubt lead to much greater scrutiny of states' definitions of proficiency, and it is likely that more states will follow Indiana's lead of seeking an external review.

Achieve undertook two distinct but mutually reinforcing studies, one on the alignment of the tests with the standards and one analyzing the reasonableness of the passing scores for each test. A comprehensive technical report with supporting details and item-by-item data has been provided to the Indiana Department of Education. This report must be kept confidential since it contains references to secure test items.

THE ALIGNMENT STUDY

Alignment is a measure of the extent to which standards and assessments are in agreement and work in conjunction to guide and support student learning. It is not a "yes or no" question; rather, alignment is a considered judgment based on a number of factors that collectively determine the degree of match between a state's standards and the assessment used to gauge if students are meeting those standards. At its core, the Achieve analysis answers two key questions: Can everything on the tests be found in the standards? And conversely, do the tests do an effective job of measuring the knowledge and skills set forth in the standards?

METHODOLOGY

To determine how closely each Indiana test was aligned to the related grade-level standards, Achieve convened two teams of content experts who followed a subject-specific, stepwise procedure or protocol that Achieve has used with numerous tests in more than a dozen states.

In the first phase of the review process, a team of content experts evaluates each individual test question to determine 1) if it actually measures the indicator to which the test developer assigned it; 2) how well it matches the content and performance described in the related standard; 3) whether it is constructed fairly; and 4) how intellectually challenging it is. These are key issues. The information gained from a test is no better than the collection of items that make it up. If an item measures content and skills beyond what is contained in the standards, it is less likely that it will have been taught in classrooms. Similarly, an item that is flawed for such reasons as having no right answer or two right answers or a misleading graphic will not give accurate information about students' performance. Tracking individual items for the level of cognitive demand each poses also is critical. If a test truly is standards-based, it should have a mix of basic and more challenging items that reflect the range of concepts and skills spelled out in the standards so differences in the performance of proficient and non-proficient students can be detected. In summary, Achieve's item-by-item analysis not only yields valuable information about critical aspects of alignment, but also provides quantitative data that contribute to the judgments made in the remainder of the review.

In the second phase of the alignment study, content experts take a more holistic view of the test in order to judge if the test is balanced overall and if it is appropriately rigorous for the grade level. Moving away from the item level, reviewers consider the test a standard at a time — such as literary response or geometry — and look at the collection or set of items that are meant to assess each standard.

To judge how balanced the set of items mapped to each standard is, experts ask, "Does this set of items succeed in measuring the breadth and depth of content and skills described in the standard?" Or said another way, "to what extent does the set of items assess the key content and skills in the standard?" Because a single on-demand test cannot assess all the indicators that make up a state's standards, it is critical to determine how well the items on a test sample the indicators. Content

experts also examine the reading passages as a set to ensure that both literary and informational text are represented fully.

In evaluating the rigor of a test, content experts follow the same general procedure they used when evaluating balance. They compare the overall intellectual demand encompassed by a set of items to the level of intellectual demand described in the related standard. Looking at each standard in turn, they ask, "Does doing well on the item set that measures this standard mean the student has mastered the challenging material contained in the standard?" Because experts rated each item earlier in the process as to its level of cognitive demand, they can determine if an item set has a span of difficulty appropriate for the grade level. Content experts also review the reading passages as a set to determine if they have a span of demand appropriate to the grade level tested.

At the close of the analysis, reviewers look across the standards, at the test as a whole, to determine how good a job the test does in measuring the knowledge and skills described by the standards and how rigorous the test is overall.

FINDINGS

The following is a summary of findings from the alignment study. The findings in English/language arts and mathematics differ, so we have broken out each subject separately to give Indiana policymakers a clear understanding of the strengths and weaknesses in each subject.

Strengths of Tests Overall

- Alignment: As noted earlier, alignment is a two-way street. On the one hand, everything on the test should be able to be traced back to the standards. But on the other, it is equally important that the breadth and the depth of the standards be effectively measured by the tests. Achieve determined that Indiana's assessments in English/language arts and mathematics meet the first test of assessment-to-standards alignment: Everything on the tests can be found in the standards. We discuss both of these aspects of alignment in the subject area discussions below.

