

Using P-Value Statistics to Determine the Believability of Equating Results

Richard Hill

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

Paper presented at the National Conference on Student Assessment
Orlando, Florida

June 17, 2008

Background

At last year's CCSSO conference, I presented a paper that outlined several issues states should look at before concluding that gains observed in their test scores are due to improvements in their educational system, and not an artifact of other explanatory factors¹. My experience with states' data this year has made me realize that paper starts a little too late in the process; as a first step, states need to satisfy themselves that the gain score changes they are observing from year to year are a result of students' achievement levels and not some anomaly in the equating process. This paper outlines a process for doing that.

A Simple First Step

The first step in looking at a state's results across years should include the data in Table 1. I have entered sample data into the table for the first three equating items. Typically, there will be considerably more questions administered in common across years, and the data should be entered for each question. Then, the average change in item position and p-value should be calculated as well as the standard deviation of p-value differences. For this first example, we are going to assume that a state administers its equating items in matrix sampled forms, that the forms are uniformly distributed across students, and all the questions are of multiple-choice format. Later, we will discuss how to review the data when there are other administration procedures, but we will start with the simplest example.

¹ *Comparability of Assessment Results across Years*, available at http://www.nciea.org/publications/CCSSO2_RH07.pdf.

Table 1

Data Necessary to Examine Results across Years

Item	Year 1					Year 2					Position diff	p-value diff
	Form	Sequence		p	N	Form	Sequence		p	N		
		Sess.	Item Position				Sess.	Item Position				
1	3	2	10	51.6	5214	6	2	10	52.3	5315	0	+0.7
2	5	2	8	67.8	5186	4	2	8	66.6	5312	0	-1.2
3	6	2	9	87.2	5210	7	2	9	87.7	5310	0	+0.5
...												
Ave.											0	+0.3
SD												1.0

This table provides a simplest case. All the equating items are in the same position as they were the previous year, the numbers of students taking each item is approximately equal across forms within year, the numbers of students taking each item is slightly larger in Year 2 than in Year 1 (perhaps because of a modestly growing school population in the state) and the changes in p-values across years are fairly similar—and small—for all equating items. Note that the data requested in the table allows one to make all the statements in the previous sentence—and one must be able to make all those statements in order to simply interpret the results. Given these data, the mean scaled scores across years *must* be approximately equal, regardless of student performance on all the other questions in the tests the two years. No information about any of the other questions on the test is needed to know that the mean scaled score across the years cannot have changed—and if it were reported that the mean scaled scores had changed, the above data would be sufficient to know that those results should be questioned.

So the very first point to make in taking equating “out of the black box” is to note that one needs only the information in Table 1 to determine whether mean scaled scores across years should be basically unchanged, increasing or decreasing. No item response theory statistics are needed—just simple descriptive statistics about performance on the questions across years. And the table provides information that IRT information would not, such as whether the Ns across the years are comparable and whether the equating items changed position—and such data is critical to the interpretation of change across years.

Thus, all the information in Table 1 should be the first thing to look at for any person who is responsible for test score changes over years. It tells you:

- whether the numbers of students who took each item is consistent with the assumption that the forms have been uniformly distributed throughout the state,
- whether the equating items have been maintained in a constant position,
- whether performance statewide has changed, and
- whether that change is generally consistent across all the equating items.

When the answers to those questions are positive, it is easy to determine what the changes in scaled scores should be. When the answer to any of them is negative, it makes it more difficult to make that determination, but it is important to do that, as the following example will show.

An Example of Resolving a Problem

The example I am going to present shows the value and versatility of the information in Table 1. When everything is straightforward (equating items are distributed over matrix sample forms that are uniformly distributed throughout the state, equating items are presented consistently across years, etc.), Table 1 allows us to make a reasonable estimate of the changes we should expect to see in mean scaled scores. This example, however, will show how Table 1 can provide the evidence that things are not straightforward in some cases, and point the direction to follow to uncover some very subtle equating issues.

This example came to my attention through a call from a state testing director. It was his opinion that test scores his contractor were reporting had increased too much from 2006 to 2007 to be believable, and he wondered whether there might be something wrong with the equating. While he was pleased that scores had gone up, he was concerned that if indeed there was something amiss, it could lead to a decline in scores from 2007 to 2008; before he was faced with that situation, he wanted a deeper understanding of what had caused the increase.

