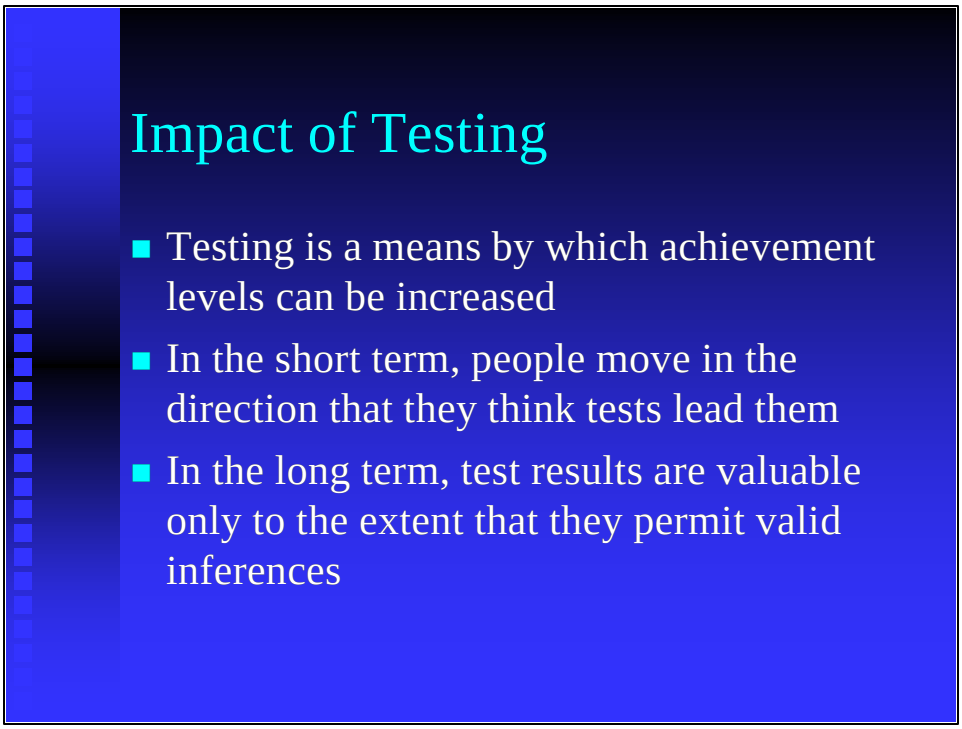
# Critical Design Issues in Assessment: A primer for policy makers

Richard Hill
National Center for the Improvement
of Educational Assessment, Inc.

1998 Policy and Practices Forum
Sponsored by the Education Commission of the States
October 12, 1998


The purpose of this presentation is to discuss issues that should be thought about when designing an assessment program. While some of the issues are technical in nature, most are issues that policy-makers can and should have some understanding of.

These seven topics are all issues that can be resolved in several different directions, and each choice can and should influence the design of an assessment. The rest of this presentation provides an overview of the issues for each of these topics in turn.

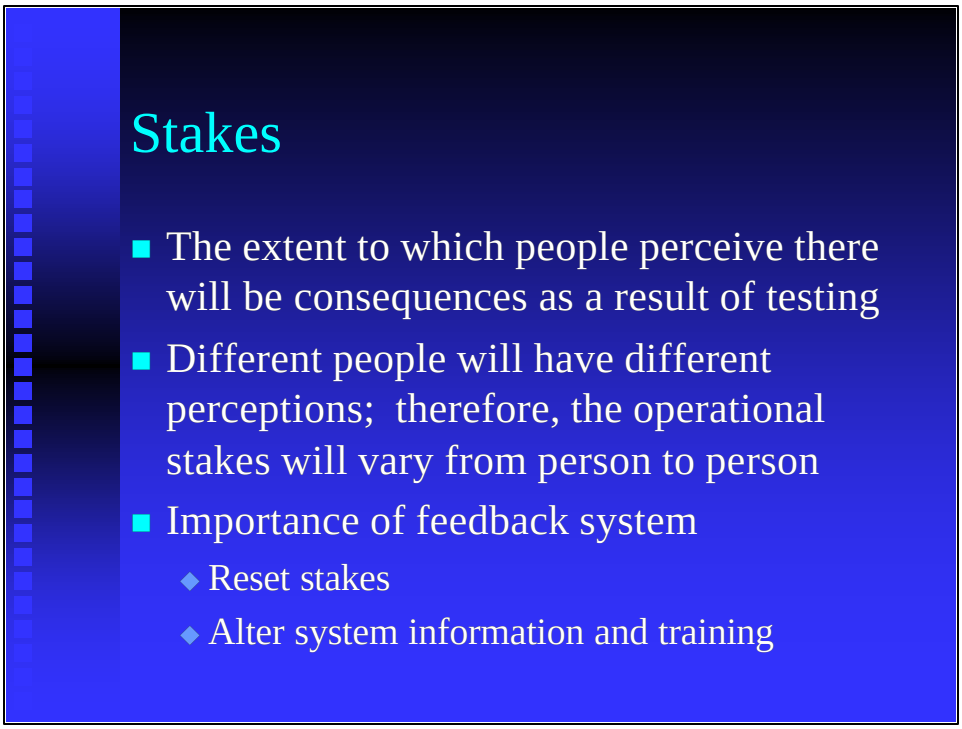



Policy-makers have become more interested in assessment programs over the past decade as a means to bring about improvement in the educational system. Testing can have this effect, although the answer to improved achievement does not reside alone in testing. But testing can play a significant role in a systemic effort at reform.

An important point to understand is that teachers will teach to what they think the test is. At the beginning of the testing program, a number of myths about what the new tests test, and how to best prepare students to take those tests, will emerge. Good designers will take advantage of this process to guide instruction even further in the intended direction than would be warranted by a close examination of the tests.

This advantage will last only a short while, however--perhaps a few years. After that, teachers will understand the tests more thoroughly and will have developed their own ideas on how to best prepare students for the tests.

People assume that test scores are valid. That is, that a group that scores higher than another in fact has higher achievement. There are many reasons why this might not be true--unreliability, inappropriate preparation, and inappropriate test-taking procedures are at least three. The higher the stakes, the more important it is that the results are valid without further interpretation--but the less likely that this will be the case.

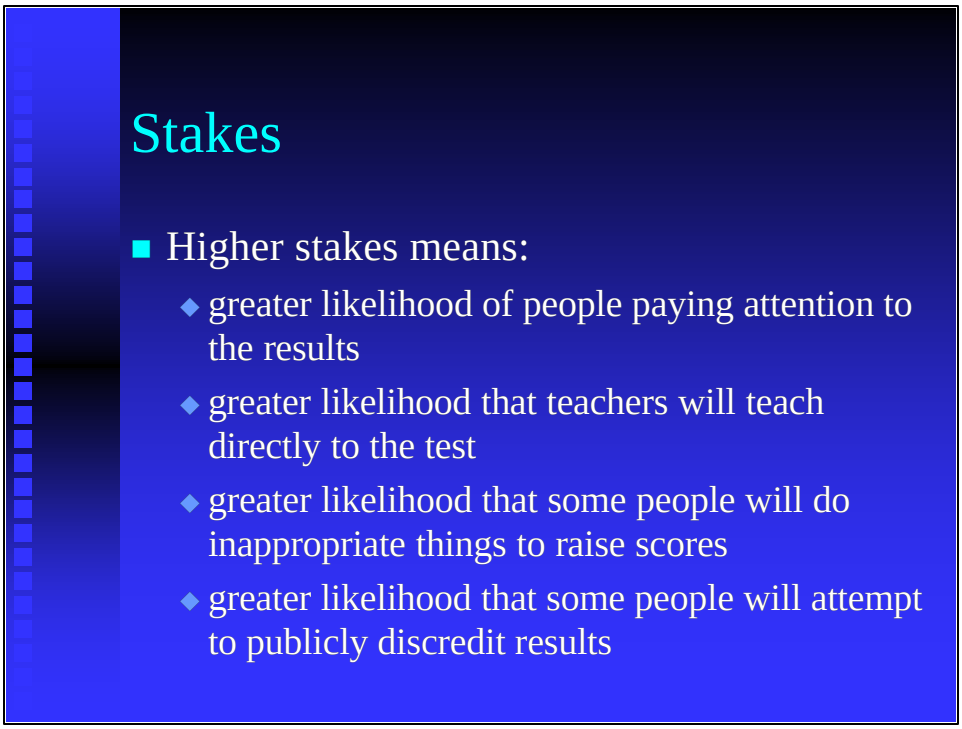



The point here is that stakes are what people perceive them to be--not what the policy-makers believe they are. And what one person sees as low stakes may very well be perceived as high stakes by someone else. Three main points need to be made:

1. As a group, teachers tend to be concerned about the impression their school has in the community. Thus, what most would see as fairly low stakes (the results will be printed in the newspaper) are generally seen as high stakes by teachers.

