Comparability of Assessment Results across Years[1]

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

Paper presented at
The Large Scale Assessment Conference
Council of Chief State School Officers
Nashville, TN

June 18, 2007

When state test scores are released each year, improving scores typically are attributed to some increasing effectiveness in the educational system, while decreasing scores typically lead to some finger pointing that presumes that the educational system has deteriorated over the past year. The purpose of this paper is to outline a series of issues and related data analyses that should be run to eliminate alternative explanations for the increase or decrease in test scores. This paper is not going to address issues of operational errors or blatant cheating, either of which could be a reason for test scores changing over years; those topics would merit papers of their own. But even when there are no operational errors or differential cheating across years, there are a myriad of reasons why test scores could change.

In order to compare test results across years, an *equivalent test* must be given to an *equivalent population* under *equivalent conditions* with *equivalent scoring*. This paper is organized to explore each of those topics in turn.

Equivalent Test

By an equivalent test, we mean a set of test questions that when administered in two different years, will yield the same ability estimate for students with the same level of achievement. While this topic will be discussed more fully under "Equivalent Conditions," it should be noted that administering the same test questions in two different years cannot qualify as an "equivalent test," since the exposure of the questions during the first year's administration makes the test a different one the second year. Therefore, states typically conduct some form of equating to facilitate comparisons across years, with the majority of test questions being different from administration to administration.

An initial requirement for two years' tests to be equivalent is that they have been drawn from the same framework and are designed to match the same test blueprint. States should, and typically do, routinely document in their technical manuals that this has been done. Unless there has been a planned change in the test design, states generally do have equivalent tests across years.

If there has been a planned change in the test design, however, changes in test scores must be interpreted with caution, at best. Often, the change in the tests is significant enough that results

[1] This paper is a draft. Comments and suggestions for revision are welcome, and should be e-mailed to Richard Hill at rhill@nciea.org. The paper will be revised over the summer; the revised paper will be posted on the NCIEA website in the fall.

cannot be compared across years. In that case, the trend line associated with the old test must be discontinued and a new one started.

Often, it is not a clear decision whether the test has changed enough to warrant discontinuation of the trend line. One data analysis that we have found helpful is to compute the average correlation within and across three sets of items: the old items that will not be on the new test, the new items that would not have been on the old test, and items that are used both years. So, for example, suppose last year's test consisted of 50 items, 20 of which measure skills not eligible to be included in this year's test (call that "Set A"). Suppose further that this year's test also consists of 50 questions, 30 of which are carried over from last year's test (Set B) and 20 of which are new (not just different questions from last year, but questions that are drawn from a different framework and would not have been considered for inclusion in last year's test—Set C). There are 95 unique pairs of correlations that could be computed among the items within Set A and within Set C, and 435 within Set B. In addition, there are 600 unique pairs of correlations that could be computed between the items in Set A and Set B, and the same number for Set B and Set C. No correlations can be computed among the items in Set A and those in Set C, because the two sets of items were administered to two different sets of students.

The first step would be to compute the 95 + 95 + 435 + 600 + 600 correlations. Then, each correlation should be converted to a Fisher's *z* by the transformation, $z = .5[\ln(1+r) - \ln(1-r)]$. This puts the correlations on an interval scale which in turn permits the means to be appropriately calculated. Now, compute the means and standard deviations of the *z* scores for these five groups of correlations. If the means for all five sets of correlations are the same, then there is evidence that the construct of the test has remained essentially unchanged. However, if the correlations of the items within each set are considerably higher than the correlations across the sets (and especially if the difference between A and B is different from the difference between B and C), there is strong evidence that the construct of the test has changed across years and that scores between the two years cannot be compared. Interpretation of these differences will be facilitated by dividing them by the standard deviations of the correlations within each set in order to determine an effect size.

But short of a major change in the design of a test, the greatest obstacle to creating tests whose scores are equivalent across years is equating. Typically, states either "pre-equate," in which items are drawn from a pool among which all of the relationships were established from an initial field test, or they "common-item" equate by administering one set of questions to students in two consecutive years. On its surface, pre-equating has fewer problems associated with making comparisons across years, because all the relationships among the items are established at one time. However, those relationships often change across years as teachers and students become comfortable with the frameworks used to generate the item pool, so most states use a common-item equating design.