- Inclusion of constructed-response items: Indiana has made a wise choice in including short-answer and constructed-response items on its tests — a hallmark of high-quality assessments. These item types are essential because of their ability to measure more advanced skills and their positive influence on how material is taught and learned in classrooms.

- Release of applied skills tests: Indiana's decision to release publicly its applied skills assessments along with scoring rubrics, student work and descriptions of the ways that each piece of student work meets the criteria for each of the score points is highly commendable. It helps make tests transparent for teachers and students and helps ensure that assessment is in fact an instructional tool.

English/Language Arts

STRENGTHS:

The ISTEP+ assessments are good tests of basic reading and writing skills that should be expected of students entering grades 3, 6 and 8. In particular, they have the following exemplary characteristics:

Alignment: On the whole, the questions on Indiana's English tests measure the knowledge and skills set out in the state standards. That is, everything on the tests can be traced back to concepts in the standards. We did find, however, that in a fair number of cases the tests measured standards above or below the grade level targeted. We explain the significance of this below.

Reading passages: The reading passages used in all three grades include a wide range of genres — poetry, narrative and informational text in grade 3, for example — and come from authentic sources. In addition, the accompanying illustrations and graphics served to aid student comprehension rather than impede it as we have seen in many other tests. The difficulty of the reading passages was appropriate for each grade being tested but probably would not challenge advanced readers (particularly at grade 8).

Direct assessment of writing: Indiana has made the right choice to assess writing directly. In particular, the grade 3 assessment and the released scoring materials for all three grades stand out. Given this commitment to direct writing assessment, the state should reconsider the use of fill-in-the-blank questions to measure *writing* standards. This procedure is a type of assessment format developed for measuring *reading* comprehension and is best used for that purpose.

Item construction: With a few exceptions — two writing prompts and one reading item at each grade — the questions were fair evaluations of students' knowledge and skills that did not contain tricks or inappropriate tasks.

CONCERNS:

Mapping of test items: As stated above, while all of the items on the tests can be matched to Indiana standards, some items measure standards other than the ones the state set out to measure. For example, roughly a third of the grade 3 items measure grade 1 (rather than grade 2) standards; a similar percentage of grade 8 items measure grade 5 (rather than grade 7) standards. At the other end of the spectrum, one-fifth of the grade 3 items measure grade 3 (rather than grade 2) standards. In addition, a number of items are mapped to the correct grade level but assigned to the wrong indicator. This may be due in part to the fact that Indiana has decided to include a number of anchor items from previous years' tests in order to maintain trend data.

Nonetheless, the issue of items that miss the standards and/or indicators they purport to measure is not simply a technicality. First, if items measure knowledge and skills beyond those to which

students should have been exposed — for example, content from grade 3 standards when the test is designed to measure performance through the end of grade 2 — it is likely to catch teachers and students off guard. Students should be expected to demonstrate mastery only of content they have been taught.

Second, if the standards and/or indicators that are being measured are misidentified, score reports sent to teachers and students can create distorted views of student achievement. Teachers and students may believe literary analysis is being mastered, when in fact it has not even been measured. If instruction is altered as a result, important knowledge and skills inadvertently may be given short shrift.

Balance: Literary response and analysis is one of the seven principal categories included in Indiana's English/language arts standards and contains significant content. As such it deserves to be assessed with items that are true to the knowledge and skills described in the indicators. On all three tests, Achieve found that this important standard was not well measured.

Rigor: Overall, the Indiana tests are less challenging than the state standards imply they should be. Several factors contribute to this lack of rigor:

1. As mentioned above, a number of items measure standards below the grade intended. It is true that Indiana's standards are cumulative (i.e., the expectations for each year build on the skills and knowledge of the previous year), so arguably the inclusion of questions that measure standards from grades lower than the grade being assessed is an understandable and acceptable consequence. This is particularly true since, in the past, the tests have been given only in key grades. However, the percentage of these items is rather high and not reflected in the test blueprints. In re-examining this practice, Indiana's policymakers should keep in mind the new context in which ISTEP+ will be given. As required by the federal No Child Left Behind Act, Indiana soon will be implementing a new grade-by-grade testing system. The inclusion of many below-grade items on a series of assessments given in each grade would diminish the value of grade-by-grade test results and would tend to lower rigor across the board.