2. If students don't have stakes, they may not try their hardest (and certainly will be perceived by many teachers as not trying their hardest). Consequently, without student stakes, test results may be perceived to be invalid.

3. Because there will be a broad range of responses to stakes, there will be ample anecdotal evidence to support two positions: that the stakes are too high and also too low. As with many elements of a new testing program, it is critical to have a system of surveys in place to provide accurate and representative information that policy-makers will need to make modifications to the system. Also needed is feedback on what teachers are doing in response to the system, so that adjustments can be made in the information on student scores and training for teacher improvement that is being provided.




Stakes should not be chosen lightly. While raising stakes will have some positive benefits, there are significant costs to stakes--and the higher the stakes, the higher those costs will be. Therefore, it is imperative to choose stakes that get the appropriate attention, but no higher. In general, it is my opinion that stakes in large-scale assessment programs are set too high for teachers, that additional, externally-imposed stakes for elementary school children are largely unnecessary, and that there is no good answer to setting stakes for high school students--especially juniors and seniors. Again, this is why it is important to get accurate information on the effect of stakes. Perhaps more than any other item in an assessment design, it is critical to get this part of the program right--stakes that are too low will result in an ineffective (unnoticed) program; stakes that are too high will lead to results that cannot (and should not) be believed.


As noted in the previous slide, high stakes have costs to the validity of the program. But there are other costs that need to be recognized as well.

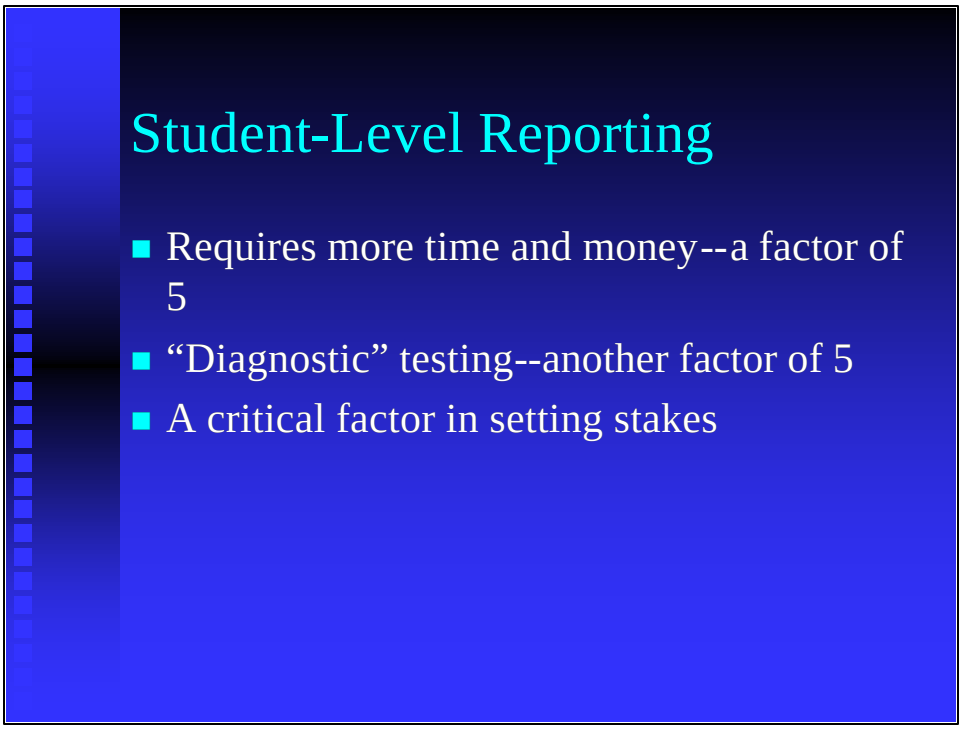
Programs with higher stakes need systems to ensure all rules are being followed. They need public relations efforts to ensure that misinformation is responded to. They need to ensure that results are not misapplied, and they need to investigate complaints of unfairness.

Tests need to be released after each administration. This promotes public dialogue about the tests, allows those affected by the stakes to thoroughly review their results, and helps prepare everyone for future editions of the test. This, of course, means that new questions must be developed for each administration, which must be equated to previous versions.

Because the program will come under greater scrutiny from the public, and often from the courts, all decisions must be documented to a greater degree.

Putting all this in place, and preparing both the educational community and the public for it, takes considerably more time than the typical preparation for a new testing program.

Finally, the higher the stakes, the less flexibility a state can have in the design of its program. A decade ago, when stakes were much lower than they are now, statewide testing programs varied widely. Today, with stakes generally much higher, programs look more alike.



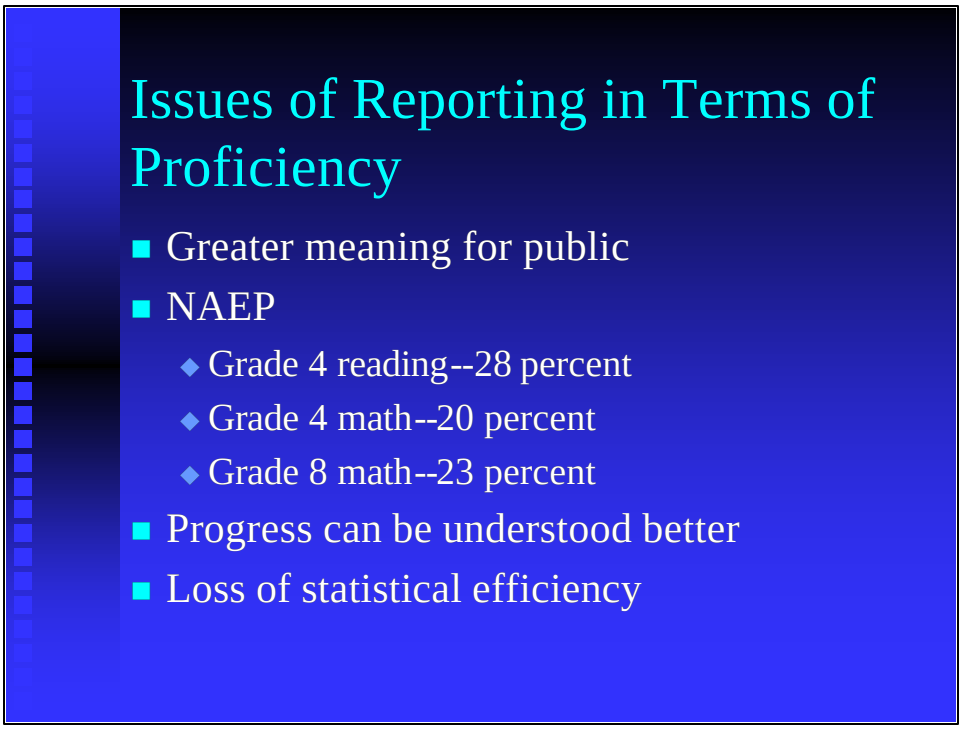

Choosing a level of reporting is a critical decision. A decade ago, many statewide testing programs were reported at the school level and not at the student level. For many good reasons--including the fair distribution of consequences for performance--there is more demand today that reporting be done at the student level.