The story of how we uncovered what had happened is a long one, and it was only after it was over that I realized how valuable the data in Table 1 would have been had we looked at that first. Therefore, rather than following the story chronologically, I will show what we would have done if we had had the wisdom to create Table 1 first.

The state from which the example is taken administers tests in both mathematics and reading to all students at grades 3-8. The tests are mostly multiple-choice; students also take a few constructed-response questions. Each student takes a common section in each content area; the test also contains matrix-sampled sections, embedded in the middle of the common section. The equating items are drawn from the matrix-sampled section of the previous year's test. Because the constructed-response questions were such a small part of the test, they will be ignored for this example. However, we will devote a section at the end of this paper to special issues associated with such items.

While test scores increased from 2006 to 2007 at every grade in our example state, they changed the most in grade 5 reading. The mean scaled score increased by 17 points (the standard deviation of student scores was around 70, so this was an increase of over .2 standard deviations); the percentage of students at Proficient or higher increased from 39 to 52. For this reason, we decided to focus our attention first on grade 5 reading to see whether we could uncover what was going on.

The reading test in this state is administered in two sessions. In the first session (Session 2), students are presented with five passages. Each passage has six associated multiple-choice questions; two of the passages also include a constructed-response question. The first two passages are common, the second two are matrix-sampled, and the last passage is another common passage. In Session 3, students are presented with four passages, all common. Again, each passage includes six multiple-choice questions; two also have a constructed-response question. Thus, each student takes 42 multiple-choice and three constructed-response questions in common; the matrix-sampled section includes 12 multiple-choice and one constructed-response question.

Table 2 provides the information defined by Table 1 for all the equating items.

Table 2
Grade 5 Reading Results for Equating Items

Item	2006					2007					Position diff	p-value diff
	Form	Sequence		p	N	Form	Sequence		p	N		
		Sess.	Item Pos.				Sess.	Item Pos.				
1	3	2	14	67	4087	0	2	7	71	32965	-7	+4
2			15	64				8	67		-7	+3
3			16	53				9	57		-7	+4
4			17	74				10	77		-7	+3
5			18	82				11	85		-7	+3
6			19	66				12	67		-7	+1
7	8	2	14	75	3965	1	2	21	64	4158	+7	-11
8			15	81				22	70		+7	-11
9			16	77				23	66		+7	-11
10			17	60				24	54		+7	-6
11			18	69				25	61		+7	-8
12			19	44				26	40		+7	-4
13	8	2	14	75	3965	2	2	21	72	3996	+7	-3
14			15	81				22	81		+7	+0
15			16	77				23	76		+7	-1
16			17	60				24	59		+7	-1
17			18	69				25	71		+7	+2
18			19	44				26	43		+7	-1
19	7	2	21	52	4089	3	2	21	54	3980	0	+2
20			22	65				22	70		0	+5
21			23	62				23	66		0	+4
22			24	72				24	76		0	+4
23			25	54				25	56		0	+2
24			26	83				26	86		0	+3
25	7	2	21	52	4089	4	2	21	55	4047	0	+3
26			22	65				22	69		0	+4
27			23	62				23	65		0	+3
28			24	72				24	76		0	+4
29			25	54				25	57		0	+3
30			26	83				26	87		0	+4

The information in Table 2 reveals some basic information about the equating. Three passages, with six associated questions each, were drawn from three matrix-sampled forms in 2006 to form the equating link. One passage was inserted into the common section of the test, which required that it be moved one passage (seven items) forward in the test. Of the remaining two passages, each was included in the matrix-sampled section of two forms in the 2007 test. One remained in the same

position as it was in 2006, while the other was moved back one slot (to counter-balance the effect of moving the common passage up one slot).

Patterns in the data will be easier to see if we average results across the items within each passage. Those averages are presented in Table 3 (all the results are taken directly from Table 2).

Table 3
Averages of Results across Items within Passage

Passage Set	2006			2007			Average Diff. in Position	Average Diff. in p-value
	Form	Average p-value	N	Form	Average p-value	N		
1	3	68	4087	Common	71	32965	-7	+3
2	8	68	3965	1	59	4158	+7	-9
3	8	68	3965	2	67	3996	+7	-1
4	7	65	4089	3	68	3980	0	+3
5	7	65	4089	4	68	4047	0	+3