2. Most of the test items make low or modest cognitive demands on students. At grade 8, for example, 82 percent of the items require only low inferential skills, placing them at the low end of the Achieve four-point scale for cognitive demand. In addition, many of the English/language arts indicators describe compound tasks such as "recognize and assess" (5.2.3), "draw inferences and support them"(5.2.4), and "identify and correctly use" (5.6.1, 5.6.3 and 5.6.4). The tests tend to measure the least-demanding performance where more than one performance is described by a single indicator. When there is a pattern of assessing the least-demanding performances on a test, for example, asking students to only "recognize" but not "assess," the overall rigor of the test is undermined.

3. As mentioned above, the indicators for literary analysis are not measured adequately. For example, at grade 3, of the 30 items dedicated to assessing reading, nine were intended to measure literary response and analysis, but in fact none did; instead they measured general comprehension. Since literary response and analysis is one of the most intellectually demanding standards, the absence of items assessing this standard suppresses rigor.

Writing prompts: While the writing prompts and scoring guides for grade 3 are outstanding, the prompts and related scoring guides at grades 6 and 8 are not as strong. We encourage the state to ensure that future prompts present a clear and fair set of tasks to students.

Mathematics

As noted earlier, Indiana's revised math standards are rigorous and offer the opportunity to build challenging tests.

STRENGTHS:

Alignment: As with English/language arts, all of the skills and knowledge measured by the math tests can be found in Indiana's standards. Unlike English/language arts, however, the issue of questions measuring standards from grades lower than those targeted by the tests is far less prevalent in math; less than 5 percent of the items mapped to lower grades. Also, the fairness issue, raised by items measuring higher-grade knowledge and skills in English/language arts, is all but absent on the math assessments.

Special features: In addition to open-ended items, Indiana includes "gridded-response" items at grade 8, which require students to write the numerical answer directly. This item format demands more thinking on the part of students since gridded-response items do not provide students with a choice of answers — they have to come up with the answers themselves. The state also has gone to great lengths to include the use of manipulatives, to partition the test effectively to permit some use of calculators and to provide grade 8 students with formula sheets.

CONCERNS:

Breadth at the expense of depth: The math assessments in all three grades seem to emphasize breadth over depth, suffering from the mile-wide, inch-deep phenomenon noted in the Third International Mathematics and Science Study. With a few exceptions, most of the skills measured on the tests are assessed only by one or two items, which invariably limits how thoroughly those concepts and skills can be measured. As a result, very few of what we would consider essential knowledge and skills are measured with the level of depth and precision that would allow Indiana educators and policymakers to conclude that students have mastered these skills.

Problem-solving skills: Indiana commendably attempts to measure students' ability to use their mathematical knowledge to solve problems they might face in their everyday lives. However, the way in which some of the test items are mapped to problem-solving indicators first and to content indicators second is a cause for concern. As with English/language arts, the current maps of items could create misimpressions, in this case about students' ability to problem solve. Teachers could conclude students cannot solve problems well, when in fact their difficulty may arise from trouble with the math content that is at the heart of the questions. The current mapping schema may have the effect of minimizing the importance of the mathematical content being measured. We recommend reversing the emphasis by mapping these items first and foremost to content indicators and only secondarily to problem-solving indicators.

Use of open-ended items: As mentioned earlier, Indiana made a wise decision in including gridded-response and constructed-response items on its tests. However, the state does not make the best use of these test items by targeting them to aspects of the standards that are not assessed easily by multiple-choice items. In many cases constructed-response items required only straightforward application of procedures — easily measured with multiple-choice items — rather than skills such as reasoning, developing a sequence of steps or weighing two approaches to the same problem. For example, the grade 7 standards call for students to graph a line given its slope and one or two points. This important performance was not assessed, even though it is an ideal topic for a constructed-response item.