This change increases the cost of assessment programs by a factor of 5 and increases the amount of testing time needed by a factor of 5. If tests were purely multiple-choice, the increase in cost might not be problematic; but when costs for assessment programs are going up dramatically anyway because of the addition of performance-based assessment, this additional cost can be a real concern.

This increase in cost and time is to obtain a general statement of performance level for students. Many people call for "diagnostic" testing of students, which usually makes no sense for statewide testing programs. The costs are far too great, and the information generally not as useful as necessary. First of all, unless the tests are computer-adaptive (impractical for current statewide testing), all students, even those who have already mastered an objective, need to take all items. This means that testing time will be largely wasted for many students. Second, for a test to be truly diagnostic, it needs to be generated directly from the curriculum to which results will be applied. Since there generally is no statewide curriculum, it usually would not be possible to construct a truly diagnostic test, even if the time and money were available. Finally, large-scale performance-based assessments take months to score. By the time results are returned to schools, they have little diagnostic value.


In the past, the primary reporting statistic was a school mean. Student scores were reported as percentile ranks or scaled scores.

For many good reasons, including vastly improved communication with the public, many statewide testing programs are moving to reporting performance relative to level of proficiency or standard. This is a change with several implications.

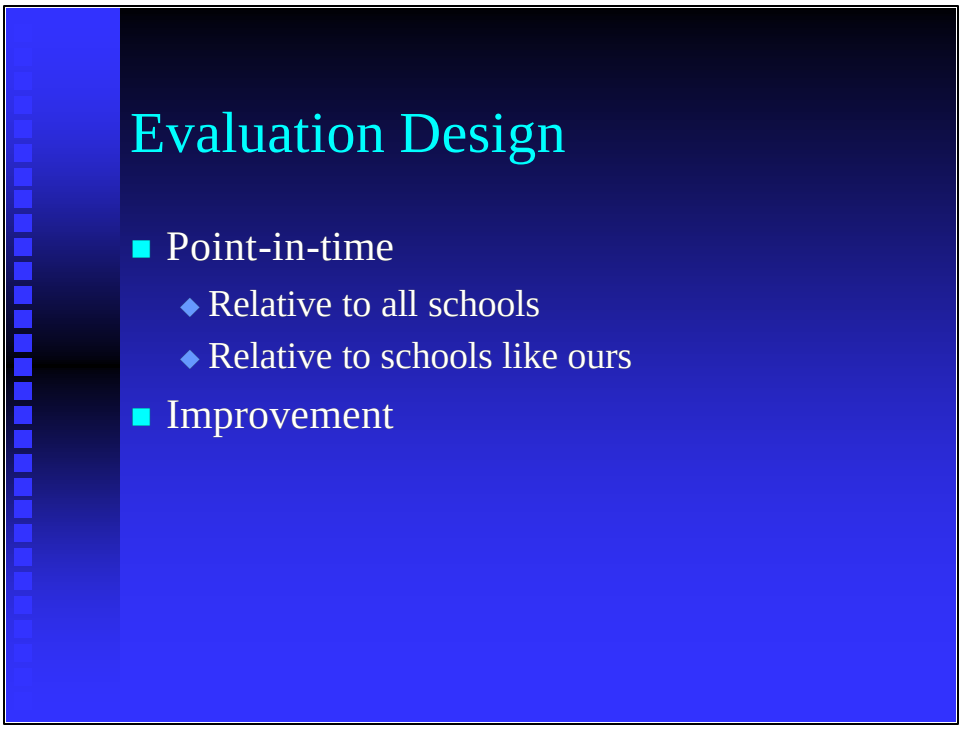



Reporting in terms of standards allows us to engage the public in a discussion of "how good is good enough." The materials from Arkansas demonstrate the power of using samples of student work to (1) communicate what the standard actually is, and (2) engage the public in the discussion.

Laying all the information bare to the public involves a major planning decision, however. NAEP standards are often talked about as "high" standards. In fact, when people get a chance to look at actual student work, they don't think that those standards are particularly high. Yet, only about a quarter of students nationwide meet even those standards. If the public is to be engaged in the discussion of what is acceptable work, much planning must be done to prepare them for the very high percentages of students who will not meet those standards, even in the best of states. In particular, policy makers must not paint themselves into a corner selling a program on the basis of high stakes for students when they will be unable to deliver on those promises when very high percentages of students fail to reach the desired standards. There are many issues that can be addressed on this topic, including maintaining constant standards and distinguishing between standards and consequences.

A great advantage to reporting in terms of standards is that progress can be understood much more readily by the public and others. To know, for example, that the percentage of Proficient students has increased from 15 to 30 percent means much more to people than knowing that the average percentile rank of test scores has gone up 10 points.

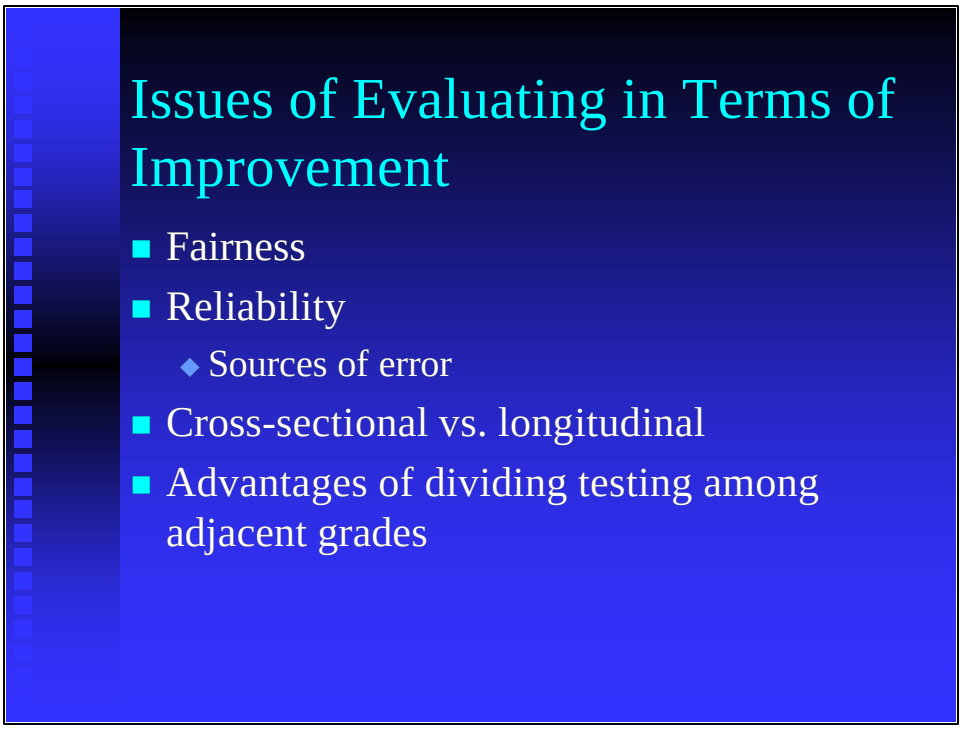
Although it is a technical issue, policy makers should be aware that multi-level accountability indices have only about 70 percent of the efficiency of means--and pass/fail indices have far less efficiency than that.




In the past, statewide testing programs generally evaluated schools on the basis of how they did in one particular year, or their average over a few years. Since quality of teaching is only one factor in the scores schools attain, such an evaluation system can be unfair. Low scoring schools are not necessarily doing a poor job of teaching, and high scoring schools are not necessarily doing a good job--a school's score is significantly affected by the achievement levels of the students before they start school and by the quality of their lives at home.