Finding out that forms were not randomly distributed to students. The first immediate observation to make is how different the results were for Forms 1 and 2 in 2007 on the equating items—and in contrast, how similar the results were for the equating passage used in Forms 3 and 4. Students taking Form 2 scored between 3 and 11 percentage points higher on the same questions as students taking Form 1, while the results for the equating items on Form 3 were within one percentage point of those for Form 4. If one uncritically accepted the results of the first equating passage (the one moved into the common section) or of the ones moved into Form 3 and Form 4, one would believe statewide performance had increased a fair amount. On the other hand, looking at the results for Form 2 would lead one to believe that achievement had dropped a small amount—and the results for Form 1 make it appear as though achievement in the state plummeted across the years. Before taking an average across all these disparate results (note that I left that last row out of Table 2), it is important to resolve these differences. An average of several vastly different results makes no sense until one understands why the results are so different and takes those differences into account.

An important first step in looking at data such as these is to compute the standard error, given that the results are drawn from samples. The formula for the standard error of any item across years is as follows:

$$SE_{p_2-p_1} = \sqrt{\frac{p_1 * (1-p_1)}{N_1} + \frac{p_2 * (1-p_2)}{N_2}}$$

In this case, the standard error for the items used in the matrix-sampled section of the test both years is about 1 percentage point; for the items moved to the common section, it is considerably smaller than that. So the variation we see between Form 3 and Form 4 is within the range of reasonable fluctuation and can be attributed to that—but the variation we see between Form 1 and Form 2 is well outside that range, so we need to look for an alternative explanation for those data.

A subtle clue to the answer lies in the differences between the number of students taking Form 1 versus the other forms. While the increased N (about 4 percent) is not enough to explain the differences in results between Form 1 and Form 2, it leads one to naturally question why more students would take Form 1 when, so far as the state knew, the forms were being randomly distributed to students within classroom.

It turned out that a rumor had started within the state that Form 1 was easier than the other forms (which of course makes no sense, since all the items that count in a student’s score come from the common items). As a result, it was common practice to reserve Form 1 for all students who needed a special administration of the test. Fortunately, all students had taken a set of items in common, so we were able to use that information to look at the differences among the achievement levels of the students taking the various forms. Those results are provided in Table 4.

Table 4

Deviation from State Average on Common Items in Grade 5 Reading in 2007,
Reported as Percentage of Total Score, by Form

Form	Deviation from State Average
1	-9.1
2	+0.2
3	+2.3
4	+1.9
5	+1.4
6	+2.0
7	+1.8
8	+1.3

The data in Table 2 now start to make sense. The students who took Form 1 were about 10 percentage points less able than the students who took Form 2, who in turn were about 2 percentage points less able than the students who took Form 3 and Form 4. If one subtracts 2 percentage points from the Form 3 and Form 4 results, and adds 9 to the Form 1 results, the picture is much more consistent, but not entirely so. After this adjustment, there is a small gain for the Form 3 and 4 results, but a small loss for Form 1 and 2.

Having realized how important it was control for the differences in achievement by form in 2007, we looked at the 2006 data and found there were small but significant differences among the forms in 2006. Those are provided in Table 5.

Table 5

Deviation from State Average on Common Items in Grade 5 Reading in 2006,
Reported as Percentage of Total Score, by Form

Form	Deviation from State Average
1	-4.5
2	-0.2
3	+0.9
4	+1.6
5	+0.6
6	+1.8
7	+0.7
8	-0.6

To properly adjust the scores, one must add the 2006 deviation and subtract the 2007 deviation to the observed results shown in Table 2. Those calculations are provided in Table 6.

Table 6

Results for Equating Items after Adjustments

2006 Form	2007 Form	Ave. Change in p-value	2006 Adjustment	2007 Adjustment	Net
3	Common	+3.0	+0.9	0.0	+3.9
8	1	-8.5	-0.6	+9.1	0.0
8	2	-0.5	-0.6	-0.2	-1.3
7	3	+3.7	+0.7	-2.3	+2.1
7	4	+3.8	+0.7	-1.9	+2.6

Finding out that changes in item position are important. These results are getting closer to telling a consistent story, but the three different sets of equating items are still providing somewhat different results. If one just looked at the first equating passage, one would believe performance improved significantly across the two years; the last passage shows some improvement across the years, but substantially less than the first one; and the second passage shows either no gain or a loss. Note that Table 2 also provides information about position of the equating items in the test. From that table, we can see that the first passage was moved to an earlier position in the test, the last passage remained in the same position, and the second passage was moved to a later slot. These changes in item position were relatively small (one passage position, 7 items). Could it be possible that these small changes would explain the differences in results among the equating passages?