Item construction: The number of test items with technical problems is greater in math than in English/language arts, with several at each grade level offering an inappropriate source of challenge for students. The largest proportion of problematic items (13 percent) occurred in grade 3. If an item has a technical problem, students may either get the right answer for the wrong reason or get the wrong answer when in fact they have the knowledge to answer the item correctly. The greater the number of such items on a test, the greater the likelihood that teachers will draw incorrect conclusions about what their students have and have not learned.

Rigor: As was the case in English/language arts, the level of challenge posed by the math assessments does not match the rigor represented in Indiana's standards. This happens for several reasons:

1. Most of the items at all three grade levels stress recall or the application of a simple procedure or algorithm, rather than a deeper, more conceptual understanding of math. As in English, these items fall at the low end of Achieve's scale for cognitive demand. While a small proportion of questions require the use of reasoning skills, the skills needed tend not to be of a complex or abstract nature. The grade 6 test had the least-challenging items.

2. The items themselves tend to lower the cognitive demands made on students. Achieve found, for example, that multiple-choice questions tended not to have distracters tied to common mistakes students make. This characteristic makes guessing easier, but it's not the only effect. When distracters are based on common student errors, analyzing the

patterns of student responses to an item can help pinpoint students' source(s) of confusion. Consequently, beyond the impact on the test's rigor, this shortcoming also deprives educators of important information that could diagnose learning gaps.

3. The balance of items leans heavily toward repeatedly measuring easy-to-assess indicators and testing the least-challenging aspects of other indicators.

4. Across all three tests, questions measuring geometry tended to be the least demanding, while those measuring data analysis and probability were the most demanding.

Balance: As described above, Indiana's math tests tend to measure easy-to-assess indicators multiple times at the expense of other significant concepts and skills. Examples of concepts and skills not measured include skip counting and the comparison of unit fractions (Indicators 2.1.1 and 2.1.9) in grade 3; simple percent, perimeter and surface area (Indicators 5.1.4, 5.5.2 and 5.5.4) in grade 6; and estimation, use of mental arithmetic, and construction of two-dimensional patterns or "nets" for three-dimensional objects (Indicators 7.2.4, 7.2.5 and 7.4.4) in grade 8.

EVALUATION OF PASSING SCORES FOR ISTEP+

The aim of a standards-based education system is for all students to acquire the knowledge and skills described by a state's content standards. Tests are the tool for measuring how well students have mastered the expected knowledge and skills and scores established for passing represent the level of mastery a state deems satisfactory. For the system to work, the tests must do a good job measuring the standards, and the passing scores must be meaningful.

The major difference between "standards-based" and "norm-referenced" assessment systems is how student performance is evaluated. In a norm-referenced assessment system, students are compared to one another and to a representative "norming" sample. These tests are meant to compare student performance across many states; as a result, they may be tied only loosely to any one state's standards. Because achievement is judged in relation to other students, scores often are reported as percentile rank (e.g., the percentage of students that an individual outscored).

In a standards-based system, on the other hand, students and schools are judged based on how closely they meet state content standards. The emphasis in reporting (at the school level) often is on the percentage of students actually meeting a particular performance level (e.g., basic, proficient, advanced). Performance levels provide meaningful targets for the educational system, and when students can meet the highest performance levels, there is evidence of real educational achievement.

Performance levels are delineated by cut scores. An important feature of meaningful cut scores is that students who perform at or above that score possess *qualitatively* different knowledge and skills than students who fall below that score. In other words, students should not be considered higher performing simply because they answer a greater number of similar questions correctly on a test than do lower-performing students; rather, they should be considered higher performing only if they perform better on questions that require more advanced knowledge or skills, deeper understanding, or greater fluency**.**

METHODOLOGY

It is highly unusual for a state to invite this level of scrutiny of its passing scores from an independent, third-party reviewer. As stated earlier, states and publishers typically set passing scores using technical, secure procedures. As a result, it has been difficult for educators, parents and policymakers to fully understand what passing students actually know and can do. Indiana is leading the nation in this regard.