To improve on this system, assessment programs began to develop statistics that allowed schools to compare themselves to "schools like theirs." While an improvement on the old system, there were many reasons why this system was considered unacceptable. For one thing, it often took on the appearance of "excuse-making;" i.e., that low scoring schools from poor areas were doing OK, since they were doing at least as well as other poor schools.

The current trend is to hold all students and all schools to one standard, but to make the system more fair by evaluating schools on the basis of progress.

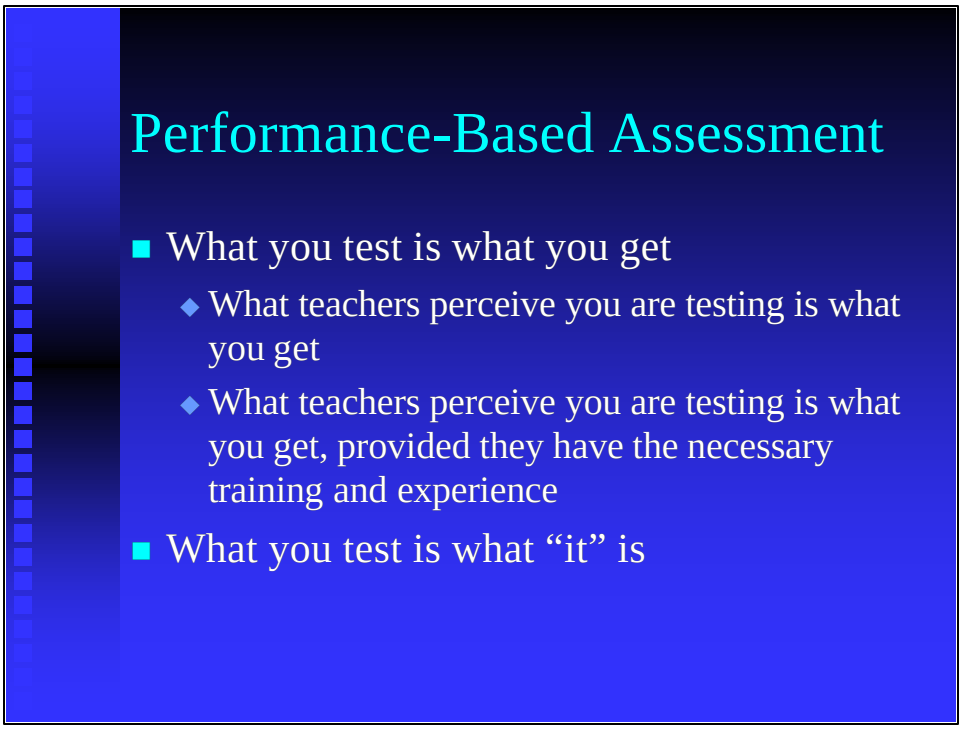



Such a system is probably more fair, but progress is more difficult to detect than point-in-time status, since there is error associated with both the pre-test and the post-test, and errors are additive.

An important point to understand is that uncertainty in a school's score comes from many sources, but 80 percent of the uncertainly in a typical school comes from the fact that students change from year to year, and a school may well be judged as improving or declining simply because the students have changed from one year to the next.

Some have called for a longitudinal design--for example, testing the third graders one year and the fourth graders the next. This usually is not the answer. There is no way to assure the comparability of standards from one year to the next, schools are not held accountable for students who move between pre-test and post-test, and testing costs are doubled.

A design that is just beginning to be thought of and implemented involves dividing the testing across grades--say, giving the reading and math in grade 4 and the science and social studies in grade 5. This is a model that dramatically reduces uncertainty, has trivial increases in cost over single-grade designs, and divides testing burden and accountability over more grades. This is an idea that designers should keep in mind.

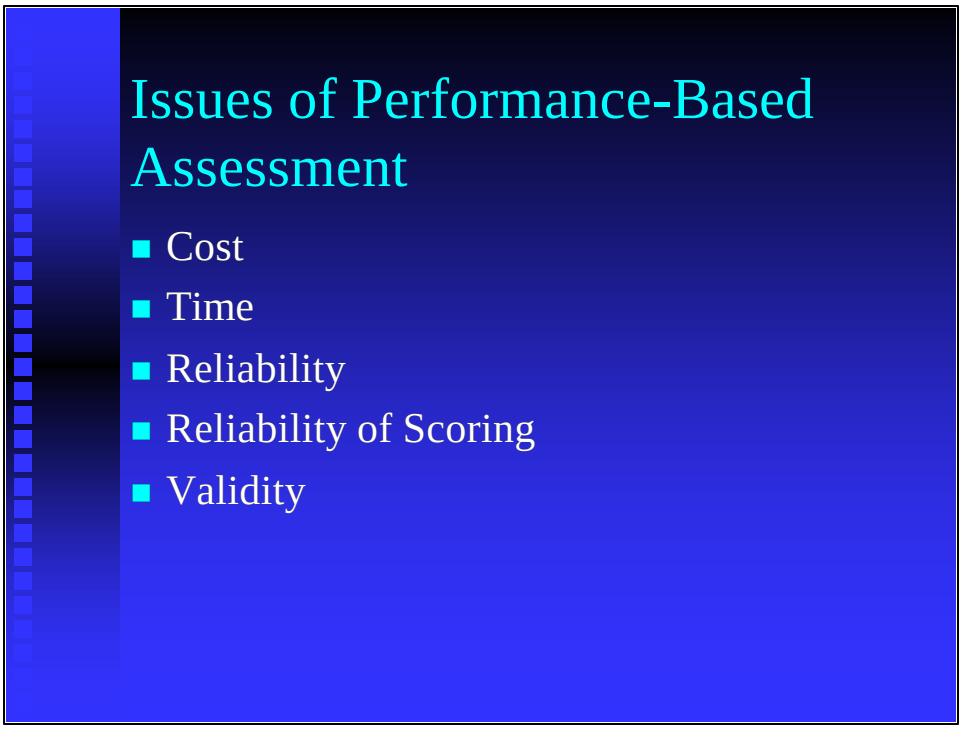



Lauren Resnick made the phrases "What you test is what you get," and "What you don't test is what you don't get" popular. They are accurate reflections of the point of view that what you include in a statewide testing program will attract the attention of teachers--and the higher the stakes, the more attention you will attract.

Experience has shown that this is not the whole story, however. As noted earlier, teachers develop their own mythology about what they need to teach in order to have their students score well on the tests. Thus, what they think the test is about is what they will teach.

And finally, we often overestimate the capacity of teachers to change, even when they understand that they should. If they don't have the training and experience to change, their attempts to do so will be unsuccessful. It is when this frustration is combined with a high stakes testing situation that the worst examples of inappropriate behavior occur.

Test operationally define the content standards. Teachers and others often don't understand what they are supposed to be teaching just from reading content standards. But when they see the test, it all becomes much clearer. There are several implications to this, such as the importance of having the tests reflect the instruction desired, and the need to have tests out in the public for some time before holding people accountable for improving scores on them.




Given the potential that statewide tests have to influence instruction, most states are including at least some performance-based assessment tasks in their tests. This is not done without incurring additional costs in a variety of ways.