With just three passages, one should not draw a conclusion about this effect for one grade. Fortunately, the same pattern of moving some passages forward and some back was used at other grade levels, and we therefore could see whether the result we observed at grade 5 would hold up at the other grades. Those results are provided in Table 7.

Table 7

Statistics for Equating Items on Reading Test

Position Change	Statistic	Grade				
		4	5	6	7	8
+7	# Items	6	6	6	0	6
	# Forms	2	2	2	N/A	2
	Change	-2.8	-0.7	-3.9	N/A	-3.4
0	# Items	6	6	12	18	12
	# Forms	2	2	2	2	2
	Change	-0.5	+2.3	+1.3	+1.0	+0.6
-7	# Items	6	6	6	6	6
	# Forms	8	8	8	8	8
	Change	+2.3	+3.9	+2.9	+3.8	+2.2

Table 7 shows that the pattern was consistent from grade to grade. Although the grades showed different amounts of change, at every grade the net adjusted gain for the items that were moved forward was more positive than the gain for the items that remained in the same position, and a loss was shown for the items that were moved backward. On average, the effect of moving items forward one passage slot (seven items) was a little over 2 percentage points; the effect of moving items backward the same amount was large—over 3.5 percentage points. On average, the effect of moving items 14 positions was about 5.5 percentage points, or about 0.4 percentage points per item position. This is rather dramatic evidence of how critical it is to take changes in item position into account when looking at test results across years.

There are two footnotes to the above finding. First, I was presented with the reading test data from a very large state (so that the standard errors of the observed p-value results were very small) where there appeared to be some minor anomaly in the findings across years. While the anomaly was small, it was bothersome because the standard errors were even smaller than that. It turned out that the equating items had moved a small number of positions from one year to the next. When the rule of thumb of “changing an item position of 1 leads to a change in p-value of 0.4” was applied, the anomaly completely disappeared—that is, the rule of thumb completely explained the apparent discrepancy in the results across years, thus reinforcing the finding from the original state. The second footnote is that none of what I found for reading in this state applied to their mathematics results. Items shifted in position across the years in ways that were systematic enough to study, but no pattern emerged. In mathematics, sometimes performance declined when items were moved later in the test, and sometimes it improved. In addition, there was no discernable pattern across grades. So as convincing as the finding appears to be for reading, it seemingly applies to that content area only and should not be generalized to others.

Checking the reasonableness of the reported gains. We have digressed considerably from the original charge given to me by the state testing director. The original problem was to uncover why there was such a large gain in scaled scores from one year to the next, but in the process, we have learned two important things about the state’s equating data: first, the forms are not uniformly distributed across the state as was supposed; and second, changing the position of equating items (at least reading items) has an effect on performance that was far larger than was anticipated. The path to uncovering both things was triggered by carefully examining the information in Table 2. But now,

back to the original question, we know that once we take these two factors into account, performance on the grade 5 equating items increased a little more than 2 percentage points. Is that amount of improvement consistent with the reported scaled score gain of 18 points?

That question can be answered through the common items. The standard deviation of grade 5 reading scaled scores is about 70 points; the standard deviation of those same scores, when expressed as a percentage of total raw scores, is about 20. A gain of 2 percentage points (about .1 standard deviations) should translate into about 7 scaled score points (also about .1 standard deviations). So, no, the increase in performance on the equating items, while impressive, is not consistent with the increase in the reported scaled scores.

The equating issue. The intent of this paper is to show a method for checking the reasonableness of equating results, not to identify particular equating methods that may or may not be problematic. Therefore, this section will not go into any detail, but instead summarize our findings.

First, the non-uniform distribution of matrix-sampled forms was not the source of the discrepancy. The purpose of equating is to find the *relative* difficulty of items. So if a subpopulation on which is the equating is done is not representative of the population as a whole, that usually will not be a problem so long as the relationships among the difficulties of the items is the same for the subpopulation as it is for the larger group. That was the case here, so that was not a concern in the review of the equating process. While it was important to understand the differences in achievement among the subgroups that took the various matrix-sampled forms when we looked at the p-values, that was not a factor in the equating.

On the other hand, the fact that moving the equating items had an effect on their difficulty did influence the results of the equating. However, that alone was not the reason for the difference between the data in Table 2 and the increase in scaled scores. The change in item position affected both the p-values and the scaled scores. One point noted in this paper is that it was easier to uncover that issue from Table 2 than it would have been by looking at IRT statistics. Indeed, until we started looking more deeply into the issues through the p-values, no one was aware of the impact of item position on item difficulty.