The analysis of Indiana's passing scores proceeded in two stages. First, we set out to determine what students need to know and be able to do to pass each test. Next, we made a judgment as to whether that passing score represents a rigorous but reasonable expectation for students at the grade level. (Indiana refers to this as "solid academic performance.")

In responding to Indiana's request that we describe the knowledge and skills associated with proposed passing scores in English/language arts and math, Achieve and NCIEA went beyond an evaluation of the knowledge and skills of the lowest-performing students who still pass — those students who perform right at the passing score. Due to measurement error alone, these "just-passing" students have a strong likelihood of landing on either side of the passing score if they were to take the test again. To better understand the knowledge and skills characteristic of a passing performance, Achieve examined the test results of students who would likely pass the test if they took it repeatedly — students who scored at least one-half standard deviation above the passing score. We refer to those students as "clearly passing." We also examined the performance of "clearly failing" students, those who fell at least one-half standard deviation below the passing score. In doing so, we analyzed patterns of student performance on the different standards and also on the items that measure each standard to determine if performance on particular standards and/or items helped to differentiate the skills and knowledge that each group is likely to possess. By contrasting the performance characteristic of borderline passing students with the performance of those who clearly passed and those who clearly failed, we also were able to examine the likely effect of moving the passing score up or down.

In order to pinpoint the knowledge and skills students need to pass the tests, we first set out to determine which questions on each test students needed to answer correctly to pass. If we had that information, our content experts could simply look to see which concepts and skills those questions were measuring, and that would tell us what students need to know to pass the test.

The analysis proved to be much more complicated than anticipated. Like many other states, Indiana uses a "compensatory" standard-setting procedure, which allows high scores in some areas to compensate for low scores in others. *This means that there is no set of questions that all students have to answer correctly to pass.*

By looking at all of the possible combinations of items that passing students answered correctly, we hoped to find a small number of patterns that would reveal differences between the knowledge and skills passing students have as opposed to those who do not pass. Unfortunately, we found no common scoring patterns. *In fact, there are almost as many ways to reach a passing score as there are students who pass.*

Given this lack of common scoring patterns, we probed further to examine whether there are general patterns of average number of correct responses of students scoring at different points along the scoring continuum (e.g., students who were "clearly failing," "just passing" and "clearly passing"). To do this, we compared the "effect-sizes" for the various standards for groups of students scoring at different total score intervals. This comparison enabled us to determine in which standards higher- and lower-performing students had similar average scores and in which standards their performance differed. Examining these patterns led to an understanding of how the content and skills in different standards helped to spread students' scores above and below the passing scores. That is, it showed which standards had the most impact in separating the three groups of students. Contrasting the types of performance characteristic of just-passing students

with the performance of those who clearly passed and those who clearly failed allowed us to examine the likely effect of moving the cut score up or down.

Stage two of the analysis required content experts to reflect on the results of stage one — the knowledge and skills required to pass each test — and make a judgment as to whether the passing score represents "solid academic performance." However, our ability to do this was hindered by the fact that Indiana does not have definitions of "solid academic performance" that clearly delineate the knowledge and skills students performing at that level should possess. In contrast, other states have made these decisions up front, clearly articulating what students performing at each level should be able to do, and have communicated this to test makers and to educators and the public as well. The National Assessment of Educational Progress (NAEP) also has done this for their basic, proficient and advanced achievement levels.

Without a clear definition from the state, Achieve and NCIEA were left to judge whether the passing scores are rigorous but reasonable, based on the following factors: the content included in Indiana's standards, how well the general patterns of student responses related to identifiable knowledge in the respective discipline and our experience working with other states.

Although the primary task in this study was to examine the passing scores for "pass" on ISTEP+, Indiana officials also asked Achieve and NCIEA to comment on whether the scores for "pass+" reflect "exemplary" performance. While we did not do as thorough an analysis of this as we did for pass, we do comment on pass+ in this report.