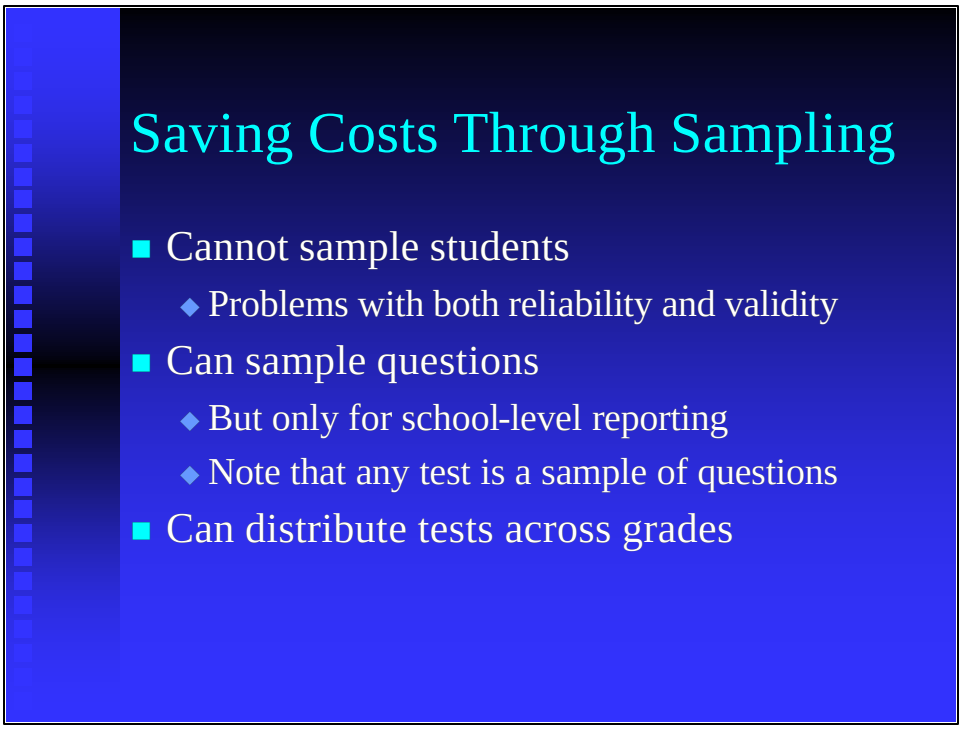
The first obvious cost is the cost of task development and scoring. Both are significantly higher than similar costs for multiple-choice tests.

The good news about open-response questions is that they provide as much information as 3-4 typical multiple-choice questions. The bad news is that a single open-response question takes over 10 times as long to answer as a typical multiple-choice question. As a result, if an open-response test is constructed to have the same reliability as a multiple-choice test, its administration time will about about 3 times as long. That usually is a workable number, but is a significant increase in testing time.

People generally assume that the talk they have heard about the inadequate reliability of accountability systems (e.g., Kentucky) is due to the unreliability of scoring. That is not true. Scoring (actually, coding) of open-response tests is quite accurate. Unreliability, where it occurs, is primarily due to the sampling of students, and that will be true whether the test is multiple-choice or open-response.

Such talk also ignores the considerable sources of unreliability associated with multiple-choice questions not present in performance-based assessment--the most obvious example of which is guessing. The probability that a student will correctly guess the answer to a multiple-choice question is far higher than the probability that a scorer will miscode a student's response to an open-response question.

As noted in the previous slide, tests must be true or valid reflections of desired outcomes if teachers are to change instruction in the intended direction.

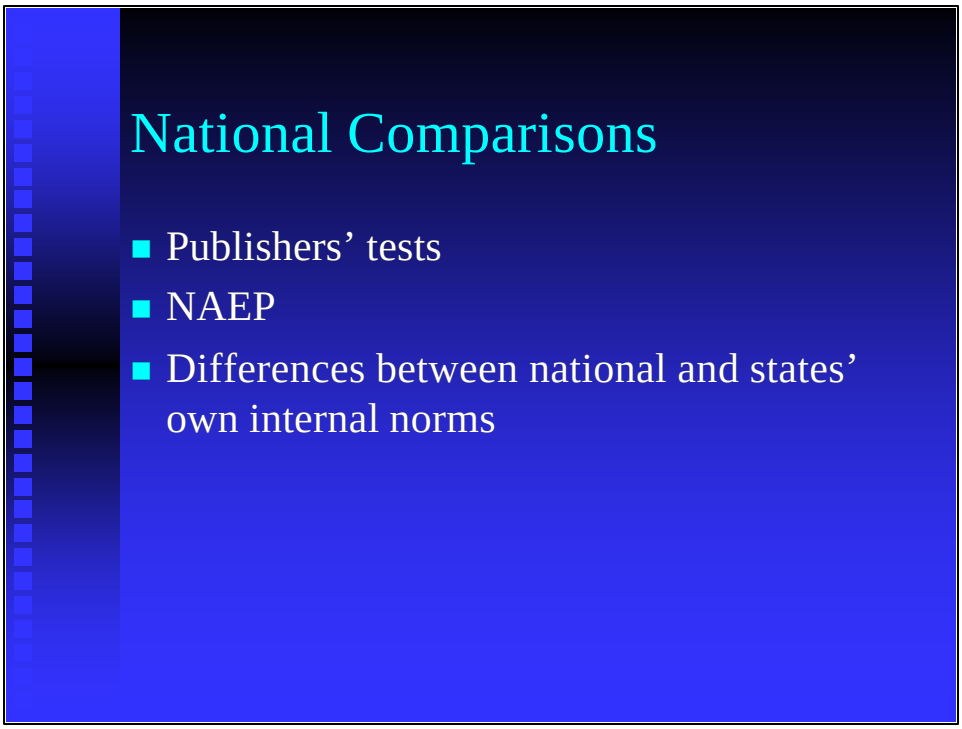



Given the increased costs of a performance testing program, it is good to ask how those costs might be reduced. In the past, sampling has been an answer for some programs.

One cannot effectively sample students within schools for a high-stakes program. Given that sampling of students is the major source of error in school-level data, one must test all the students available. In addition, with a high-stakes testing program, one cannot assume that the sample any school would provide would be representative.

Sampling questions is a good way to reduce time spent in testing, but this works only for school-level data. When students are the unit of analysis, it is very beneficial to have all students take the same questions. An approach that Maine and Kentucky have used is to administer a set of questions in common across all students, and then to supplement those questions with additional matrix-sampled questions. Remember that any test is a sample taken from the domain of all possible questions--so the issue isn't whether sampling is being done, but rather, what type of sampling makes the most sense.

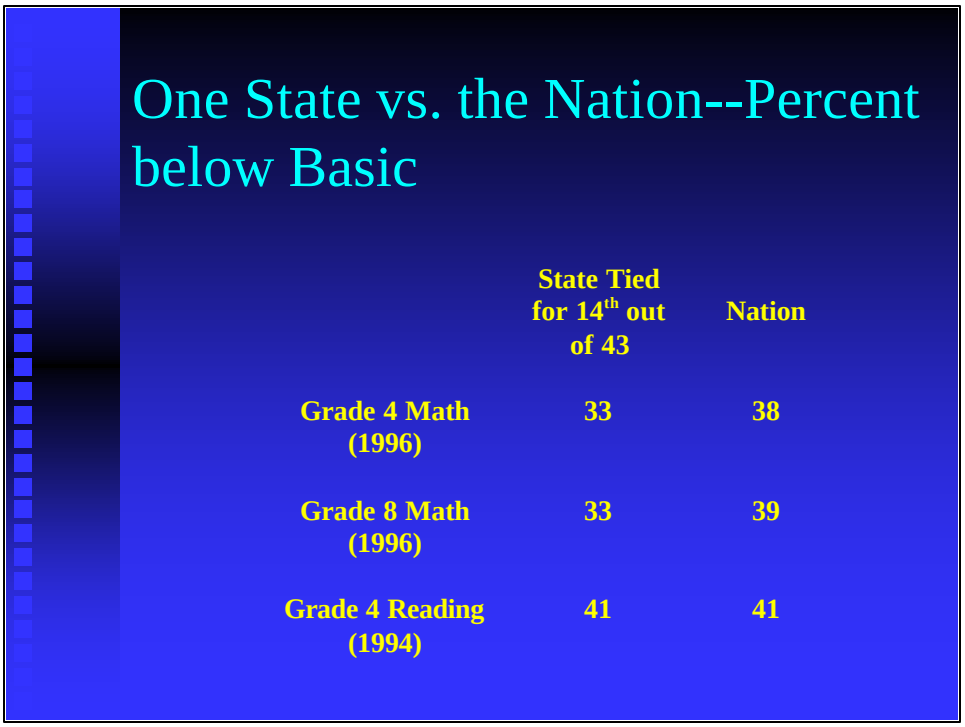
As noted in an earlier slide, distributing tests across grades is an excellent way of distributing accountability, reducing the amount of time any one student is tested in a given year, and dramatically reducing the error due to sampling students.