It turned out that the reason the gains in scaled scores did not match up to the gains in p-values was a complex interaction of item position, the use of equating items in the matrix-sampled portion of the test vs. the common section, and the choice of equating model. When an alternative equating model was chosen, the gain in scaled scores dropped to 11 points—not exactly the 7 points we believed could be justified from the gain in p-values, but considerably closer than the 18 points that had been originally reported.

Using Constructed-Response Questions in Equating

In the example, we ignored the data from the constructed-response questions even though they were available. That was due, in part, to the fact that the intent of the study was to determine the believability of reported scaled score gains, as well as the fact that the constructed-response questions consisted of a small portion of the test. However, in general, the information from constructed-response questions should be used in equating, and should be looked at in much the same way as the multiple-choice questions.

However, the fact that constructed-response questions are scored by humans adds another dimension to the data that must be looked at. If performance on a constructed-response question improves, it may be because students are doing a better job of answering the question, but it also might be because the scorers this year are scoring it differently from previous years.

While those conducting the scoring sessions will often argue that the scoring is consistent because of the consistency of training across years, those assurances are insufficient. The only way to know for sure whether the scoring has remained consistent across years is to incorporate a sample of papers from the previous year's scoring into the current year's process, and have them rescored blindly. The scores those papers received in the previous year should be compared to the scores they received from the current year's scorers. Any significant average discrepancies would be an indication that changes in the scores of these questions cannot be accepted uncritically. At worst, the items would need to be considered unusable for equating purposes, since the scores of one year's questions do not match those of another.

Comparing Percentages of Proficient Students across Years

Throughout this paper, I have talked about comparing mean scaled scores—rather than the percentage of students scoring at the Proficient level or higher (PAC)—across years to see whether scores have actually changed. This was not an oversight. While PACs are a fine reporting statistic (and with the advent of NCLB reporting, a predominantly used statistic), comparing them across years can be misleading. The problem with PACs is that the legitimate changes that might occur across years often are less than the percentage of students scoring at a particular raw score point around the cut. That is, it might be reasonable that the percentage of Proficient students in a state might increase by 2 or 3 percentage points from one year to another, but that 5 percent or more might receive a raw score just above or below the cut. If the difficulty of a test changes slightly, the equated raw score might change by a tenth of a raw score point or so. However, the cut score that determines the PAC is usually a whole number. As a result, it is possible for student performance to improve (as determined by changes in the mean scaled score) and still have the PAC go down—or up far more than the changes in mean scaled score would indicate it should. The granularity of the raw score scale makes PAC too crude a statistic to use effectively in the examination of score changes across one year.

Summary

The recommendation is to start the review of every year's results by obtaining the data in Table 1 for every subject area at every grade. As a minimum, first look at the average change across years in p-values of the equating items. Any average difference of more than 2 percentage points for any subject/grade combination should be questioned—that value would be a very significant increase for a state. In addition, the standard deviation of the differences should be small, especially if the equating items have been administered to thousands of students. If the standard deviation is much more than a point, one should look at the changes for individual items and look closely at items that have changed substantially more or less than the average item. Were those items that measured content for which there was significantly more or less attention statewide this year compared to previous years? If so, perhaps the differential change is justifiable. If not, however, the administration of those items should be looked at carefully. Were they located in the same position as the previous year? Was the presentation of the item *identical*? (There are plenty of examples where the slightest change—even the color of the ink used—caused performance on items to change substantially.)

If the change in average p-value of the equating items is not believable (and I would suspect any change of more than 2 percentage points in a year), then one should look at the list of issues to double-check that were included in the paper presented at last year's conference (referenced on Page 1); e.g., making sure the tested populations were equivalent, and that administration rules remained constant. That is, check whether there is something fundamentally wrong with the data that have been collected that equating cannot fix. The equating process simply will assume these data are accurate and proceed from that point. Equating cannot fix data that are fundamentally flawed.

If the change in average p-values seems realistic, then the next question would be whether the change in mean scaled scores is consistent with the change in p-values. A reasonable rule of thumb would be that scaled scores should change by one-twentieth of a standard deviation for each percentage point the equating items have changed. Deviations substantially more or less than that amount should be questioned, and the equating process should be reviewed in detail until there is an explanation for the apparent discrepancy between the two statistics.