There are several issues to be careful about when using common-item equating. First, it must be noted that in this design, *all* the change to be attributed across years comes from these items. If performance on the equating items changes, that same change will be statistically applied to all the other items. Thus, it is crucial that one has confidence interpreting the changes due to those equating items. There must be enough of them so that the standard error of the equating is relatively small, and they must be representative of the entire item set. Note that one way of determining their representativeness is to conduct the analysis of means of item-intercorrelations that was outlined above. One way this need for representativeness often is violated is by conducting equating on multiple-choice questions only and eliminating all constructed-response questions from the equating pool. When this happens, any change that takes place in student achievement relative to the

constructed-response questions is “equated out”—performance on these items has no impact on the change calculated across years.

Even if the items chosen for the equating set are representative of the items as a whole, they must be administered in a way that ensures the conditions for their administration are comparable. This especially means that the position of each item relative to the entire test administration must be equivalent, but it also means that the security of the equating items must have been maintained across the two administrations and that no event occurred during the year that led to changing the difficulty of the equating items relative to each other or to the other items in the test.

A responsible equating process will check to see whether student performance has changed differentially on a particular item or on a subset of items. If so, that item often is eliminated from the equating set before the final equating is run. But this practice should be dependent on the reasons for the relative change in the item—that is, whether it can be inferred that this change occurred in other items within the domain of items that could be assessed (in which case the item should be left in the equating set) or the change is idiosyncratic to this particular item (in which case the item should not be used for equating).

An additional condition that must be met to ensure the equivalence of the equating items across years is their position within the test. Any change in the administration conditions of the equating items, such as placing them earlier or later in a session, changing the session within the entire administration window, or placing easier or harder items in front of them can lead to changes in the difficulty of those items, which in turn leads to erroneous conclusions from the equating. That is, if a set of equating items is made more difficult in the second year of their administration by moving them later in the test booklet or by replacing the items preceding them in the test booklet with more difficult test questions, student performance will appear to have declined from one year to the next when in fact that might not be the case.

Even if all the equating is done properly (so that the mean estimate of change is exactly correct), there are at least two sources of random variation that may make the results of a year appear to go up or down a misleading amount. First, the equating items in both years are taken by a sample of students. Second, the equating items are only a sample (and often a fairly small one) of all the items that could have been administered across the years. Third, there is a degree of granularity in the test results because there is a finite number of possible raw scores that students could attain on a test. That is, suppose the cut score between Basic and Proficient one year is 42.0 raw score points. The next year, the test is slightly harder, and the equated cut score is 41.5 raw score points. Since students can only get 41 or 42 points on the test, but not 41.5, the score chosen for the cut point the second year is 42—the same as the first year—but the test is slightly harder. In this case, if actual student achievement is identical across the two years, performance will appear to have declined somewhat.

As a result of these issues, states should always calculate the amount of random variation due to equating that is present in their results, and interpret changes (whether up or down) as non-consequential when the amount of change is less than the amount of random variation that one might reasonably expect to find from year to year. The Center worked with a state that had experienced a decline in test scores from one year to the next after several years of reported increases; our original charge was to determine whether there had been an equating error. We found that the equating process had been sound, but that the amount of decline had been less than the error we would have expected from the equating process. However, that statement also could have been (and should have

been) applied to the positive results from the previous several years. Given that the state had not informed the public about the amount of error involved in the equating when the results had been positive, it became difficult to include that information the one year the results had been negative. What the state should have done was track the changes over years and shown that the total amount of change over a considerable period of time (say, four or five years) was significantly more than the random variation due to equating. Then, when the results had declined in one particular year, it would have been easier to attribute that change to equating error.

In addition, it generally is assumed that all students are taking the same test. However, in recent years, it has become more typical for a statistic such as "percentage of students Proficient or higher" to be an amalgam of results across different tests. The results of students taking an alternative assessment, for example, often are integrated into the statewide results as if they were equivalent to the results for all other students. However, if the rules associated with scoring those tests or the percentage of total tests those alternative tests represent have changed across the years, that would have an impact on the results that should be noted. If, for example, all students take the regular test one year, but the next year an alternative assessment is in place and some percentage of students take that test instead, the results across the years will not be directly comparable.

Equivalent Population

Even if the equating of the tests across years could be done without error, changes in performance could well be due to changes in the population taking the test rather than changes in the academic achievement of an equivalent population. Students move in and out of a state during the year; if those moving in are higher achieving than those moving out, scores will go up if the educational system remains unchanged. If the inclusion rules are changed across two administrations, results might change as a consequence. A change in the students present for testing (say, due to differential dropout rates, illness or some dramatic event that causes more students to be present or absent) also could lead to significant changes in test results.