Today, most states are calling for national comparisons for their tests. Some are retreating to publishers' norm-referenced tests in order to obtain those norms. The norms provided by publishers' NRTs have been known to be greatly inflated in the past, and there is no assurance that the norms provided with the latest versions of publishers' tests are any better.

Accurate national norms are available through NAEP. In fact, the most accurate comparisons of states to the nation as a whole are available from the Trial State Assessments that NAEP conducted in 1992, 1994 and 1996.

An important issue to be addressed about "national comparisons" is what it is we would like our students to be compared against. When different state tests are based on different content specifications, and there are different levels of motivation to perform well on the test, comparisons may be meaningless. For example, if a state selects an NRT to make its national comparisons, but then establishes high stakes for performance and uses the same edition of the test year after year, the "national comparisons" it obtains have little relationship to how their students would do compared to the peers nationwide if another test were administered under different conditions.

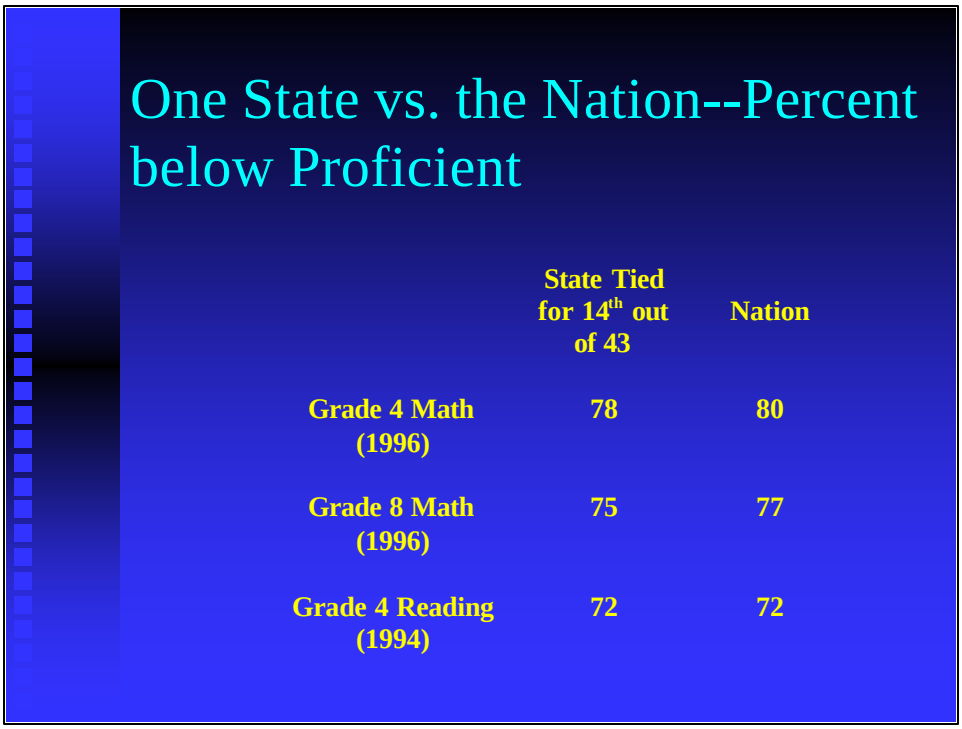



These data are for a state that tied for 14th out of 43 states that participated in the 1996 NAEP Trial State Assessment in mathematics at grade 4. In this state, 33 percent of the students scored below Basic on grade 4 math in the 1996 Trial State Assessment; the comparable statistic for the country as a whole was the 38 percent. Thus, if in-state norms were used instead of national norms, a student who was reported at the 33rd percentile actually would have scored at the 38th percentile if national norms had been used.

Suppose a student scored at the 50th percentile in this state and we told his/her parents that, and suppose further that we even left the impression that those were national norms. We would not be telling the whole truth--in fact, the student would be scoring above the 50th percentile when compared to students nationwide. But how much above the 50th percentile? Something less than 10 points; and probably something closer to 5 or 6.

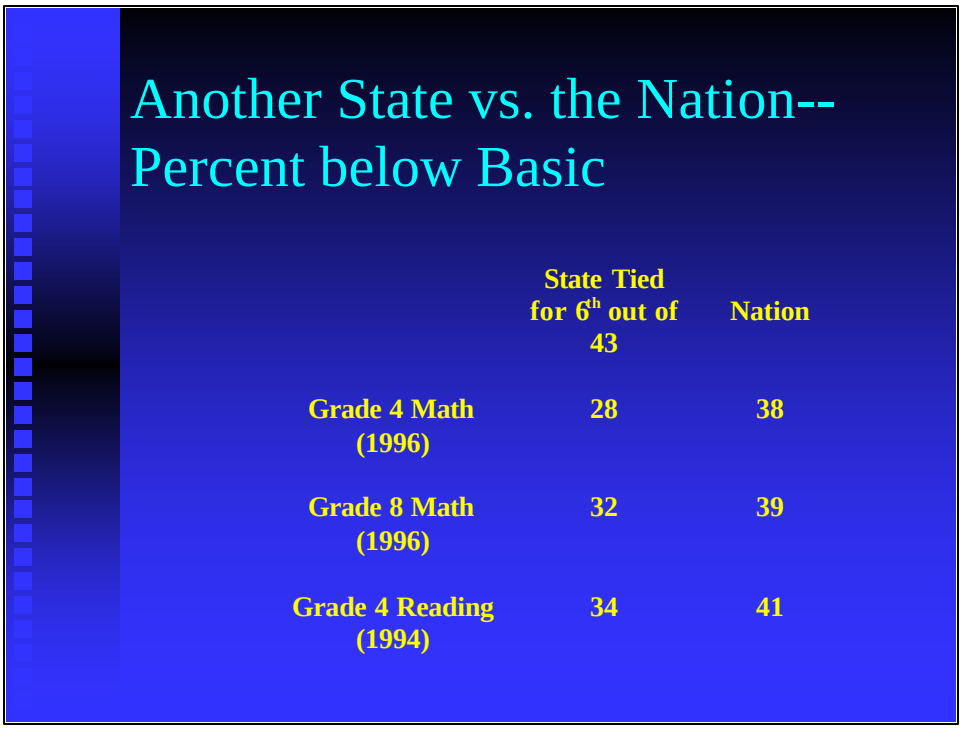
Now here is the big question. If we told a set of parents that their child was scoring at the 50th percentile when the child really was at the 57th percentile, how much of a difference would that make? Would anyone really make a different operational decision, on the basis of one test, on the basis of 7 percentile points? Hopefully not, since the standard error of measurement on the test is likely to be far higher than that.

If it doesn't make sense to change decisions on the basis of a few percentile points, why spend the money and change the design of the assessment to collect the data? Why not report highly accurate state percentile ranks (available for free) instead of national percentile ranks with doubtful accuracy that come at great cost, and are likely to be different only in trivial ways from the state percentile ranks?