FINDINGS

One of the most important findings to emerge from this part of the study is that the issues of test design raised in the alignment analysis — the quality, rigor, focus and depth of the test questions themselves — has a considerable impact on the state's ability to describe categories of performance represented by different passing scores. In other words, in describing what it takes to pass the ISTEP+ tests, or to achieve at the pass+ level, we are completely dependent on what the tests are actually measuring. If some concepts are not measured well, we are unable to say that students passing the test have learned them. In math in particular, this proved to be the overriding issue.

A summary of findings is provided below. As was true for the alignment analysis, the results differ for English/language arts and mathematics.

- ***What Does It Take To Pass the ISTEP+ Tests in English/Language Arts and Mathematics?***

ENGLISH/LANGUAGE ARTS

In English/language arts we found a distinction between what passing and non-passing students know and can do in terms of reading comprehension skills. Items mapped to comprehension and

literary analysis had the greatest impact in separating students, particularly in grades 6 and 8. This pairing of standards should not be surprising given the alignment study's finding that most items described as literary analysis would more accurately be described as general comprehension questions. To a lesser degree in those two grades and to a greater degree in grade 3, vocabulary questions helped to differentiate students.

We did not see a similar distinction between passing and non-passing students' performance on the writing portion of the tests. Reading rather than writing skills proved to be the key determinant in whether students pass the tests.

The reading skills represented by the set of items that passing students tended to answer correctly are different from the set that characterize non-passing students. Students who score at or above the passing score are able to make sense of and draw conclusions from what they read, rather than simply retrieving explicitly stated information. For example, rather than simply identifying information or restating the main idea, they are able to infer that one event caused another or conclude what the main point of a paragraph is based on the evidence provided.

There also are differences between what "just-passing" and "clearly passing" students know and can do. In examining performance on reading comprehension items specifically, Achieve found that students in grades 6 and 8 who scored at the passing point appeared to possess *basic* reading skills such as locating information in informational text and making "local" inferences — inferences that involve two pieces of information in consecutive sentences in the same paragraph. By comparison, "clearly passing" students appeared to be able to make more "global" inferences and draw conclusions that may not be readily apparent from the information provided. When compared to the descriptions of "basic," "proficient" and "advanced" reading performance used by NAEP and states such as Illinois and New York, students who "just pass" in Indiana demonstrate basic skills, while "clearly passing" students showed skills more characteristic of proficient readers.

In contrast to grade 6 and grade 8 tests, on the grade 3 test, vocabulary skills played an important role in distinguishing among students scoring below, at and above the passing score. For example, a greater proportion of students at or above the passing score compared with those scoring below the passing score are able to recognize and use context to define words and phrases. That said, reading skills were still the main determinant of students' overall English/language arts performance. Students who just pass appear to be able to read at a basic level of comprehension, in that they are able to locate information, make simple inferences and determine the main idea of a text. Clearly passing students generally are able to understand the literal level of a text and make inferences more successfully than those students who just pass. Students who solidly passed this test appear to be stronger readers in terms of basic comprehension skills than students just passing the test, yet some of the more demanding skills that characterize proficient readers are still somewhat beyond their grasp. Finally, the writing standards contributed considerably less than reading to the total English/language arts score at both grade 6 and grade 8, but this disparity was even more pronounced for the grade 3 test results.

MATHEMATICS

In contrast to English/language arts, the study revealed *no* clear distinction between what passing and non-passing students know and can do in mathematics, with the possible exception of number sense in grade 6. Students scoring higher on the test, whether at the passing score or clearly passing, simply have a higher probability of getting more items correct than students who are lower on the achievement continuum. We are not able to say that higher-scoring students possess a qualitatively different set of knowledge and skills than lower-scoring students. We found that the questions that discriminate well between clearly passing and just-passing students are not linked to a major concept or important skill, but rather appear to be items distributed across a number of constructs.