Unless the test is taken by an entire population of students, changes in the tested population across years are almost a certainty. For years it has been known that comparisons of SAT or ACT scores across years are confounded by changes in the population of students taking those tests. If more students in a state take the SAT, for example, it is likely that scores will decline, since the additional students taking the test often have lower levels of achievement than the more select subset of students taking it in previous years. This phenomenon has not applied to most state tests in the past, since the tests are expected to be taken by all students, but likely will become an issue in the near future as states replace their existing testing programs with end-of-course examinations.

Another related phenomenon has taken place in recent years. Some states had fairly relaxed inclusion rules—when a local school or district felt that a student should not be tested, they were allowed to exclude that student without penalty. In some states, over 10 percent of the students were being excluded in statewide results. That percentage has dropped drastically in those states with the introduction of the No Child Left Behind inclusion requirements. One would expect that testing a more inclusive population would cause results to go down, since students excluded from testing usually perform below average on those tests when they are required to take them.

In a similar vein, one must be especially careful in comparing the results of high school students across years. Test results typically reflect the scores only of those who are present to take them.

Given that some students drop out of high school (and therefore are not present to take the tests), a change in dropout rates could easily cause non-comparability of tested populations across years.

Equivalent Conditions

For results across years to be truly comparable, the conditions under which the tests are administered must also remain unchanged. Some of the issues have already been mentioned, such as noting that administering the same test questions across time does *not* mean that results are comparable, particularly if some of the items have been exposed. However, there are many more changes that might take place from year to year that would limit the comparability of results.

When the stakes associated with a test change, it is almost certain that the conditions under which students are taking the test will change as well. It is often observed that the biggest changes in conditions occur when the stakes change for the test administrators, not the students. That is, a new requirement that students pass a test in order to be promoted to the next grade certainly increases the stakes for students, and that can be expected to have an impact on the test results. But a new requirement that schools have certain average test scores in order to earn rewards or avoid sanctions is likely to change the testing conditions and impact the test results even more.

Increased stakes are likely to have an impact not only on the effort teachers and students put forth during the actual administration of the test, but also in preparation for the test. In particular, efforts to anticipate the actual test questions and prepare students to answer those increase as the stakes increase. Therefore, issues such as the security of items used for equating or coming from an item bank become more likely to affect test results when stakes increase. For example, some states draw their test forms from a large but publicly available item pool; the logic behind this often is expressed by stating, "If the students are willing to learn the answers to all these questions, then they really do know the subject matter." That's a position that probably is fair when stakes are relatively low. But if stakes are high, it would not be unusual to find students that know the answers to the questions in the pool, but have little knowledge of the subject matter beyond those questions. In such cases, results across years when the stakes change could not be considered to reflect comparable changes in real achievement.

In order for conditions to be equivalent across years, the accommodations that students are permitted must remain constant. If students are prohibited from using calculators one year, but then allowed to use them the next, the results across those years will not be directly comparable.

Similarly, equivalence of conditions requires that students be given equal time. If students are allowed 30 minutes to complete the test one year, but then given unlimited time the next year, the results across years will not be comparable.

An interesting question is how one would consider changes when it is known that course-taking patterns have changed across years. Suppose, for example, that the percentage of students taking more rigorous mathematics course in high school increases, and over that same period of time, test scores increase as well. Certainly, if everything else were equivalent, we could persuasively argue that mathematics achievement had increased. But we would not know whether the increase was due to more effective teaching within the existing course, or the greater percentage of student who had an opportunity to learn advanced mathematics skills.

Another issue related to opportunity to learn is the time during the school year that the test is administered. Obviously, scores across years cannot be directly compared when there is a change in the time of the year that the test is administered. But often the time issues are more subtle than that. For example, if the number of schools operating on a year-round schedule increases, but the test administration window remains fixed, it will be unclear how much of the school year passed before each student took the test—and the number will be different, depending upon the year-round schedule the student is enrolled in. This assumes, of course, that it is the number of days of instruction that most affects opportunity to learn. But the age of the student is also a factor. Suppose, for example, districts change school schedules so that schools open two weeks earlier each year, and the state, in turn, move the testing schedule up two weeks. Students still have the same number of days of school before testing, but they now are two weeks younger when they take the test? Would that have an effect on test scores? Anyone who doubts that school administrators believe it would have no further to look than Florida. Districts have started school earlier and earlier in the year to give students more time to prepare for the FCATs. The practice finally became so problematic, and so many complaints arose from parents about the truncation of summer vacation, that the legislature proposed a date for the beginning of the school year.

A final element of "equivalent conditions" relates to external factors. For example, when there is a tragic, traumatic event in a school shortly before testing, the testing conditions in that school are not equivalent to those of prior years. Sometimes an event will be so major that it effects several schools or even an entire state. Examples of these in previous years are the destruction of the World Trade Center towers in New York City and the hurricanes that devastated sections of the Gulf Coast in 2005. As a minimum, test results for the schools that had students affected by these events need to be marked with an asterisk. While the test results might have reflected the actual lower level of academic achievement, there also was the additional element of psychological trauma that almost certainly caused the results to decline as well.