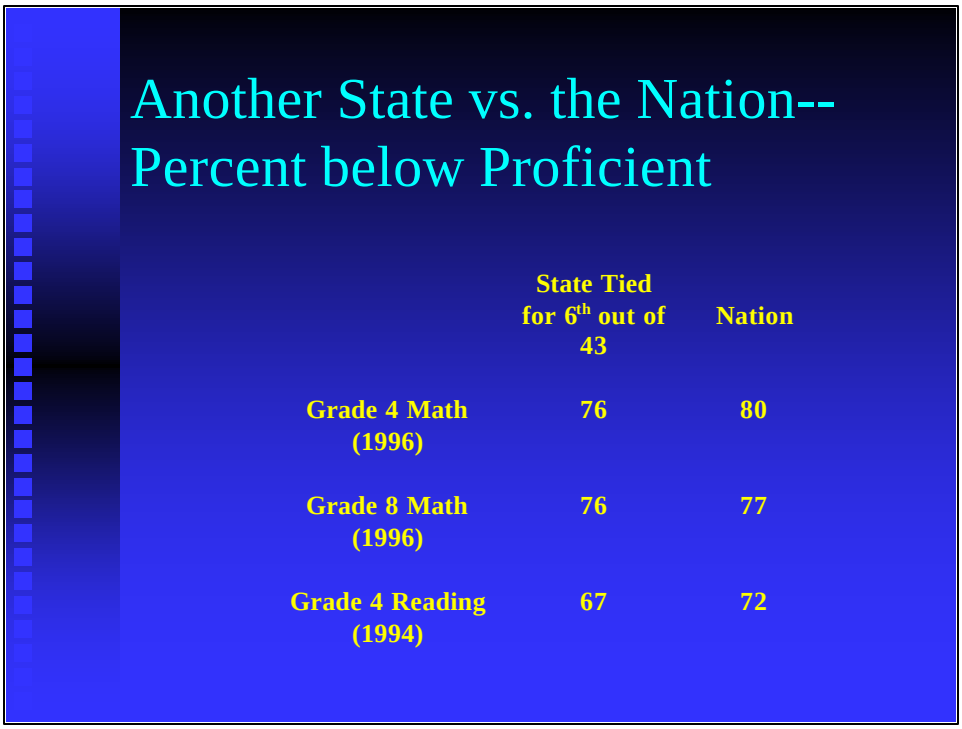

Let's examine how representative these results might be. Given, this is a state that is somewhat above national averages, but not much, since we know that overall results show that it is tied for 14th out of 43 states.

First, does the procedure that worked at the bottom of the distribution work at the top, too? Yes, and in this case, even better. When looking at the upper end of the distribution (percentage of students below Proficient), the percentile ranks we would obtain from state norms differ by no more than 2 percentage points from national norms--surely a trivial difference by any measure.

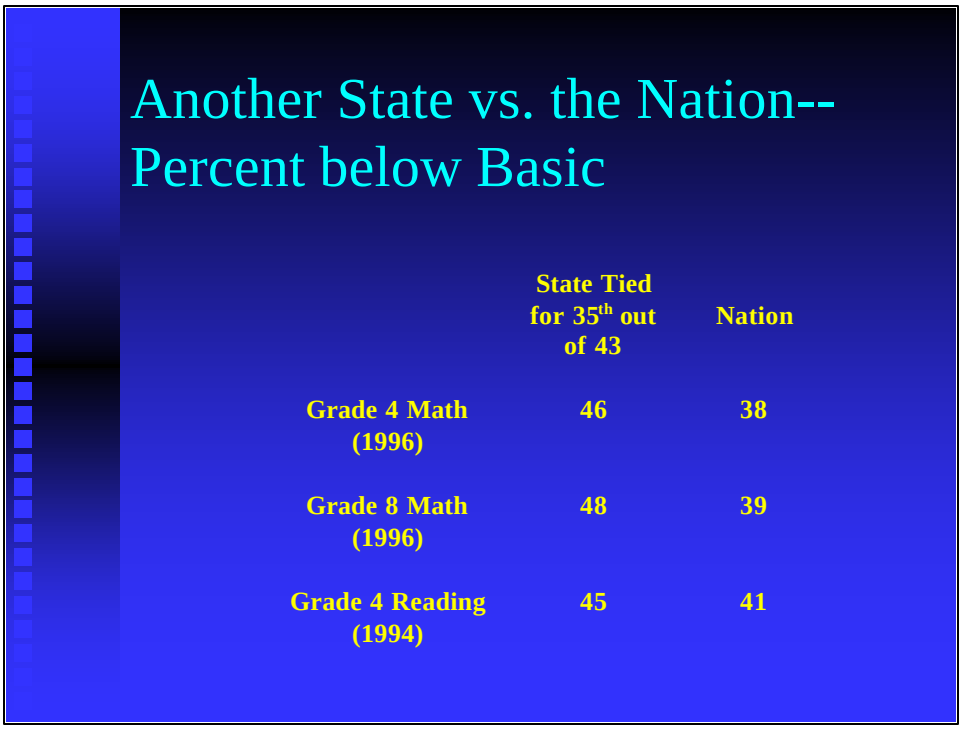



So, then what would be the case if we looked at a state that was much higher scoring? Here are the data for a high-scoring state--its overall results on grade 4 math made it the 6th highest scoring state in the nation (tied for 6th).

In this case, the state-level percentile ranks for students at the low end of the distribution are 7-10 points below comparable national-level percentile ranks. Again, one must ask the question of whether that information (1) is less accurate than similar data obtained from publishers' "national" norms, and (2) if so, whether the amount of error would lead people to make incorrect conclusions. Given that the error would be pessimistic (that is, students actually would be scoring better relative to national norms than to state norms), probably not.

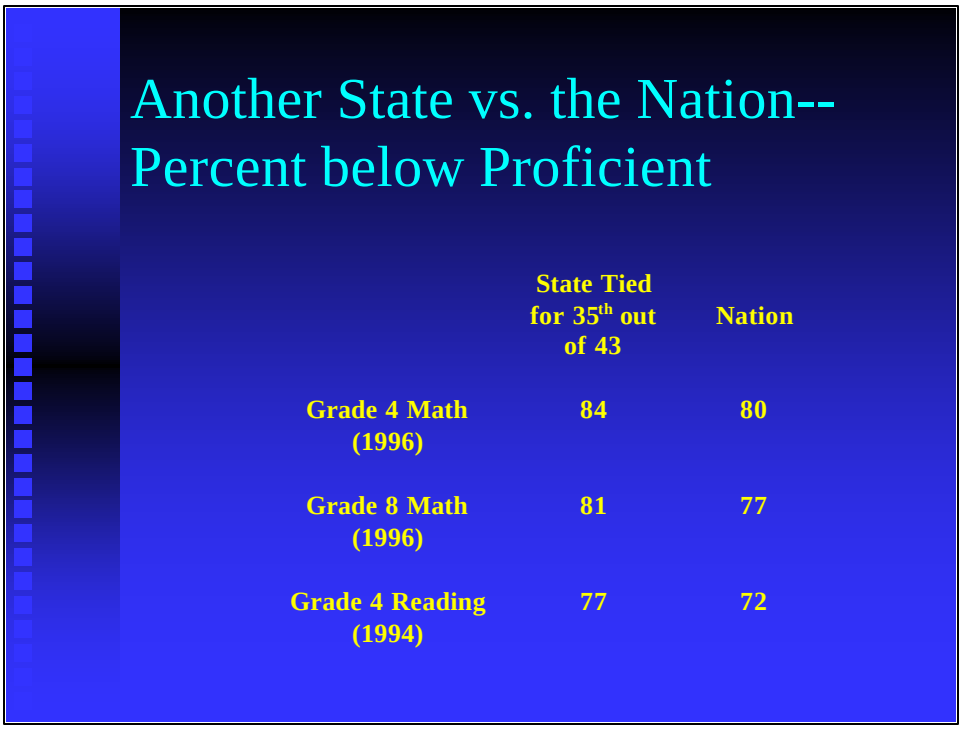



As was true for the first state, the problem is smaller at the upper end of the distribution. For the same state in the previous slide, the difference between state- and national-level norms is no more than 5 percentile points for students at the border between Basic and Proficient.