The fundamental reason for this finding lies not in where the passing scores were set, but rather in the way the math tests were constructed. As discussed in the alignment findings, the math tests do not measure most key concepts in enough depth to draw conclusions about student mastery of those concepts. It does not appear that the state set clear priorities in this regard; test items did not cluster around any identifiable "big ideas," such as proportionality or linear relationships. Instead, many topics are assessed, but only by a single item or two, hindering experts' ability to generalize to a larger set of skills. As a result, it was quite difficult to identify conclusively a set of knowledge and skills that distinguishes students who just pass, those who are clearly passing and those who are clearly failing.

As mentioned earlier, Achieve and NCIEA believe that different performance levels should be based on qualitatively different knowledge and skills, not simply on more or less of the same knowledge and skills. What seems to distinguish passing from non-passing students on the ISTEP+ math tests is that those who pass — and those who clearly pass — get a greater number of test questions correct. We cannot say with confidence that students who score higher have a better grasp of any particular domain in math, nor can we say that they have a higher level of cognitive skills.

- ***Do the Passing Scores Represent "Solid Academic Performance"?***

As mentioned earlier, because Indiana does not yet have of clear description of what "solid academic performance" means, Achieve and NCIEA made our own judgment of whether the passing scores are rigorous and reasonable.

ENGLISH/LANGUAGE ARTS

In so far as a measure of reading proficiency is concerned, we believe the current passing scores are reasonable with the caveat that literary analysis is underassessed. When Achieve's experts remapped items to the standards they measured most closely, they were able to identify several major constructs, such as literal comprehension and drawing inference from text. Some of these constructs were measured by enough test items that our experts could detect differences in the

performance of students who clearly pass, just pass and clearly fail, relative to these big ideas. To be specific, the passing score differentiates among students who possess the skills characteristic of "basic readers" from those who have yet to fully demonstrate those skills. Similarly, clearly passing students tend to demonstrate the skills more characteristic of "proficient readers," although this is more the case in grades 6 and 8 than in grade 3. Nevertheless, a passing score that separates students who demonstrate "basic" reading skills from those who do not appears very reasonable in the context of No Child Left Behind.

As mentioned earlier, writing is not a clear determinant in whether students pass the Indiana tests. Indiana clearly values writing, which is why the state assesses writing directly and has developed such exemplary supporting materials. In order to place even more of a premium on writing, Indiana could consider the following courses of action: First, work to revise both the prompts and rubrics to make sure that they are as sensitive as possible to good writing instruction. Second, weight the writing items more heavily so that writing is more of a factor in students' overall scores. Third, after conducting a separate standard-setting procedure for writing, report the reading and writing performance levels separately. And fourth, if time and resources permit, consider adding another writing prompt or replacing existing items (particularly the fill-in-the-blank items) with items that more effectively measure writing skills.

MATHEMATICS

Given the limitations of the math tests pointed out earlier, we cannot say with confidence that students who pass have a "solid" grasp of the Indiana standards. The tests would need to be strengthened in order to say this. We can, however, say that the passing scores are not unreasonable for students at the particular grade levels. That is, they are not set too high.

- ***Do the Pass+ Scores Represent "Exemplary Performance"?***

Our ability to answer this question definitively is again limited by the state's lack of specific definitions for what the term means at different grades and for different subject areas and, even more so than was the case for pass, by the nature of the tests. In order to make decisions about student performance, measurement experts try to design tests that provide a good deal of information in the region of the proposed passing scores. Indiana has done this quite well for the passing score in some areas. Unfortunately, the state has not succeeded in providing that same level of information for pass+ score in either subject.

Achieve and NCIEA found that none of the English/language arts or math tests has enough high-level items to justify a pass+ label. To be precise, no qualitatively different set of knowledge and skills are required of pass+ students than of pass students. Rather, the tests seem to "top out" without assessing the more advanced concepts and skills that we would expect exemplary students to have mastered.

The way to address this problem is to increase the cognitive demand of the assessments and then examine whether or not the more advanced cut scores truly reflect exemplary knowledge and skills. As the alignment study revealed, very few of the items on these tests require cognitively advanced thinking. To judge students as "exemplary," they must be judged against performances that require a command of more complex knowledge and skills. Until the tests are retooled, Indiana policymakers should decide on the percentage of students they think deserve this distinction and then set the pass+ scores to yield that percentage. This, of course, is pragmatic, not technical advice, and it should only apply to the transition period.