Equivalent Scoring

Multiple-choice questions will be comparably scored across years, but the scoring of open-response questions and essays can vary greatly. In the past, some states and testing contractors have operated as though equivalence of scoring training has translated into equivalence of scoring, but that is not necessarily the case. The only way to ensure that scoring is equivalent is to mix actual student work from the previous year into the scoring process (in a manner so that it is indistinguishable from the student work of the current year) and then see whether the scores attained by those papers when scored this year are the same as the scores assigned the previous year. When student responses were scored on paper, this was usually not possible to do. With the advent of document imaging and the consequent scoring on computer screens, however, this usually can (and should) be done.

Equivalence of scoring, however, is a more general issue than simply assigning the same scores to the same individual item responses. "Scoring" is the assignment of a result across the entire pattern of information provided by the examinee, and that too much be consistent from year to year. An example of how this might be problematic comes from the scoring of the SAT. When results are reported each year, the score produced for each student is the highest one the student attained over all the administrations that the student elected to take. Given measurement error, the more times a student chooses to take the SAT, other things being equal, the higher the score the student will attain. Over recent years, the average number of times students take the SAT has declined. Over the same period of time, the national reported average of SAT takers has declined—but a significant part of that decline in scores can be attributed to the decline in SAT test-taking opportunities.

Questions to Ask

The following are questions one should ask and answer before attributing changes in assessment results to improvements in the educational system.

A. Equivalence of test

1. Were the frameworks the same both years?
2. Were the content specifications the same both years?
3. Were the equating items presented identically? To answer this question, one should look at the actual test booklets themselves, seeing that the questions were identical (e.g., that the distractors were the same, the font used was identical, and the ink was equally readable in both administrations), and that the questions surrounding them did not lead to changes in the equivalence of the equating items.
4. Were the equating items located in the same position, both within the booklet and within the entire administration process?
5. Were the equating items representative of all the questions in the test? To answer this question, use the process of generating inter-item correlations presented earlier in this paper.
6. Were the students taking the equating items representative of all students taking the entire test? This question is important if the test administration process involves something other than having all students take all the equating items.
7. In the process of conducting the equating, were any items discarded? If so, which ones and for what reasons? In a similar vein, did the analysis suggest that the context for any of the equating items changed across the test administrations?

B. Equivalence of population

1. What was the total number of students enrolled at the time of each administration? What was the enrollment by significant subgroup?
2. What portion of the enrollment actually took the test for each administration? (Again, this should be answered for each significant subgroup as well as the total population for the state.)
3. In particular for high school students, did the dropout rate change, or is there evidence about other events that might have changed the population of students enrolled? Examples of this would be significant shifts of students to or from public schools to charter or private schools that are not included in the statewide results, and changes in promotion policies that lead to more students repeating certain grades.

C. Equivalence of conditions

1. Have the stakes changed for students or schools?
2. Have the accommodation policies changed?
3. Has the number of students being provided accommodations changed?
4. Has there been a change in the amount of time students are given to complete the test?
5. Has there been a change in the time of the school year that the test is administered?
6. Has there been a change in course-taking patterns of tested students?
7. If the answer to any of the above questions is positive, then have the changes been consistent from district to district, or are there some districts that have changed more than others? Note that there routinely will be fluctuations in district scores from year to year, so one would need

to look at the changes observed over two previous administrations in order to estimate what a reasonable change would be. The issue then would be to determine whether any districts have changed more than one would expect.

D. Equivalence of scoring

1. Have the constructed-response questions been scored the same across administrations?
2. Have any of the rules changed on how a student's total score is obtained?

One final caution is to ensure that the amount of change is more than one would expect due to random fluctuations. Given that the students tested each year are a conceptual sample of all the students that might be tested, one needs to recognize that there is sampling error associated with any result, but for state total results, this is likely to be a small number. If a state has 10,000 students a grade level, then a 95 percent confidence interval around the percentage of proficient students should be no larger than a point; with 100,000 students, it should be no larger than about one-third of a point. However, the error associated with equating can be considerably larger than that, so that statistic should be calculated, and changes in test scores smaller than that amount should obviously be interpreted with caution.

However, if the answers to all the above questions are negative and the observed change is more than one would expect from random fluctuation, then it is likely that the true reason for the change is a change in the effectiveness of the educational system. Congratulations to your state for accomplishing that worthwhile goal.