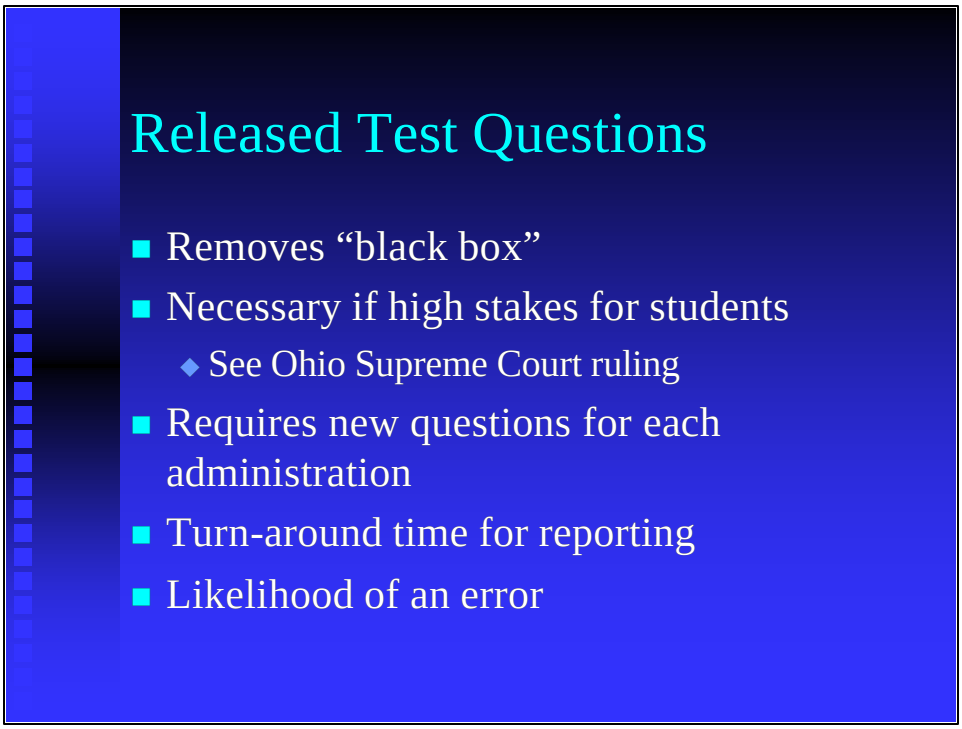



Alternatively, what would happen if we looked at a low-scoring state? Here are the data for a state that had a average scaled score higher than only six other states included in the grade 4 math TSA in 1996 (three states tied for 35th). Its state-level percentile ranks are 4 to 9 points higher than national percentile ranks. Thus, for example, a student who scored at the 46 percentile in this state in grade 4 math actually was only at the 38th percentile of national norms. Again, the questions must be asked of how much larger (or smaller) this error is than the errors that have been reported to parents on the basis of publishers' norms for years, and what the consequences would be for an error of this magnitude.

If these errors are too large, they could be reduced significantly by obtaining norms for ethnic and socio-economic groups and adjusting the state's data by reweighting these categories to national averages. But as can be seen from the data, such adjustment is unnecessary for the vast majority of states.




Again, the errors are smaller for the upper end of the distribution. Although this is one of the lower scoring states in the nation, its state-level percentile ranks in the upper end are different from national-level percentile ranks by only 4-5 points.

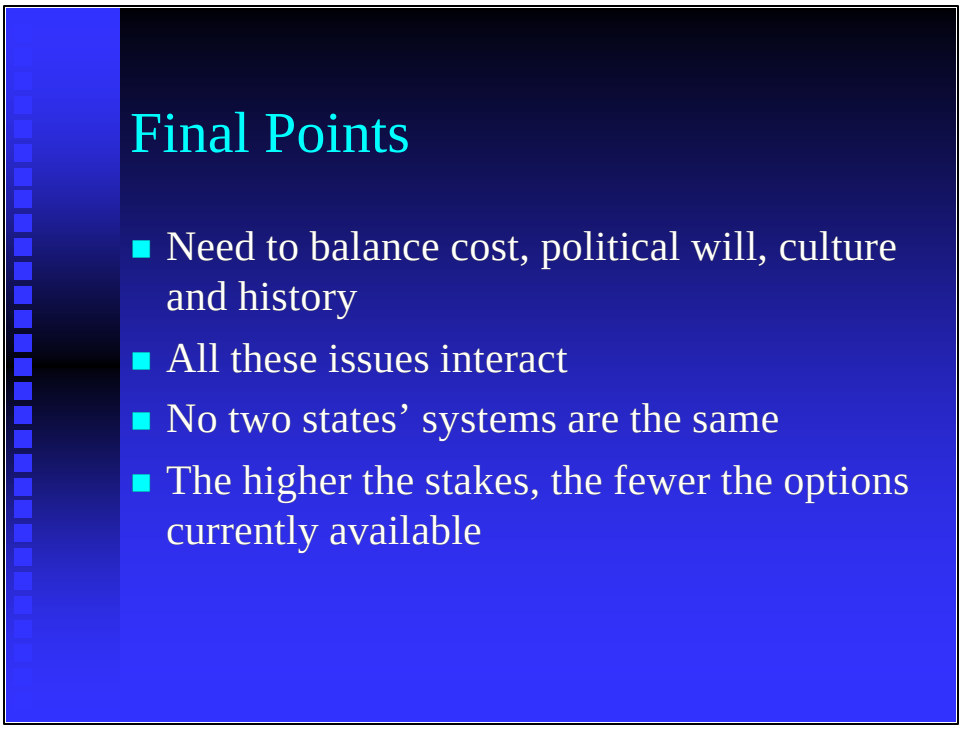



If a test is going to be used for high stakes at the student level, the questions used to compute student scores must be released at the conclusion of the administration. If the test is not going to be used for high stakes at the student level, much of the test should be released at the conclusion of the administration.

Release of test questions is an excellent way to open the dialogue with the public about what the test was designed to measure and what it is students are expected to learn. Experience in other states has shown that opponents of testing will create stories about what the test questions are. The most effective way to combat this attack is to provide the public with at least a healthy sampling of the questions. A caution here is that when the items are made public, detractors of the program may selectively point to questions that they believe should not have been included in the test. If the test is well-constructed, there should be a balance of items so that these arguments can be effectively refuted. But having the public see the balance is dependent on effective communication of the entire test.

Releasing questions, as is true for everything else discussed today, is not without its costs. The costs of developing new questions for each administration is an obvious cost. A less obvious cost is the increase in the amount of time it takes to produce reports, which is confounded with the likelihood of introducing an error. The more that things change from administration to administration, the more custom computer programming that has to be done, and the greater the likelihood of human error. If sufficient time is built into the reporting schedule, that likelihood can be minimized, but never reduced to zero.

It should be noted that NRTs typically do not have enough forms so that items can be released, so one clear advantage of a state-developed test is the ability to release questions at the state's desire.




What is the best assessment and accountability design for a state? There are as many answers as there are states. All of the issues discussed today interact, and each has trade-offs that must be evaluated in terms of the costs that a state is willing to invest in its program. The optimal investment is influenced by the demand for change within the state, the will of its political leaders to invest in such change, and the culture and history that brings the state to its current decision point.

For this reason, no two states' testing programs are the same. However, while there was great variation in testing programs in the past, that variation has diminished somewhat. With higher stakes, there is more of a requirement to include performance-based questions in the assessment. Costs and current restrictions of technology then limit options, so as stakes have increased, programs have become more similar. As technology improves, so should the flexibility in the design of programs.