- ***Should the Cut Scores Be Changed?***

Although our analysis has pointed to several areas where the passing scores are not as definitive or informative as Indiana education leaders might like, we do not recommend changing them. In English, the pass score sets an appropriate bar for reading achievement. Although we recommend making writing more of a determinant, this cannot be done by simply changing the passing score. Similarly in math, we have made several recommendations for improving the assessments so they more effectively measure the essential knowledge and skills laid out in the standards. Raising or lowering the passing score on the test without attending to these design issues will not solve the problem.

The same thing is true with the pass+ level in both subjects. The best way to ensure that students earning that score are demonstrating "exemplary" performance is to increase the cognitive demand of the test questions so that they more effectively tap higher-level skills.

RECOMMENDATIONS

Based on these findings, Achieve recommends that Indiana take these steps to improve ISTEP+:

- √ ***Address the issues that suppress the rigor of the tests*** — Indiana should make its assessments more challenging by reducing the earlier-grade content included on the tests and measuring the more cognitively demanding standards and compound standards fully so that their more rigorous aspects are assessed.

- √ ***Use constructed-response items to their fullest*** — In order to measure the more challenging aspects of the state standards, Indiana will not only need to sustain its commitment to using open-ended items, but it will need to take better advantage of those items as well. In the secure technical report, we point out a variety of instances where these items are under-utilized, most notably on the math tests. At the same time, we point to the more challenging standards that are not being measured well on the tests. We encourage the state to redirect the open-ended items so that they tap the more challenging standards.

- √ ***Leave cut scores in place for now*** — In English/language arts, the passing scores differentiate among students, with "clearly passing" students demonstrating reading skills consistent with proficient readers. We do, however, recommend making writing more of a determinant. In math, the most significant issues are not byproducts of setting the passing score, but rather of test design and item writing. Raising or lowering the passing score on the math tests only would change how many of the same sorts of items students must answer correctly to pass (or demonstrate exemplary performance). It would not reveal qualitative differences in the knowledge and skills of non-passing, passing or pass+ students.

- √ ***Define "solid" and "exemplary" academic performance clearly*** — The most important attribute of all standard-setting methods, such as the bookmarking method that Indiana utilized, is that passing and pass+ scores need to be tied closely to proficiency descriptions that clearly define the knowledge and skills characteristic of students performing at each level. These definitions should be widely shared and agreed upon, and the state should produce evidence that these cut scores are reflective of them. Indiana does not currently have such definitions, and we would encourage the state to develop them.

- √ ***Revisit the "standard-setting" process only after the tests have been refined*** — While we are not recommending that the state alter its current passing scores, we do feel it is appropriate to revisit them once the tests have been refined to address the issues raised in this report. At that point, we recommend making the process as transparent as possible so that educators, parents and the public understand the level of performance that Indiana expects of its students.

CONCLUSION

In summary, Indiana is to be commended for subjecting its academic standards and assessments to such an intense level of scrutiny. The state has consistently demonstrated its commitment to quality in its approach to systemic educational improvement, and this has helped make Indiana one of the nation's leading education reform states.

In particular, Indiana has made enormous strides in improving the quality of its academic standards; they now rank among the best in the nation. If attention is paid to the issues raised in this report, Indiana can achieve equal success in improving the quality of its tests and the meaningfulness of its passing scores. Already, the state also has made a laudable effort to develop criterion-referenced tests that measure basic as well as advanced skills. Notably, the tests include constructed-response items and writing prompts — item formats that emphasize critical thinking and solving problems. By continuing the work that has already begun to strengthen ISTEP+, Indiana can become one of the best standards-based systems in the nation — and serve as an example for other states wrestling with these vital questions.

Board of Directors, Achieve, Inc.

Achieve, Inc.
www.achieve.org

400 North Capitol Street, NW
Suite 351
Washington, DC 20001
Phone: (202) 624-1460
Fax: (202) 624-1468